Development and Objective Perceptual Quality Assessment of Monaural and Binaural Noise Reduction Schemes for Hearing Aids

Von der Fakultät für Mathematik und Naturwissenschaften

der Carl-von-Ossietzky-Universität Oldenburg

zur Erlangung des Grades und Titels eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

angenommene Dissertation

Dipl.-Ing. Thomas Rohdenburg

geboren am 28. Mai 1975

in Bremen

Erstreferent: Prof. Dr. rer. nat. Dr. med. Birger Kollmeier

Korreferenten:

Prof. Dr.-Ing. Karl-Dirk Kammeyer

PD Dr. rer. nat. Volker Hohmann

Tag der Disputation: 17.12.2008

Abstract

People with hearing impairment have great difficulties communicating in noisy and reverberant environments. They usually require a higher signal-to-noise ratio (SNR) to achieve the same listening performance as normal hearing people. But even for normal hearing people a noise reduction is desirable in these environments, especially when using modern communication systems such as mobile phones, handsfree devices, or teleconferencing systems. Hence, the development and evaluation of noise reduction algorithms is an active field of research. To achieve the maximum performance and subjective benefit, these algorithms generally need to be evaluated with objective measures that should be based on auditory models in order to predict human perception as closely as possible. The current dissertation contributes to this field by adding yet another dimension to the problem - binaural hearing (i.e., listening with two ears).

As a starting point of this thesis, some frequently used objective performance measures and novel objective measures based on recent knowledge of the auditory system are reviewed. Using subjective listening tests on signals processed by monaural noise reduction schemes, those measures are identified that exhibit the highest correlation with subjective data. Using these measures, it is possible to optimize single-channel noise reduction algorithms on a perceptual scale. However, the performance of single-channel noise reduction systems is limited and leads to a trade-off between signal distortion and noise reduction. Therefore, multi-channel beamformer algorithms are investigated throughout the main part of the thesis. In general, they introduce less signal distortions and achieve higher noise reduction for non-stationary signals. Typically, beamformers do not have binaural outputs. To provide the user with information on the spatial arrangement of signals, different strategies to partially preserve or reconstruct the binaural information are developed and evaluated with objective and subjective assessment data. For the objective evaluation of the benefit on binaural speech perception, a novel binaural speech intelligibility measure is used and compared to subjective data. For the binaural beamformer schemes, the influence of head diffraction on the performance is analyzed. It is shown, that at least simple head-models should be integrated into the algorithm design to increase the performance compared to traditional free-field designs and to make use of the acoustic decoupling due to the head shadow effects. In order to utilize the algorithm

under real-life conditions, the problem of moving target signals and head movements is tackled, that can reduce the effective benefit for the user of the microphone-array hearing aids. A self-steering beamformer is finally developed and evaluated under realistic reverberant conditions, that exhibits increased performance compared to non-steered systems at positive signal-to-noise ratios.
It is expected that both, the algorithms and objective assessment methods, will prove to be beneficial for human communication in noise in the future.

Zusammenfassung

Menschen mit einer Hörproblematik haben große Schwierigkeiten in lauten und widerhallenden Umgebungen zu kommunizieren. Gewöhnlich benötigen sie einen höheren Störpegelabstand als normal hörende Menschen, um dieselbe Hörleistung zu erzielen. Aber selbst für Normalhörende ist eine Störgeräuschreduktion in diesen Umgebungen erwünscht, besonders wenn moderne Kommunikationssysteme wie Handys, Freisprecheinrichtungen oder Telekonferenzanlagen verwendet werden. Daher ist die Entwicklung und Evaluation von Störgeräuschreduktionsalgorithmen ein aktives Forschungsfeld. Um den maximalen Nutzen und subjektiven Gewinn zu erzielen, müssen diese Algorithmen im allgemeinen mit objektiven Maßen bewertet werden, die auf auditorischen Modellen basieren sollten, um die menschliche Wahrnehmung möglichst genau vorherzusagen. Die vorliegende Dissertation trägt zu diesem Forschungsfeld bei, indem sie dem Problem noch eine weitere Dimension hinzufügt - binaurales Hören (d.h., Hören mit zwei Ohren).
Zum Einstieg dieser Doktorarbeit wird ein Überblick über einige häufig genutzte objektive Gütemaße und neuartige, auf aktuellen Kenntnissen des Auditorischen Systems beruhende, objektive Maße gegeben. Mit Hilfe von subjektiven Hörtests, mit den durch monaurale Störgeräuschreduktion verarbeiteten Signalen, können die Maße mit der höchsten Übereinstimmung mit subjektiven Daten identifiziert werden. Mit Hilfe dieser Maße ist es dann möglich, einkanalige Störgeräuschreduktionsalgorithmen anhand einer perzeptiven Bewertungsskala zu optimieren. Allerdings ist die Leistungsfähigkeit von einkanaligen Störgeräuschreduktionsalgorithmen begrenzt und führt auf einen Kompromiss zwischen Signalverzerrung und Störgeräuschreduktion. Daher wurden in dieser Arbeit hauptsächlich mehrkanalige Beamformer-Algorithmen untersucht. Im allgemeinen führen diese zu geringeren Signalverzerrungen und erzielen eine höhere Störgeräuschreduktion bei nichtstationären Signalen. Normalerweise haben Beamformer keine binauralen Ausgänge. Um dem Benutzer die räumliche Information über die Signale darbieten zu können, werden verschiedene Strategien entwickelt und anhand subjektiver Bewertungen evaluiert, welche die binaurale Information teilweise erhalten oder rekonstruieren. Zur objektiven Evaluation des Gewinns an binauraler Sprachwahrnehmung, wird ein neuartiges binaurales Sprachverständlichkeitsmaß verwendet und mit subjektiven Daten verglichen. Für die binauralen Beamformer wird der Einfluss der Beugung am Kopf auf die Leistungsfähigkeit analy-

siert. Es wird gezeigt, dass wenigstens einfache Kopfmodelle in den Algorithmenentwurf integriert werden sollten, um die Leistungsfähigkeit gegenüber traditionellen Freifeldentwürfen zu verbessern und die akustische Entkopplung aufgrund des Kopfschattens zu nutzen. Um den Algorithmus unter realen Bedingungen anwendbar zu machen, wird das Problem bewegter Nutzsignale und der Kopfbewegungen angegangen, welches den effektiven Gewinn für den Nutzer des Hörgeräte-Mikrofonarrays verringern kann. Ein sich selbst ausrichtender Beamformer wird schließlich entwickelt und unter realistischen widerhallenden Bedingungen evaluiert, der im Vergleich zu statischen Systemen eine verbesserte Leistungsfähigkeit bei positivem Störpegelabstand aufweist. Es ist zu erwarten, dass sich sowohl Algorithmen als auch objektive Bewertungsmethoden in der Zukunft als gewinnbringend für die menschliche Kommunikation im Störgeräusch erweisen.

Contents

Abstract iii

Zusammenfassung v

Contents ix

1. Introduction 1

2. Development of a Toolbox for Objective Quality Assessment of Noise Reduction Schemes 5
2.1. Introduction . . . 5
2.2. SNR Based Measures in Time and Frequency Domain . . . 6
2.2.1. Broadband Signal-to-Noise-Ratio . . . 6
2.2.2. Segmental SNR Measure . . . 7
2.2.3. Frequency Weighted SNR . . . 7
2.2.4. Signal-to-Noise Ratio Enhancement . . . 8
2.3. LPC Based Objective Measures . . . 8
2.3.1. Log-Likelihood Ratio Measure . . . 9
2.3.2. Itakura Saito Distortion . . . 9
2.3.3. Log-Area-Ratio Measure . . . 10
2.4. Perceptually Motivated Objective Measures . . . 10
2.4.1. Weighted Spectral Slope (WSS) Distance Measure . . . 10
2.4.2. PEMO-Q . . . 11
2.4.3. PESQ . . . 12
2.4.4. PEAQ . . . 14
2.4.5. BSIM . . . 14
2.5. Reference and Test Signal . . . 15
2.6. Discussion . . . 15

3. Monaural Noise Reduction Schemes - Performance Evaluation and Optimization 17
3.1. Introduction . . . 17
3.2. Comparison of Objective and Subjective Data for the Assessment of Single Channel Noise Reduction Algorithms . . . 18
3.2.1. Algorithms . . . 18
3.2.2. Signals . . . 18
3.2.3. Objective Measures . . . 19
3.2.4. Experiments . . . 20

3.2.5. Results . . . 20
3.3. Subband-based Parameter Optimization in Noise Reduction Schemes by means of Objective Perceptual Quality Measures . . . 25
3.3.1. Algorithm . . . 25
3.3.2. Perceptual Quality Measures . . . 27
3.3.3. Parameter Optimization . . . 27
3.3.4. Effects of the Noise Reduction on the Internal Representations . . . 28
3.4. Quality Assessment for Low-Power Applications . . . 30
3.4.1. Experiments and Results . . . 31
3.5. Conclusion . . . 32

4. Multi-Channel Noise Reduction Schemes with Binaural Output - Performance Evaluation and Optimization 35
4.1. Introduction . . . 35
4.2. Acoustical Setup . . . 37
4.3. Algorithm . . . 38
4.3.1. Signal Model . . . 38
4.3.2. Beamformer . . . 39
4.3.3. Binaural Output . . . 41
4.3.4. Influence of Different Propagation Models on the Beamformer Design 43
4.3.5. Algorithm Combinations . . . 47
4.4. Evaluation Methods . . . 48
4.4.1. Signal-independent Performance Measures and the Influences Of The Head . . . 48
4.4.2. Signal-dependent Performance Measures . . . 50
4.5. Experiments and Results . . . 51
4.5.1. Spatial Directivity Pattern . . . 51
4.5.2. Perceptual Optimization of the White Noise Gain Limitation . . . 52
4.5.3. Binaural Output Quality . . . 54
4.5.4. Performance Analysis of Adaptive and Fixed Beamformers . . . 54
4.5.5. Robustness Against Steering Errors . . . 55
4.5.6. Robustness Against Positioning Errors and Head Model Variation . 56
4.6. Subjective Listening Test . . . 57
4.6.1. Method . . . 57
4.6.2. Results . . . 58
4.7. Discussion . . . 60
4.7.1. Influences of the Head and Head Models . . . 60
4.7.2. Adaptive Versus Fixed Beamfomers . . . 60
4.7.3. Binaural Signal Reconstruction . . . 61
4.7.4. Objective Perceptual Measures . . . 62
4.7.5. Realistic Signal Conditions . . . 62
4.7.6. Subjective Listening Tests . . . 63
4.8. Conclusions . . . 63

5. Combined Source Tracking and Noise Reduction for Application in Hearing Aids 65
5.1. Introduction 65
5.2. Signal Model, Recorded Signals, And Binaural Multi-Channel Noise Reduction 66
5.3. Performance Of Direction Of Arrival Estimators 69
5.3.1. Generalized Cross Correlation Phase Transform (GCC-PHAT) . . . 69
5.3.2. The Spatial Response Pattern (SRP-PHAT) extension 70
5.3.3. Source Tracking Constraints 71
5.3.4. DOA Estimation Error 72
5.4. Objective Quality Assessment for the Complete Noise Reduction System . . 73
5.5. Objective Perceptual Quality Results for the Combined System 74
5.6. Discussion 75
5.7. Conclusion 76

6. Conclusions and Further Research 79

A. Tables 81
A.1. Table of Articulation Index 81
A.2. Critical Bandwidth / Equivalent Rectangular Bandwidth (ERB) 81

B. Acronyms 83

Bibliography 91

Erklärung 93

Danksagung 95

Lebenslauf 97

1. Introduction

Hearing impaired people have great difficulties communicating in noisy and reverberant environments. To achieve the same level of speech understanding, they generally require a signal-to-noise ratio (SNR) that is $5 - 10$ dB higher than a normal hearing person would need [57]. Similar problems also exist for normal hearing persons in acoustically *difficult* situations (e.g., train stations, car, large noisy rooms,...), especially when using telecommunication systems. This *communication bottleneck* under real-life conditions is a great challenge for our communicating and ageing society [42, 79].

The problem of poor listening performance in background noise has motivated a lot of research into noise reduction schemes that are applicable for hearing aids and other communication supporting systems. Many of these algorithms have their origin in telecommunication applications where computational complexity is no hindrance, primarily because of lower restrictions on power consumption. However, in modern times the boundaries between telecommunication applications and hearing aids become blurred. For all applications maximum signal quality and speech intelligibility at a relatively low power consumption are desirable. This opens the way towards more integrated research and development approaches. The current thesis contributes to this emerging field of research.

Noise reduction schemes are usually classified by the number of inputs (single channel or multi-channel) or outputs (monaural[1] or binaural). The differences of noise reduction schemes - and also their principle limitations - lie in the definition of (and assumptions on) *signal* and *noise*. The *desired target signal* in most cases is speech[2]. Thus, the terminus *speech enhancement* is often used interchangeably, whereas the definition of *unwanted noise* is often strongly related to the noise estimation method used in the algorithm. Noise reduction schemes aim at enhancing the *speech intelligibility*, the *ease of listening*, or other dimensions related to the audible signal quality, although sometimes only sub-goals can be reached.

Much research efforts have been put on single-channel (and monaural) noise reduction schemes in the short-time discrete Fourier transform (DFT) domain. These are characterized as the class of single-channel short time spectral attenuation (STSA) algorithms, and include Wiener-filtering, spectral subtraction [2, 7] and minimum mean squared error (MMSE) filtering techniques. STSA algorithms use time-varying spectral envelope filters and discriminate between speech and noise based on statistical properties of the signals. Generally, only magnitude gain factors are derived while the phase of the degraded signal is preserved in the processed signal. When the short-time spectral amplitudes can be accurately estimated, additional (and independently derived) information on the phase is of little use [76, 78, 81]. STSA algorithms generally have several parameters

[1] including dual monaural (diotic) representation

[2] For diverse applications of modern communications, other target signals are starting to attract increasing interest.

that influence the amount of noise reduction. Limitations of the single-channel STSA algorithms are given by the fact that a high noise reduction may attenuate and distort the desired speech signal. Another problem posed by STSA is that background noise can even become more annoying after the processing, e.g., by introducing random spectral peaks referred to as *musical noise* or *musical tones*. These reduce the audio-quality of the noise reduction algorithms so that a trade-off between noise reduction, speech distortion and other side-effects such as musical noise has to be found. Single-channel STSAs that rely on statistical estimates, have not been shown to improve the speech intelligibility until now [6, 50], except if they are used in conjunction with speech coders (e.g. low-bit rate codecs in telecommunication) [10] or cochlea implants [17, 46]. The current thesis contributes to objectively asses the potential benefits and limitations of these algorithms.

With directional microphones the spatial distribution of signal sources can be exploited to suppress signals deviating from the desired signal's direction. Hence, the definition of *noise* is *spatially* motivated. A directivity can be applied by a specific delayed interference of sound at the microphone sensors. In hearing aid applications *fixed* and *adaptive* directional microphones [21] have shown considerable success in speech intelligibility and quality enhancement. Microphone arrays of more than two microphones offer an even higher directivity and flexibility in signal processing strategies. Thus, much research has been done on *Microphone-Array Hearing Aids*, as summarized in, e.g., [26]. Because of their simplicity and achievement in many communication applications, several authors investigated microphone-array-based *beamformers* for hearing aid applications. The standard configuration for a single hearing aid on one side employs two or three microphones spaced closely together (maximum distance approx. two centimeters), and uses either a fixed or an adaptive beamforming algorithm [20, 27, 37] in order to typically enhance the signals emanating from a direction straight in front of the wearer's head. A higher functional array gain (i.e., improvement in signal-to-noise ratio assuming a constant microphone noise) at low frequencies can be achieved by applying *superdirectivity* [5, 11], which however increases the susceptibility to self-noise of the microphones and model errors, and by increasing the physical distance between the left-most and rightmost microphone of the array. Hence, several solutions for hearing aid wearers with external arrays, e.g., mounted in a pair of glasses [68, 69], have successfully been introduced. Generally, beamformers combine all input channels to a monaural output which disrupts the spatial impression of the enhanced signal. To preserve the binaural cues that could be used for localization, separation of acoustical objects and for spatial perception of the environment, beamformers with binaural outputs have been suggested in [13, 47, 82]. Another important point is the model of wave propagation. It was found in several studies [13, 17, 43, 51, 55, 84] that the performance of beamformers designed for free-field but used in head-worn systems was significantly reduced. However, up to now no comparative study of different types of head-models integrated into the beamformer design is available. Furthermore, the influence of head movements and moving sound sources has only been accounted for indirectly as a matter of robustness, as a detrimental effect to spatial filters with a sharp main lobe (e.g., [27]). The current thesis therefore treats these aspects in a systematic way.

Multi-channel Wiener filters (MWF) are another microphone-array based class of noise reduction schemes that are closely related to beamforming. In steady state and by assuming optimal estimates, multi-channel Wiener filters (MWFs) are equivalent to a *su-*

perdirective beamformer with a single-channel Wiener post-filter [67]. While beamformers need information about the relative microphone positions and the source direction, MWFs generally need a voice activity detection to update the noise estimate in speech pauses. Recently, MWFs have been investigated for use in hearing aids [15, 16, 39, 74, 75] including different strategies to preserve binaural cues. The decision between the using combined beamformer post-filter schemes or MWFs primarily depends on the signal conditions they are used in. However, the boundaries between the respective classes often overlap, especially for *noise-adaptive* beamformers. A comparison between adaptive and fixed and beamformers is therefore included in the current thesis.

Objective performance assessment of the aforementioned noise reduction schemes with perceptual models is a developing research field. Until the 1990s, the standard way to measure the quality of different speech processing schemes was to conduct a subjective listening test, which gives the subject group's mean opinion score (MOS) of the quality of each condition under test [61]. Testing human subjects in a controlled environment is the most valid and reliable method to measure quality, but is not feasible for the assessment of algorithms on a day-to-day basis during algorithm development. The goal of objective measurement is therefore to estimate MOS automatically based on direct measurements of a system or algorithm. In telecommunication applications many objective measures have been developed that consider signal degradations relevant for speech coding purposes. These measures have also been used in the development and assessment of noise reduction schemes. However, there is only little research on the significance of these measures for the subjective assessment of noise reduction schemes. The studies by Marzinzik [50] showed that besides signal quality the reduction of mental effort that is needed to listen to speech in noise is an important factor. Recently, binaural noise reduction schemes have been studied by many researchers. Currently, no reliable objective measure exists that includes perceptual effects of the binaural system. In summary, objective quality measures can only be evaluated reliably under realistic signal conditions and using recent noise reduction algorithms. Hence, the development and evaluation of noise reduction schemes and objective perceptual quality measures needs to be investigated simultaneously.

The goal of this thesis is to investigate possibilities of improving speech communication in adverse conditions for hearing-impaired and normal hearing listeners using hearing aids. In particular, the following research questions are investigated in this work:

- Which objective measures can predict speech intelligibility and mean opinion score of subjects for the different dimensions of signal quality of noise reduction systems, including binaural effects?
- What are the limitations of single-channel noise reduction systems under realistic conditions? How can the parameters that influence the noise reduction and signal quality be perceptually optimized with objective measures?
- How does head diffraction influence the performance of head-worn microphone arrays and can the influence be compensated?
- Is a binaural connection of microphone-array hearing aids beneficial and robust in realistic signal conditions?
- How do moving source signals and head movements influence the performance of these algorithms and is there a way to compensate?

To answer these questions, in this work the approach is taken to develop and evaluate noise reduction algorithms and objective performance measures simultaneously and check their significance with subjective listening tests.

Chapter 2 summarizes objective measures that are frequently used for the performance evaluation of noise reduction and speech coding systems and discusses measures based on current psychoacoustic models of the auditory system, including a novel binaural measure. With these objective measures, monaural noise reduction schemes are evaluated in chapter 3. Subjective assessment data is used to identify the objective measures that have the highest correlation with different dimensions of subjective quality. These measures are used for parameter optimization of novel noise reduction algorithms. In chapter 4, a class of binaural noise reduction schemes is evaluated with these objective measures and subjective listening tests. Head influences on the sound propagation, adaptation to realistic non-stationary noise fields, robustness against model errors, and different binaural output strategies are investigated. Finally, chapter 5 analyzes the problem of head movement and non-stationary moving source signals and proposes a combined direction of arrival estimation and binaural noise reduction algorithm to increase performance compared to fixed systems.

2. Development of a Toolbox for Objective Quality Assessment of Noise Reduction Schemes

2.1. Introduction

In order to alleviate speech perception and the reception of other *desired* sound sources in a noisy acoustical environment, considerable effort is being spent on provision efficient means to subjectively reduce the impact of noise on perception. Performance evaluation and patient benefit are important factors in the development of such noise reduction schemes, particularly for hearing aid applications. The most reliable assessment can be achieved by listening tests with a representative[1] group of subjects. However, subjective listening tests often are time-intensive, depend on instruction, training, assessment conditions, experience and hearing ability of the listeners and thus in practice may be inapplicable in early stages of algorithm development. Hence, much research has been focused on objective measures that quantify average subjective ratings of quality and speech intelligibility, possibly as a function of the hearing loss. The development goal is to maximize the correlation between objective measures with subjective data across a wide range of typical (and ecologically valid) listening conditions. Because audition is a complex process, most objective measures include more or less complex models of speech production and perception. This chapter provides a short overview of various objective measures proposed in the literature [59] that have been used to evaluate noise reduction or speech enhancement systems. Although historically most measures have been developed for the assessment of speech coding systems [45] these can (with modifications) also be applied to noise reduction schemes for speech-in-noise input signals.

All measures presented here need a reference signal (usually the undistorted desired signal or speech) and a test signal (generally, the processed output signal). Some measures additionally need the speech and noise component separated at the input and the output of the tested system. These measures are usually denoted as *intrusive* measures [61] in opposite to *non-intrusive* measures that need no reference signal and can, e.g., be used in a live network. Hence, for the *intrusive* measures investigated in the following, it is assumed that the desired signal s and the background noise n, are available separately before and after the processing. This type of processing is often used in simulated systems where all test signals are known and the separated signals can be processed with the previously saved filters that were calculated for the mixed input signal x (see Fig. 2.1). These systems are sometimes referred to as *master/slave* or *shadow-filter* processing schemes that allow for a relatively exact objective evaluation.

[1] e.g., listeners for which the algorithms under test are intended

The aim of this chapter is to provide a framework of different objective measures that have been used to evaluate speech enhancement systems of various types in research. A subset of these measures together with real-world audio signals recorded in typical acoustic environments for the hearing aid application is used to build an objective performance test-bench for the developed noise reduction schemes.

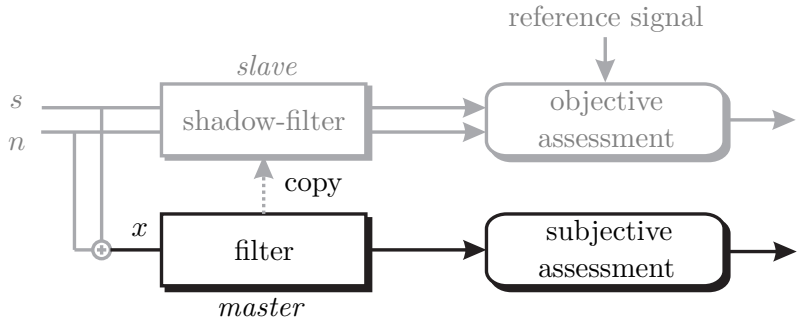


Figure 2.1.: Concept of the master/slave filtering system for objective quality assessment. The gray signal paths are generally only known during simulations and thus invisible to the algorithm indicated by the block *filter*.

2.2. SNR Based Measures in Time and Frequency Domain

The most commonly used performance measures are based on the SNR, i.e., the ratio of a *desired* signal power to the *undesired* noise power corrupting the signal, usually expressed in terms of the logarithmic decibel scale. Generally, the power is averaged either over the complete signal or over small time segments and can be evaluated for a broadband signal or in frequency bands (e.g., auditory filters). There exist several definitions of the SNR, depending on the way the signal and noise powers are calculated, averaged and weighted. Thus, a comparison of absolute SNR values over different studies is often difficult. For this reason, some researchers make their toolboxes of objective measures available in MATLAB®code [29, 45, 59]. Some of the most commonly used measures based on the SNR are defined below. In general, measures based on the SNR are not capable to assess all perceptually relevant target signal distortions, especially, if intermodulations between signal s and noise n occur.

2.2.1. Broadband Signal-to-Noise-Ratio

Generally, speech pauses need to be excluded before averaging the SNR. For the broadband SNR used here this is done as follows. First, the DC-component is removed from the *discrete*[2] speech and the noise signal, respectively.

$$\tilde{x}(k) = x(k) - \mu_x \tag{2.1}$$

[2]If not stated otherwise, all signals are adequately sampled and processed according to the sampling theorem.

Then the expected values of the signal and the noise power are estimated averaging only values that exceed a specific threshold, e.g., $thr_{dB} = -80$, thus excluding (speech) pauses.

$$\tilde{x}'(k_{thr}) = \begin{cases} \tilde{x}(k) & \text{if } |\tilde{x}(k)| \geq 10^{(thr_{dB}/20)} \\ [\,] & \text{otherwise} \end{cases} \tag{2.2}$$

$$\mathrm{E}\left\{|\tilde{x}'(k_{thr})|^2\right\} \approx \frac{1}{N_{thr}} \sum_{k_{thr}=0}^{N_{thr}-1} |\tilde{x}'(k_{thr})|^2 \tag{2.3}$$

The expected values of s and n are calculated by (2.2-2.3), thus the broadband SNR is:

$$\mathrm{SNR_{dB}} = 10\log_{10}\left(\mathrm{E}\left\{|s(k_{thr})|^2\right\}\right) - 10\log_{10}\left(\mathrm{E}\left\{|n(k_{thr})|^2\right\}\right). \tag{2.4}$$

2.2.2. Segmental SNR Measure

For segmental SNR measures the SNR is first calculated for signal portions of 10 to 30 ms and then averaged over these time segments (see eq. (2.5)). The segmenting makes the average SNR independent on the absolute level of the utterance in the short-time segment. However, the measure poses a problem if there are intervals of silence in the speech utterance [59]. In these segments any amount of noise will give rise to a large negative signal-to-noise ratio for that segment which would bias the overall SNR. To overcome this, either a limitation to a minimum SNR value can be applied or segments with low speech energy can be excluded from the averaging.

$$\mathrm{segSNR} = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \frac{\sum_{k=Nm}^{Nm+N-1} s^2(k)}{\sum_{k=Nm}^{Nm+N-1} n^2(k)} \tag{2.5}$$

In (2.5) M is the number of segments, N is the segment length and k is the sample index.

2.2.3. Frequency Weighted SNR

There exist many different frequency weighted SNR measures, some of them are summarized in [45, 59]. The following frequency weighted SNR has been suggested in [73]

$$\mathrm{fwSNRseg} = \frac{10}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} B_j \log_{10}[F^2(m,j)/|F(m,j) - \hat{F}(m,j)|^2]}{\sum_{j=1}^{K} B_j}, \tag{2.6}$$

where B_j is the weight placed on the jth frequency band, K is the number of bands, M is the total number of frames in the signal, $F(m,j)$ is the filterbank amplitude of the clean signal in the jth frequency band at the mth frame and $\hat{F}(m,j)$ is the filterbank amplitude of the enhanced signal in the same band. Hence $|F(m,j) - \hat{F}(m,j)|^2$ is an estimate of the remaining noise power after processing in the respective frame and band. Various forms of weighting functions B_j can be suggested, one possibility is to choose the weights B_j based on articulation index (see table A.1). In the following, another definition of an ERB-band

weighted SNR is used:

$$SNR_{\mathrm{ERB}} = \sum_{b=1}^{B} \frac{10\log_{10}\sum_{k=0}^{k=N-1}\left(\sum_{n=f_c(b)-\mathrm{ERB}(b)/2}^{f_c(b)+\mathrm{ERB}(b)/2} S(n,k)\right)^2}{10\log_{10}\sum_{k=0}^{k=N-1}\left(\sum_{n=f_c(b)-\mathrm{ERB}(b)/2}^{f_c(b)+\mathrm{ERB}(b)/2} N(n,k)\right)^2} \tag{2.7}$$

where $S(n,k)$, $N(n,k)$ denote the short time Fourier transform (STFT) of the discrete time signals $s(t), n(t)$ respectively, $f_c(b)$ and $\mathrm{ERB}(b)$ denote the center frequency and the the equivalent rectangular bandwidth (ERB) of band number b, respectively, and B is the total number of auditory filter bands. For the calculation of the center frequency $f_c(b)$ and bandwidth $\mathrm{ERB}(b)$ see A.3.

The main differences between the definitions (2.6) and (2.7) are as follows. In (2.6) the frequency- weighted short-term SNRs are added which may lead to a stronger bias if speech-pauses and fluctuating noise occur. On the other hand the difference term $|F(m,j) - \hat{F}(m,j)|^2$ in (2.6) also includes signal distortions between unprocessed and processed target signal components. However, also a broadband gain would bias this term which may be of less relevance for signal quality. Thus in summary, (2.7) seems to be a more robust estimate even though it assumes that a separate estimate of the noise is available.

2.2.4. Signal-to-Noise Ratio Enhancement

Often, one is not interested in the absolute SNR but in the difference of SNRs before and after the processing with an algorithm, e.g. a noise reduction scheme. The signal-to-noise ratio enhancement (SNRE) or ΔSNR is the difference between the SNR measured at the output of a noise reduction system and the SNR at the input. In the following, most of the (segmental) SNR measures defined above will be used as relative enhancement measures between unprocessed and processed signals which is indicated by a Δ or the suffix "E".

2.3. LPC Based Objective Measures

Several objective measures have been proposed that are based on dissimilarities between linear predictive models of unprocessed clean speech signal and processed enhanced (mixed) signal. These measures assume that - for short time intervals - speech can be represented by an pth order all-pole model of the form

$$x(k) = \sum_{i=1}^{p} a_x(i)x(k-i) + G_x u(k) \tag{2.8}$$

where $a_x(i)$ are the coefficients of the all-pole filter determined by linear prediction techniques, G_x is the filter gain and $u(k)$ is a unit variance ($\sigma_u^2 = 1$) white noise excitation. For voice coding purposes (LPC Vocoder) the all-pole filter is interpreted as a model of the vocal tract that filters the signal produced by the glottis. This excitation signal $u(k)$ can either be white noise, a periodic excitation with a certain pitch or a mixture of both. In general, the filter coefficients are determined by linear prediction with the side condition

that $u(k)$ has a maximum flat (=white) spectrum. As most of the speech information is contained in the filter coefficients $a_x(i)$ and the filter gain G_x, these parameters are used to determine several distance or distortion measures (see below). The parameters are usually calculated for signal frames of $10 - 30$ ms duration.

2.3.1. Log-Likelihood Ratio Measure

The log-likelihood ratio (LLR) is defined as:

$$d_{LLR}(\boldsymbol{a}_x, \boldsymbol{a}_d, n) = \log \frac{\boldsymbol{a}_d^T(n)\boldsymbol{R}_{xx}(n)\boldsymbol{a}_d(n)}{\boldsymbol{a}_x^T(n)\boldsymbol{R}_{xx}(n)\boldsymbol{a}_x(n)} \tag{2.9}$$

where $\boldsymbol{a}_x^T = [1, -\alpha_x(1), -\alpha_x(2), \ldots, -\alpha_x(p)]$ are the LPC coefficients of the clean signal in segment n, $\boldsymbol{a}_d^T = [1, -\alpha_d(1), -\alpha_d(2), \ldots, -\alpha_d(p)]$ are the LPC coefficients of the the enhanced signal in segment n, and $\boldsymbol{R}_{xx}$ denotes the squared auto-correlation matrix of dimension $(p+1) \times (p+1)$ of the signal segment for which the optimal predictor coefficients $\boldsymbol{a}_x$ have been computed.

The nominator of (2.9) can be interpreted as the energy of the prediction residual that remains when the clean speech signal is filtered with the linear predictive coefficients calculated for the enhanced speech. This measure shows how well the current observed signal x is represented by the speech model coefficients (i.e., the most important components) of the enhanced signal. The denominator is the prediction residual of the clean speech filtered with its optimal LPC coefficients. Thus, the denominator is always smaller than the nominator and the LLR measure is always positive.

2.3.2. Itakura Saito Distortion

The Itakura-Saito Distortion measure has another weighting of the LPC-coefficients and is defined by:

$$d_{IS}(\boldsymbol{a}_x, \boldsymbol{a}_d, n) = \frac{G_x^2(n)}{G_d^2(n)} \frac{\boldsymbol{a}_d^T(n)\boldsymbol{R}_{xx}(n)\boldsymbol{a}_d(n)}{\boldsymbol{a}_x^T(n)\boldsymbol{R}_{xx}(n)\boldsymbol{a}_x(n)} + \log\left(\frac{G_d^2(n)}{G_x^2(n)}\right) - 1 \tag{2.10}$$

where G_x and G_d are the all-pole gains of the clean and enhanced signals, respectively. The all-pole gains G_x, G_d can be computed as follows:

$$G_x = \left(\boldsymbol{r}_x^T \boldsymbol{a}_x\right)^{1/2}, \quad G_d = \left(\boldsymbol{r}_d^T \boldsymbol{a}_d\right)^{1/2}, \tag{2.11}$$

where $\boldsymbol{r}_x^T = [r_x(0), r_x(1), \ldots, r_x(p)]$ contains the autocorrelations of the clean signal. The same accordingly applies to the enhanced signal. The Itakura-Saito Distortion measure penalizes the differences in the all-pole filter gains, i.e., differences in the overall spectral levels of clean and enhanced signals [45]. Usually, the Itakura-Saito Distortion is averaged over all signal frames n. Another definition of the Itakura-Saito distortion is given in [77]:

$$d(\boldsymbol{a}_x, \boldsymbol{a}_d, n) = \frac{(\boldsymbol{a}_x(n) - \boldsymbol{a}_d(n))^T \boldsymbol{R}_{xx}(n)(\boldsymbol{a}_x(n) - \boldsymbol{a}_d(n))}{\boldsymbol{a}_x(n)\boldsymbol{R}_{xx}(n)\boldsymbol{a}_x(n)}. \tag{2.12}$$

For the LPC measures various variants and weightings exist in literature. Quackenbush et al. [59], e.g., evaluated over thirty different measures of that kind and identified the Log-Area-Ratio (LAR) measure (see below) to have high correlation with subjective data. For the Itakura-Saito Distortion used in this study, the definition due to (2.10) and [29] was used.

2.3.3. Log-Area-Ratio Measure

The Log-Area-Ratio (LAR) Measure is a performance measure that had shown high correlation with subjective quality measurement in [50, 59]. It is frequently used for evaluation of noise reduction systems [4, 23, 29]. The distance measure is based on the tube model of speech production (see [78]). First, the reflection coefficients also known as partial correlation (PARCOR)-coefficients are calculated based on the Levinson-Durbon recursion from the all-pole filter coefficients [36] of a signal frame l. Second, the area-ratio coefficients are calculated by

$$g(p) = \frac{1 + k(p)}{1 - k(p)} \qquad p \in \{1, \ldots, P\}. \tag{2.13}$$

Finally, the LAR is calculated for each frame n as the log-ratio of the coefficients g_x for clean speech and g_d for the distorted (processed) speech.

$$LAR(n) = 20 \cdot \sum_{p=1}^{P} \log_{10} \left| \frac{g_x(p,n)}{g_d(p,n)} \right|. \tag{2.14}$$

2.4. Perceptually Motivated Objective Measures

2.4.1. Weighted Spectral Slope (WSS) Distance Measure

In a psychoacoustical study by Klatt [40] about the distortion of synthetic vowels it was found that subjects assigned the largest perceptual distance to vowels with changes in spectral peak (formant) locations while ignoring other spectral changes such as spectral tilt or overall level. The weighted spectral slope (WSS) measure was developed to emphasize these changes which is realized by following processing steps: First, the spectral slopes of each critical band spectra (see A.2) are calculated by

$$S_x(b) = C_x(b+1) - C_x(b) \tag{2.15}$$

$$S_d(b) = C_d(b+1) - C_d(b), \tag{2.16}$$

where $C_x(b)$ and $C_d(b)$ denote the critical band spectra in dB and $S_x(b)$ and $S_d(b)$ are the spectral slopes for the clean speech and the processed speech for band b, respectively. For the calculation of critical band spectra based on STFT-spectra compare (2.7) and section A.2. Then a weight is calculated that emphasis spectral slopes that are near to a spectral peak or valley:

$$W(b) = \frac{K_{\max}}{K_{\max} + C_{\max} - C_x(b)} \frac{K_{loc\ max}}{K_{loc\ \max} + C_{loc\ \max} - C_x(b)} \tag{2.17}$$

where $C_{\max}$ is the largest log-spectral magnitude among all bands, $C_{loc\ \max}$ is the value of the peak nearest to band b, and $K_{\max}, K_{loc\ \max}$ are the constants which can be adjusted using regression analysis to maximize the correlation between the subjective listening tests and values of the objective measure. The experiments in [40] showed a high correlation for $K_{\max} = K_{loc\max} = 1$. Finally, the WSS measure is calculated for each frame of speech as:

$$d_{WSS}(C_x, C_d) = \sum_{b=1}^{B} W(b)(S_x(b) - S_d(b))^2 \tag{2.18}$$

where B is the number of critical bands used.

2.4.2. PEMO-Q

The preceding objective measures contained only limited knowledge about the perceptual aspects of the auditory system. The LPC-based measures basically include a model of speech production instead of a speech perception model. The WSS measure analyzes the signal in critical bands but the measure only focuses on the distortion of spectral peaks and thus it does not cover all effects influencing the perceptual quality. Exploring the potential of more detailed auditory models for improving quality estimation is therefore highly desirable.

Objective measures based on a psychoacoustic model were developed by Huber [33, 34] and summarized under the name "PEMO-Q". Primarily, PEMO-Q estimates a perceptual correlation between a reference and a test signal and additionally calculates some deduced measures that include some cognitive effects and allow a direct mapping of objective data to subjective rating scales, e.g., the objective differential grade (ODG) versus subjective differential grade (SDG). However, in the following evaluation of noise reduction schemes only the basis of the deduced measures, namely, the perceptual similarity measure PSM, is used from the PEMO-Q model.

PEMO-Q is an advancement of the speech quality measure q_C by Hansen and Kollmeier [30] who successfully applied their method to predict subjectively rated speech transmission qualities of mainly low-bit rate speech codecs. The idea of this measure is to transform the original and the distorted signal into an internal representation (which is thought of as the information that is accessible to higher neural stages of perception) and compare them in this "auditory domain". This transformation is based on the quantitative model of the "effective" signal processing in the auditory system by Dau et al. [12**?**], excluding its final detector stage and is shown in the right panel of Fig. 2.2.

The left panel of Fig. 2.2 shows the block diagram of the quality estimation which in general is also valid for other perceptual model based measures. In a first step the signals should be delay-compensated and long-term level aligned. The level alignment depends on the experiments that are made: Regarding noise reduction schemes, the block "lossy processing" in Fig. 2.2 which connects the reference and the test signal includes both, the interference of clean speech with additive noise and the enhancement by a noise reduction scheme. Thus, a level alignment between the clean speech and the more energetic noisy mixture would be biased by the absolute noise level in the test-signal. In cases were the residual noise level is relevant the level alignment should be omitted or calculated only

based on the level of the desired speech component in both, reference and test signal. Subsequently, the signals are transformed by an auditory model into *internal representations* which is shown in more detail on the right panel of Fig. 2.2. In the first stage of the auditory processing the signals are filtered by a gammatone-filterbank [32, 56] which consists of overlapping 4th- order bandpass filters with critical bandwidth (see section A.2). A half-wave rectification and lowpass-filter of 1 kHz simulates the basic characteristic of the inner hair cells (IHC) that only emit an action potential if bended to one direction [30]. For signals below 1 kHz a phase coding in the primary auditory fibers can be observed whereas for higher frequencies primarily the envelope of the auditory filters is used. The next stage accounts for temporal adaptation and dynamic compression. It consists of five cascaded feedback loops that have lowpass time constants in the feedback path ranging from 5 to 500 ms. The input of the first feedback loop is limited to a lower bound accounting for the absolute threshold of hearing that is assumed to be 100 dB below the maximum input level. The adaptation loops have been introduced by Püschel [58] to quantitatively model the temporal masking effects. The output of the adaptation loops is filtered by an 8 Hz lowpass filter according to [**?**] which reduces amplitude modulations and was found by detection and masking experiments for temporal integration. Alternatively, a modulation filterbank [12] accounting for modulation frequencies up to about 160 Hz can be applied. The modulation filterbank mode allows a more accurate estimation of perceptual quality, particularly if the perceptual differences between the clean speech reference and the processed signal are small. Here, the simpler lowpass version was used as the differences between reference and the processed test signal are relatively large compared to, e.g., high quality speech codecs. Returning to the block diagram on the left panel, the internal representations of reference and test signal may be biased to account for further perceptual effects (like, e.g., "Beerends-Berger-assimilation", see [33]). Finally, a cross correlation coefficient is calculated from the resulting signals and denoted as a perceptual similarity measure (PSM). For values of PSM very close to 1 the differences between reference and test signal are inaudible whereas lower values indicate an audible difference. Thus by choosing an appropriate reference signal (generally: the clean speech signal) any PSM value lower than 1 can be interpreted as a *quality reduction*. The selection of the reference and test signals is discussed below.

2.4.3. PESQ

The evaluation of speech quality measure (PESQ) described in [60] was selected as the ITU-T recommendation P.862 in a competition between different objective measures for speech quality estimation across a wide range of codec and network conditions. The general idea of PESQ is similar to PEMO-Q in that a test signal and a reference signal are transformed by means of an auditory processing model and a distance measure is calculated from a comparison of the signal representations. Compared to PEMO-Q (or its predecessor q_C) the preprocessing attributed to the auditory model is much simpler and has a more refined distance calculation [31] specialized on telecommunication aspects such as limited speech bandwidth, packet loss and coding artifacts. There are diverse parameters in PESQ that are trained based on the subjective ratings of a large speech database. Thus PESQ is an effectively trained objective quality measure that includes single aspects of the auditory

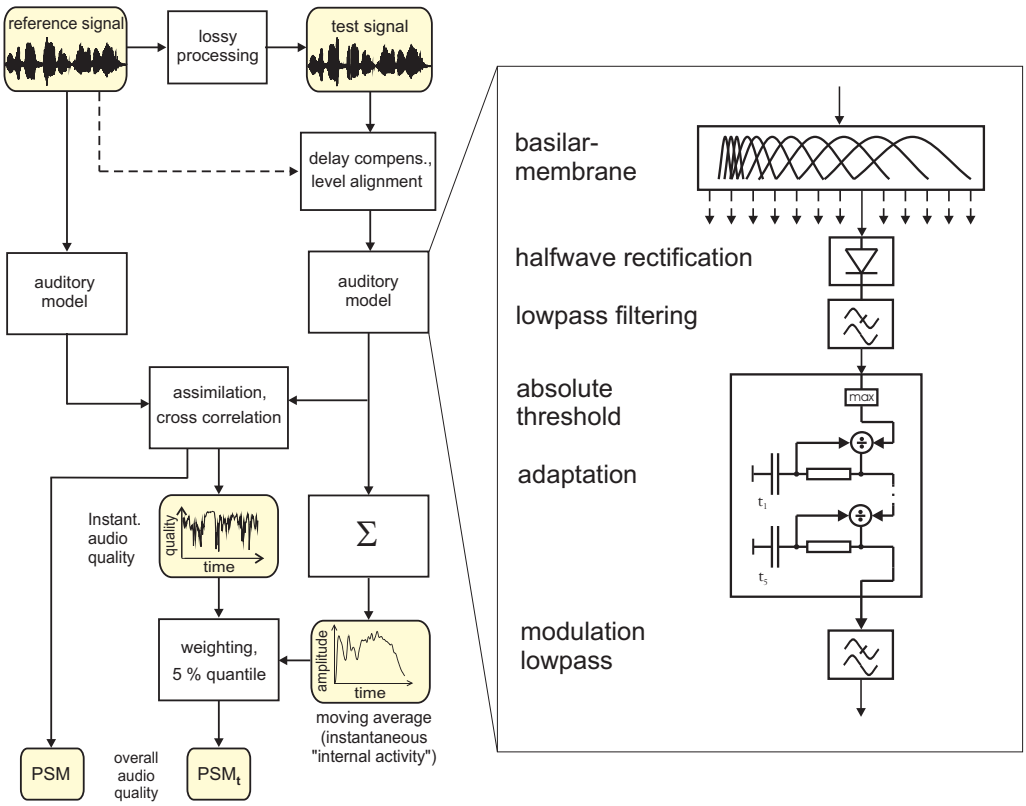


Figure 2.2.: The processing scheme of PEMO-Q [34]

system but does not aim to transform the signals into an internal representation in the sense that it models the information accessible to the neural system.
Several transformation steps are applied to the signal: First, the Bark-band spectra based on fast Fourier transform (FFT) are calculated for both, reference and test signal (see A.2). In a frequency equalization step the long-term bark spectrum of the test signal is applied to the reference signal to override spectral differences that are system-related (e.g., reduced bandwidth) and thus have minor effect on the interesting perceived quality criteria. Then, a short-term (broadband) gain equalization between the "audible power" of reference and test signal spectrum is applied, i.e., the gain is only calculated based on the importance-weighted frequency bands. Afterwards, the bark spectra are transformed to a sone loudness scale [72, 86] with a compression factor. Finally, the difference between original and degraded loudness spectrum is calculated. Different weights are applied to negative and positive values, because their influence on the perceived quality is different. Negative differences are perceived as new signal components in the test signal which are perceptually more salient than missing components (positive differences). The weighted differences are summed to a single time varying quality measure.

2.4.4. PEAQ

The Perceptual Evaluation of Audio Quality (PEAQ) measure has been defined as the new ITU-R recommendation BS-1387 to fulfill the requirements of a more general quality measure that is not restricted to narrow-band speech signals. Similarly to PESQ, it is a composite and expansion of the best elements from different quality measures and characterized by a high degree of optimization and adaptation to a single task [34]. The main purpose of PEAQ is the prediction of subjective quality ratings of low-bit-rate coded audio signals.

2.4.5. BSIM

The preceding objective physical and psychoacoustical-model-based performance measures are single channel (monaural) measures that do not adequately take into account the respective benefit of binaural cues in the output signal that can be utilized by the human user of the binaural processing scheme. The binaural speech intelligibility model developed by Beutelmann and Brand [3] tries to estimate the speech intelligibility from binaural signals. It combines two established models, the binaural equalization-cancellation (EC) processing by Durlach [19] with the monaural speech intelligibility index (SII, [1]). For the calculation of the measure the binaural signals of the speech and the binaural noise signals are needed separately as input to the model. In a first stage the signals are transformed by a gammatone filterbank [32, 54]. To account for individual hearing thresholds, uncorrelated gaussian noise which is spectrally shaped by the individual pure tone audiogram for left and right ear is added to the signals. In the EC stage, the SNR of each auditory frequency channel is maximized by first applying an optimum relative delay and amplitude gain between left and right ear channel (equalization step) and second subtracting the right from the left channel (cancellation step). The optimum delay and gain which maximizes the SNR can be found by a minimum search (see [3]). If the maximized (binaural) SNR is lower than the SNR observed at a single ear, this higher SNR is chosen. This takes into account the *better ear* effect which is the ability (of a normal hearing subject) to focus on the sound at the ear nearest to the signal of interest and ignoring the noise or disturbance at the averted side. After the model parameters that a lead to an optimal SNR are found, the auditory filterbank signal can be resynthesized according to [32] and analyzed by the speech intelligibility index (SII) model according to [1].
The EC stage of BSIM looks quite similar to a beamformer or a multi-channel wiener filter. However, the most apparent differences are that the model (i) has a priori knowledge about all four input signal components (desired speech left/right, undesired noise left/right), (ii) knows of the better-ear effect, (iii) uses an auditory filter bank and (iv) is inserting additional *internal noise* to account for restrictions within the individual hearing ability. For a detailed description of the model and its artificial processing errors see [3].
For the assessment of noise reduction schemes the performance improvement achieved by the algorithm is of major importance. Thus, the difference between the binaural speech intelligibility measure (BSIM) at the output of the scheme and the input, namely the ΔBSIM needs to be calculated:

$$\Delta\text{BSIM} = \text{BSIM}_{out}(s_L, S_s, n_L, n_R, \text{params}) - \text{BSIM}_{in}(s_L, s_R, n_L, n_R, \text{params}). \quad (2.19)$$

2.5. Reference and Test Signal

An important aspect for all quality measures presented here is the type of reference and test signals that are used. For objective noise reduction and speech coding assessment in most cases the unprocessed speech signal, i.e. the *clean speech*, will be used as a *reference* signal. This reference is compared against the *test* signal – usually the processed *noisy speech* or *coded speech* signal. The aim is to mimic the assessments done by subjects who often have an idea of the *desired* (reference) signal which in many cases may be the clean speech[3]. However, if the system under test is producing output signals that are perceptually far-off the optimum clean speech signal it is reasonable to use a different reference signal. In this case it is suggested to use a signal that was processed by an optimal (coding or noise reduction) system which has access to a priori information that in practise is unknown to the real-world system. A possible application scenario is the assessment of low-rate speech codecs and noise reduction schemes that do not aim to produce an output signal similar to the unprocessed clean speech. Some objective measures, e.g., PESQ (section 2.4.3) pursue this strategy by adjusting the reference signal to account for system-related long-term spectral differences between reference and test signal.
For the application of noise reduction systems it may also be interesting to have a *perceptual speech distortion measure* that is more or less independent of concurrent *noise reduction* quality aspects. This measure should also account for masking effects of the residual noise in the processed noisy speech signal. Thus, a *noisy reference* signal is suggested here that has the same signal-to-noise ratio as the processed output signal. The correlation of different objective measures based on auditory models with subjective data using the noisy reference is proven in section 3.2. By measuring the distortion of the speech component alone, the beneficial masking effects of the residual noise that may allow for a higher acceptable speech distortion would be ignored.
Besides the definition of applicable reference and test signals it is of major importance to consider realistic signal conditions, ideally, real-world recorded signals in acoustical scenarios for which the algorithms under test are intended. Thus in the following studies, these factors influencing the objective quality estimation and performance of noise reduction schemes have particularly been considered.

2.6. Discussion

The preceding list of objective measures (except the newer auditory model based measures) has frequently been used in the research field of speech coding and speech enhancement. The extensive studies by Quackenbush et al. [59], identified measures that were highly correlated with subjective data gained from assessments of many different artificial signal distortions, including additive noise. However, no signals processed by noise reduction algorithms were included in the list of signals, so that the typical artifacts applied by, e.g., spectral subtraction, were not evaluated. In a study by Hansen and Pellom [29] different single-channel noise reduction algorithms were evaluated with objective measures (ISD, LLR, LA, segSNR, WSS, see above). Novel objective measuring methods where suggested

[3] In case subjects have a stable internal reference, *direct* comparison tests are possibly not needed.

as a standard criteria for algorithm performance comparison, including audio test-files that are applicable for noise reduction, but no subjective listening tests were carried out to identify the most significant measures in that study. Moreover, no measures based on sophisticated auditory models were available at that time. Because only little knowledge on the significance of the abovesaid measures for subjective benefit is available by now (except for the studies done by Marzinzik [50], see below), all measures[4] are compared against subjective data in the study described in section 3.2. *Subjective* assessment methods for noise reduction schemes, on the other hand, have been studied quite well. It is known, that the assessment of subjective *overall* quality of noise reduction schemes is a complex interaction of diverse (quality) dimensions such as *speech distortion*, *noise reduction*, *ease of listening*, *speech intelligibility* and others. Hence, the ITU-T Standard P.835 [35] recommends to rate different concurrent aspects at the same time in subjective listening tests. This subjective test method was carried out to identify the most significant objective performance measures from the preceeding list (see section 3.2). As multiple dimensions of signal quality are evaluated simultaneously, it is expected that different measures can have high correlations with subjective data, depending on the quality parameter that is addressed.

To conclude this chapter, an *ultimate* selection of applicable measures cannot be given at this point. However, the following chapters will show which measures are covering the interesting aspects of signal quality and hence, are predestinated to form a test-bench for objective quality assessment.

[4] excluding the binaural measure BSIM which is used for binaural algorithms in chapter 4. The measures WSSD and PEAQ have been calculated but are not included in table 3.1.

3. Monaural Noise Reduction Schemes - Performance Evaluation and Optimization

3.1. Introduction

Objective measures for the assessment of different aspects of signal quality have been summarized in chapter 2. As quite a few measures have their origin in the speech coding area and have been designed for detecting and evaluating signal degradations that typically occur in these applications, it is unclear if the measures are also appropriate for the assessment of noise reduction systems. Thus, in section 3.2, single channel noise reduction schemes based on short time spectral attenuation (STSA) are evaluated by objective measures and subjective assessment. The subjective listening tests have been performed according to a test methodology by the ITU-T [35] that recommends to assess the quality of noise reduction systems with three absolute categorial rating (ACR) scales in parallel, each for one dimension of signal quality. These subjective data are compared to objective evaluation data estimated by all measures summarized in chapter 2 (except for the binaural measure). With a correlation analysis between objective and subjective data the measures that are appropriate to predict the mean opinion score (MOS) of normal hearing subjects will be identified. Compared to similar studies by Marzinzik [50], who evaluated listening effort and speech intelligibility for normal hearing and hearing impaired subjects, the current study evaluates a broader range of output signal qualities and noise types processed by different single-channel STSA algorithms. The significant objective measures are then combined to a test-bench that is useful for the development and parameter optimization of the algorithms in the following sections.

In section 3.3 a noise reduction scheme is proposed that replaces the STFT of a noise reduction algorithm by an auditory filterbank. It has a high dimensional parameter space and therefore poses a problem for the optimization based on a single objective perceptual measure. Thus, as a possible solution a sub-band quality measure is proposed that can reduce the degrees of freedom for optimization.

Finally, section 3.4 suggests to use the objective measures for a combined audio quality assessment and a low-power optimization that is desirable for hearing aids.

3.2. Comparison of Objective and Subjective Data for the Assessment of Single Channel Noise Reduction Algorithms[1]

The short time spectral attenuation (STSA)-algorithms contain parameters that are affecting the amount of noise reduction. However, maximizing the attenuation of noise with these parameters in general leads to a distortion of the desired signal which can only be tolerated to a certain amount.

In subjective quality assessment tests of noise reduction schemes subjects often have difficulties in rating the overall quality which seems to be a trade-off between the amount of background noise removal and speech distortion. Another point is that background noise can even become more annoying if processed by a suppression algorithm, e.g., by introducing musical tones or amplitude fluctuations.
The same difficulty exists for predicting the overall quality with objective measures. While it should be feasible to quantify the amount of noise reduction or to measure speech distortion separately, the prediction of the overall quality seems to be more complex. The idea is to find objective perceptual measures that have a high correlation with the results from subjective ratings. This measures can then be used as a test-bench for evaluation and parameter optimization in noise reduction schemes. The results of this section have been presented in [62].

3.2.1. Algorithms

STSA algorithms according to Ephraim and Malah's weighting rules [22] where employed as single channel state-of-the-art algorithms. These algorithms are characterized by a strong reduction of noise while introducing only little of the well known *musical tones* or *musical noise* that result from subtracting an average noise spectrum from a non-stationary frame-based spectral estimate. A detailed description of the involved filter parameters can be found in Cappé [9]. The most important parameters are two signal-to-noise ratio (SNR) estimates: An instantaneously estimated (a posteriori) SNR and an a priori SNR estimate that is calculated by a recursive smoothing of preceding a posteriori values. The considered algorithms need a reliable noise power estimation. Here, the minimum statistics method (MinStat) by Martin [48] and a Voice Activity Detection (VAD) algorithm by Marzinzik [50] are used.

3.2.2. Signals

The speech signals used here were taken from the Oldenburg Logatome Speech Corpus (OLLO) [83] and consisted of six sentences spoken by german male and female speakers. The noise signals were speech-shaped noise, cafeteria noise, icra7 noise (speech like modulated noise) and white gaussian noise. All signals had an approximate duration of 20 seconds and a sampling rate of 16 kHz. In the simulation system the signals were mixed at a SNR of 0 dB and 5 dB. The calculation of the time-variant filter was made on this

[1] Parts of this section have been published as "Objective Perceptual Quality Measures for the Evaluation of Noise Reduction Schemes" in proceedings of 9th International Workshop on Acoustic Echo and Noise Control (IWAENC), 2005 [62]

mixture while the filtering process was also done on the separate speech and noise signals for subsequent quality assessment and the calculation of the SNRE and other quality measures.

3.2.3. Objective Measures

For the objective quality estimation the following measures are calculated that are described in chapter 2. In addition to the signal-to-noise ratio measures SNRE, the segmental SNR (segSNR), and frequency weighted SNR (based on equation (2.7)) a linear coherence coefficient is calculated between the time signal of the processed output and the clean speech. These performance measures can be considered as *technically* motivated measures that primarily aim to predict the amount of noise reduction achieved by the system.

The measures based on linear predictive coding (LPC), Log-Area Ratio (LAR), Log-Likelihood Ratio (LLR), and Itakura Saito Distance (ISD), are motivated by a speech production model (see section 2.3) and thus primarily aim to predict the quality of speech or the overall quality. These measures are calculated for segments of 30 ms duration and the short-time estimates are averaged over time.

From the perceptually motivated measures, PSM (PEMO-Q) and PESQ, different versions are derived. First, the perceptual similarity measure (PSM) is calculated between the processed output and the clean speech (see section 2.4.2). Second, PSM is also calculated for the the noisy input and the clean speech (PSM_{in}). The difference between input and output PSM is referred to as ΔPSM. It shows the increase in perceptual similarity between input and output signal that is achieved by the noise reduction scheme. Positive ΔPSM values predict a higher quality of the processed signal compared to the unprocessed signal, whereas negative values indicate a signal degradation. A parameter that influences the weighting ("Beerends-Berger" option) is optionally switched off in the measure (PSM_b), (as this option is primarily intended for the prediction of moderate coding artifacts). Similarly, for the perceptual speech quality measure PESQ (2.4.3), the absolute value and the relative increase, ΔPESQ is calculated.

Both psychoacoustical measures, PESQ and PSM, aim to predict the overall quality if calculated between the clean speech and the output signal. As the noise reduction schemes often lead to speech signal distortions, it is also interesting to estimate the *perceptual* speech distortion in the presence of residual noise using these perceptual measures. To achieve this, it is suggested here, to use a noisy reference signal with the same signal-to-noise ratio as the output signal. From the *technically* viewpoint the SNR enhancement between noise reference and output then is zero. But for the perceptual measures, the quality reduction applied to the target signal may pop out. In the following, these measures using the noisy reference are referred to as SNR_PSM and SNR_PESQ.

A comparison analysis between the objective predictions and the subjective data will identify the significance of these measures.

3.2.4. Experiments

Signal Processing and Objective Quality Assessment

The recursive smoothing parameter τ for the a priori SNR-estimate in Ephraim and Malah's algorithms has great influence on the noise reduction strength. In order to find an optimal setting and to cover a broad range of qualities for a subsequent correlation analysis of objective and subjective measures, τ was varied in the range from $0 - 800$ms. All signals were processed with the noise estimators Minstat and VAD, respectively. For each setting, the above mentioned quality measures were calculated for a number of speech signals mixed with different types of noise (see Section 3.2.2). For subjective listening tests a subset of 7 time-constants per noise type, algorithm and input-SNR was chosen.

Subjective Listening Tests and Quality Assessment

The subjective listening tests were done according to the ITU-T Recommendation P.835 [35] which describes a methodology for evaluating the subjective quality of speech in noise and is particularly appropriate for the evaluation of noise suppression algorithms. The methodology uses separate absolute categorial rating scales (ACR) to independently estimate the subjective quality of the speech signal alone, the background noise alone and the overall quality. 16 normal hearing subjects were tested. The whole test consisted of 8 sessions with 15 trials each and took approximately 1 hour. One trial was composed of three sub-samples. Each sub-sample consisted of two sentences, male and female talkers, of 3.25 seconds duration each. In the first sub-sample the subjects were instructed to attend only to the background noise and rate it on a five category scale from "1 - sehr störend" (very disturbing) to "5 - gerade wahrnehmbar" (just noticeable). In the second sub-sample subjects were instructed to attend only to the speech signal and rate it on a scale from "1 - sehr stark verzerrt" (very much distorted) to "5 - unverzerrt" (not distorted). In the third sub-sample subjects were instructed to listen to the speech + background and rate it on a five category overall quality scale from "1 - schlecht" (bad) to "5 - ausgezeichnet" (excellent). The ratings were done with an ACR - software using sliders that allowed a sub-categorial rating in 0.1 steps.

3.2.5. Results

Fig. 3.2 shows the subjective data in the left panels for both noise estimators and two of the four noise types. Initially, it can be stated that all subjective tests - independent of noise type, input-SNR or noise estimators - show consistent behavior in the way that the perceived speech-signal qualities (red dotted line) decrease and the amount of perceived noise reduction (green dashed line) increase monotonically by increasing the smoothing constant τ. As stated before, the overall quality ratings (black solid line) seem to be a trade-off between both rating tasks. Obviously the subjects prefer in virtually all cases a certain amount of smoothing. Another point is that the two noise estimators, MinStat and VAD, show different performance for fluctuating noise, e.g., speech-modulated icra7-noise, but similar behavior for stationary noise while the mean opinion score (MOS) for the overall quality is almost the same. The subjects reported that - especially in cases

of fluctuating noise - they were uncertain what to prefer - reducing noise and accepting signal-distortion or the opposite. This may be the reason why there was no preference for one of the noise estimators observable although the outputs were very different.

To find out which objective measure describes the respective quality rating tasks best, the correlation between different objective measures and the subjective data were evaluated (see Tab. 3.1). The highest correlations of all objective measures are indicated with bold black numbers, the highest negative correlations are printed in red. The first four columns show the correlation for each noise type separately. The last column contains the overall correlation for all signal types, algorithms, and input-SNR's. Rows 1-7 show more technically measures, i.e. these measures are not based on a complex psychoacoustic model. Rows 8-12 contain the perceptual measures and their relative enhancement representations (ΔPSM, ΔPESQ) all with clean speech reference. The last rows contain the perceptual measures but with an output-SNR-aligned noisy reference signal, indicated by the prefix "SNR_".

As for the background noise rating, the highest correlations are gained by the SNRE. This means that SNRE is a good measure to rate the amount of noise reduction by an algorithm, independent of the speech signal quality. Also, high correlation values are gained by the ΔPSM measure if different types of background noise are considered separately. The correlation for ΔPSM with the subjective data can be seen in Fig. 3.1. The functional relationship between subjective and objective measures varies across different types of background noise, hence the overall correlation is less. As a consequence the objective measure should incorporate some noise dependent scaling to better model the subjective data.

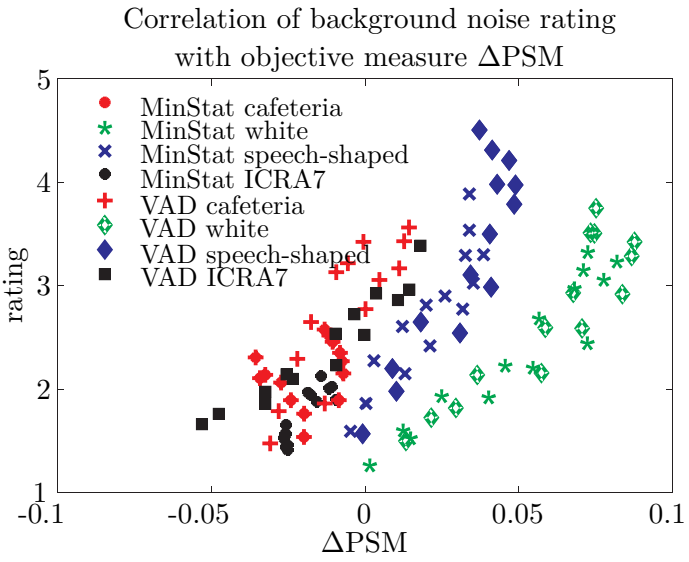


Figure 3.1.: Noise dependent correlation between objective and subjective data for different noise reduction algorithms. ACR data for normal hearing listeners are plottet over the objective measure ΔPSM.

In terms of speech-signal rating most of the correlations are negative. The strongest anticorrelated measure is the frequency weighted SNRE, which may result from the fact that noise reduction and speech distortion are competing processes in the considered algorithms. The strongest correlations are achieved for the perceptual measures with the noisy reference, especially SNR_PESQ, as expected. The best correlation in terms of overall quality rating show the perceptual measures with clean speech reference, especially PESQ and PSM_b.
The right panels in Fig. 3.2 show the prediction of the subjective data on the left panels by the objective measures that had the highest correlations for each rating task, i.e., SNRE for the prediction of perceived noise reduction, SNR_PESQ for the speech-signal degradation and PSM_b for the prediction of the overall quality. The curves have been linearly fitted to match the scaling of the MOS.

Correlation with background noise rating	Cafeteria noise	White noise	Speech-shaped noise	ICRA7 noise	Overall-Correlation
SNRE	**0.93**	0.91	0.88	**0.90**	**0.75**
Coherence	0.50	0.67	0.58	0.68	0.53
seg. SNRE	0.71	0.62	0.63	0.84	0.54
freq. wt. SNRE	0.70	0.79	0.54	0.66	0.49
mean LAR	0.33	-0.89	0.18	0.35	-0.06
mean LLR	-0.08	-0.73	0.06	0.05	-0.14
mean ISD	0.55	-0.51	0.54	0.67	0.20
PSM	0.57	0.89	0.84	0.70	0.69
PSM_b	0.52	0.66	0.70	0.69	0.60
PESQ	0.37	0.66	0.64	0.62	0.63
ΔPSM	0.76	**0.92**	**0.89**	0.83	0.62
ΔPESQ	0.42	0.81	0.64	0.84	0.56
SNR_PSM	-0.56	-0.49	-0.39	-0.58	-0.28
SNR_PESQ	-0.60	-0.81	-0.58	-0.53	-0.41

Correlation with speech signal rating	Cafeteria noise	White noise	Speech-shaped noise	ICRA7 noise	Overall-Correlation
SNRE	-0.67	-0.77	-0.94	-0.87	-0.67
Coherence	0.27	0.02	-0.21	-0.04	-0.05
seg. SNRE	-0.06	0.09	-0.32	-0.46	-0.17
freq. wt. SNRE	-0.90	-0.89	-0.79	-0.93	-0.70
mean LAR	-0.88	0.33	-0.46	-0.67	-0.06
mean LLR	-0.66	0.01	-0.41	-0.72	-0.22
mean ISD	-0.79	-0.13	-0.63	-0.89	-0.62
PSM	0.22	-0.31	-0.58	-0.07	-0.15
PSM_b	0.25	0.07	-0.38	-0.06	-0.02
PESQ	0.41	0.06	-0.33	0.05	-0.01
ΔPSM	-0.05	-0.75	-0.90	-0.49	-0.39
ΔPESQ	0.34	-0.27	-0.73	-0.52	-0.23
SNR_PSM	0.84	0.76	0.67	**0.87**	0.61
SNR_PESQ	**0.87**	**0.92**	**0.86**	**0.87**	**0.74**

Correlation with overall quality rating	Cafeteria noise	White noise	Speech-shaped noise	ICRA7 noise	Overall-Correlation
SNRE	0.35	0.66	0.41	0.29	0.35
Coherence	0.83	0.88	0.93	0.93	0.65
seg. SNRE	0.74	0.89	0.89	0.71	0.53
freq. wt. SNRE	-0.17	0.40	-0.05	-0.11	0.00
mean LAR	-0.46	-0.90	-0.45	-0.28	-0.07
mean LLR	-0.75	-0.94	-0.61	-0.68	-0.43
mean ISD	-0.24	-0.79	-0.04	-0.16	-0.34
PSM	0.82	**0.92**	0.93	0.87	0.70
PSM_b	0.85	0.91	**0.94**	0.93	0.76
PESQ	0.85	**0.92**	**0.94**	**0.94**	**0.81**
ΔPSM	0.71	0.69	0.58	0.54	0.39
ΔPESQ	**0.86**	**0.92**	0.48	0.65	0.47
SNR_PSM	0.12	-0.04	0.09	0.01	0.04
SNR_PESQ	0.09	-0.36	0.05	0.07	0.00

Table 3.1.: Correlation between objective and subjective measures for the three rating tasks (i.e., background noise, speech signal, and overall quality) and the types of background noises.

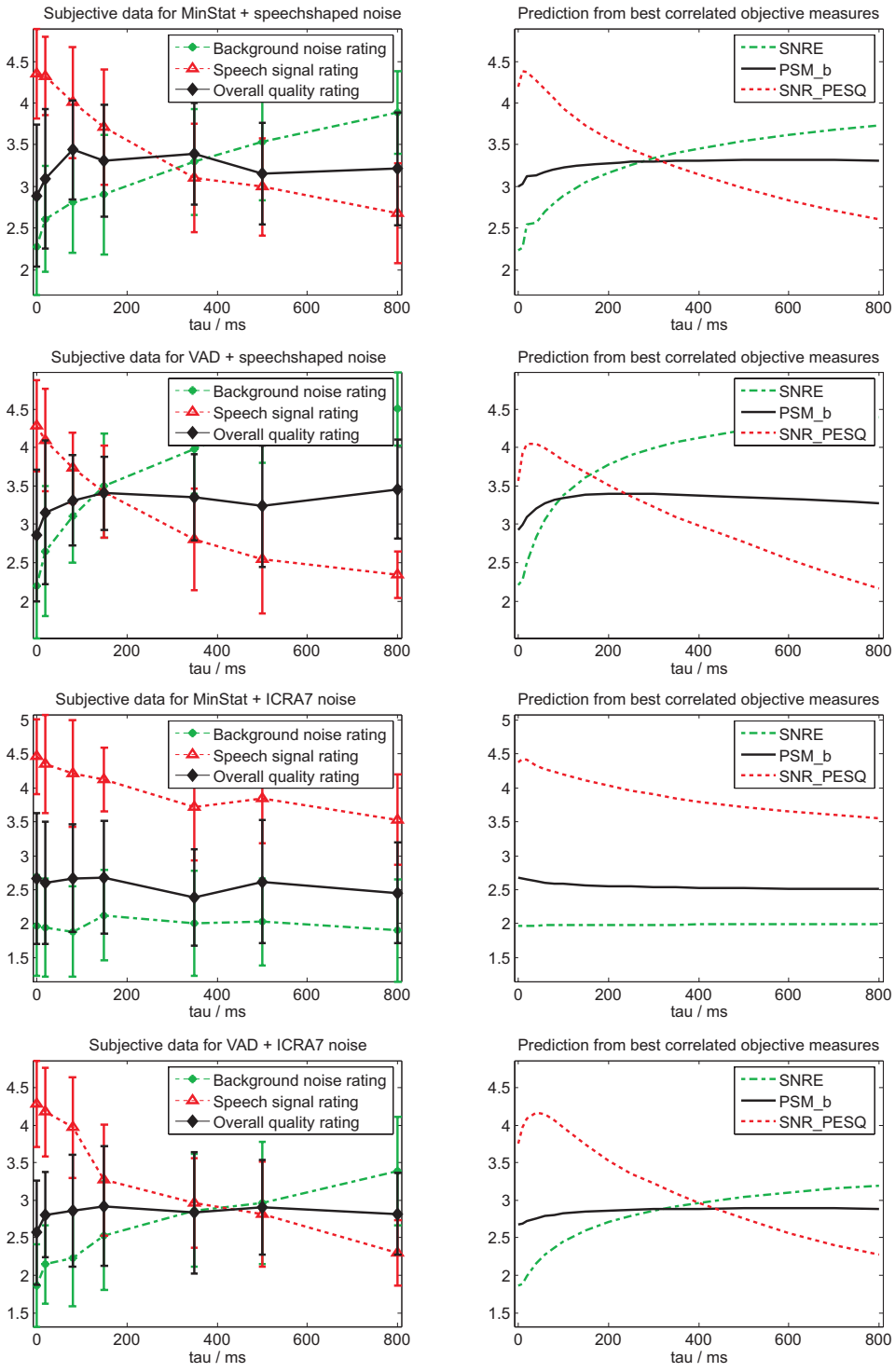


Figure 3.2.: Subjective (left panel) and objective (right panel) data for different noise types and algorithms

3.3. Subband-based Parameter Optimization in Noise Reduction Schemes by means of Objective Perceptual Quality Measures[2]

In general, noise reduction schemes for application in hearing-aids or car environments have parameters that are determined by technical distance measures or heuristically based on informal listening by the algorithm developers. In section 3.2it was shown that quality measures based on psychoacoustic models are better suited to optimize single parameters in terms of the best subjective overall quality than pure technical measures like, e.g., the signal-to-noise ratio. In other words, a test-bench based on objective quality measures and several typical noise types can support the search for the best-sounding noise reduction algorithms and their internal parameter settings. However, if the algorithms become more complex, e.g., because of frequency-dependent parameters, a single broadband measure might not be feasible to assess optimal settings because of the high dimensionality of the parameter space. In this case a subband-based perceptual quality measure might be feasible. In this study, we exemplarily apply subband-based quality prediction to parameter optimization in a noise reduction algorithm based on auditory filters.

The aim of this study is to improve the applicability of perceptual objective measures to the systematic optimization of noise reduction algorithms. In particular, perceptual measures calculated in subbands are used to optimize a multidimensional parameter set band-wise. The technique is exemplarily applied to a monaural state-of-the-art noise reduction scheme, which was adopted to work with gammatone auditory filterbank signals instead of short-time fourier transformed (STFT-) signals. The parameterized noise reduction algorithm described in section 3.3.1 is then optimized with the perceptual subband measure which is defined in section 3.3.2. To assess the effects of noise reduction on the so called internal representations we take a look at processed speech signals mixed with stationary speech-shaped noise in section 3.3.4. The results are summarized in section 3.5.

3.3.1. Algorithm

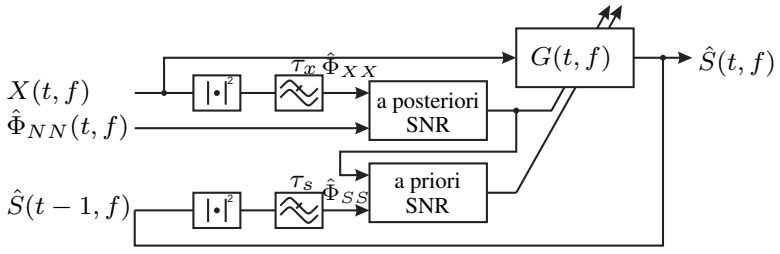


Figure 3.3.: Noise reduction scheme based on gammatone auditory filterbank

The proposed noise reduction scheme (see Fig. 3.3) is based on the idea of Ephraim and

[2]Parts of this chapter have been published as "Subband-based Parameter Optimization in Noise Reduction Schemes by means of Objective Perceptual Quality Measures", in proceeding of 10th International Workshop on Acoustic Echo and Noise Control (IWAENC), 2006 [63]

Malah's MMSE[3] log-STSA[4] [22] algorithm. Instead of the short-time fourier transform (STFT) we use a complex-valued gammatone filterbank which is supposed to have a frequency resolution similar to that of the auditory system. The gammatone filters [54] are widely used in computational auditory models for modeling the peripheral filtering in the cochlea. [32] proposes an efficient complex-valued implementation with signal resynthesis, which is used here.

Let $s(t)$ and $n(t)$ denote the speech and the noise signals, respectively. The observed signal $x(t)$ is given by

$$\hat{S}(t,f) = G(t,f) \cdot X(t,f). \tag{3.1}$$

$\hat{S}(t,f)$ can be resynthesized into a time signal $\hat{s}(t)$ with low delay by using the synthesis algorithm in [32]. $G(t,f)$ is calculated due to [9, 22] based on two SNR estimates:

$$G(t,f) = f\{\mathrm{SNR}_{\mathrm{post}}(t,f), \mathrm{SNR}_{\mathrm{prio}}(t,f)\} \tag{3.2}$$

with

$$\mathrm{SNR}_{\mathrm{post}}(t,f) = P\left[\frac{\hat{\Phi}_{XX}(t,f)}{\hat{\Phi}_{NN}(t,f)} - 1\right] \tag{3.3}$$

$$\text{with } P[x] = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{3.4}$$

$$\mathrm{SNR}_{\mathrm{prio}}(t,f) = \alpha \frac{\hat{\Phi}_{SS}(t,f)}{\hat{\Phi}_{NN}(t,f)} + (1-\alpha)\mathrm{SNR}_{\mathrm{post}}(t,f)$$

In this equations $\hat{\Phi}_{NN}, \hat{\Phi}_{XX}$ and $\hat{\Phi}_{SS}$ denote power estimates of the signals N, X and $\hat{S}$, respectively. In practice, Eq. (3.2) is precalculated and stored in a two-dimensional gain table spanned by the two SNR estimates. Eq. (3.5) is known as the *decision directed approach* [9]. The a priori SNR, $\mathrm{SNR}_{\mathrm{prio}}$, is a weighted sum of the previously estimated SNR and the instantaneous a posteriori SNR. The weighting factor α has the character of a smoothing constant with the equivalent low-pass time-constant $\tau(f) = \frac{-T_a}{\ln(\alpha(f))}$, T_a : sampling period (block period).
$\hat{\Phi}_{NN}(t,f)$ is estimated using a modified version of the minimum statistics method by Martin [49]. $\hat{\Phi}_{XX}$ and $\hat{\Phi}_{SS}$ are calculated as follows:

$$\hat{\Phi}_{SS}(t,f) = \alpha_s(f)\hat{\Phi}_{SS}(t-1,f) + (1-\alpha_s(f))|\hat{S}(t-1,f)|^2 \tag{3.5}$$

$$\hat{\Phi}_{XX}(t,f) = \alpha_x(f)\hat{\Phi}_{XX}(t-1,f) + (1-\alpha_x(f))|X(t,f)|^2 \tag{3.6}$$

It has been found experimentally that frequency dependent smoothing of the power estimates for X and $\hat{S}$ with the smoothing parameters α_x, α_s (lowpass time constants τ_x, τ_s) is useful when processing gammatone filterbank signals. In STFT-based algorithms

[3] MMSE: minimum mean squared error
[4] STSA: short-time spectral attenuation

these smoothing parameters are 0, accordingly

$$\hat{\Phi}_{XX}(t,f) = |X(t,f)|^2 \tag{3.7}$$

$$\hat{\Phi}_{SS}(t,f) = |\hat{S}(t-1,f)|^2. \tag{3.8}$$

The amount of smoothing reduces amplitude modulations and has to be selected carefully to not destroy important speech information. On the other hand, the choice of the time constants has an influence on distortions in the filtered output signal. Therefore constants will be evaluated experimentally with a perceptual quality measure that is discussed in the following section.

3.3.2. Perceptual Quality Measures

The perceptual similarity measure (PSM) described in section 2.4.2 is a broadband measure and therefore it can not directly be used to analyze noise reduction effects in subbands. To have a frequency dependent quality measure the perceptual similarity measure was calculated by omitting the integration step over all subbands which is done in the original PSM.
Let I_{tf} denote the time-frequency dependent internal representation of the estimated speech signal $\hat{s}(t)$, i.e., the transformation of the audio signal by the auditory model depicted in the right panel of 2.2 including the modulation lowpass. D_{tf} is the internal representation of the desired signal, the reference, respectively. μ_I, μ_D denote the temporal mean of the internal representations I_{tf} and D_{tf}. The subband similarity measure is then given by

$$\mathrm{PSM}(f) = \frac{\sum_t (I_{tf} - \mu_I(f))(D_{tf} - \mu_D(f))}{\sqrt{\sum_t (I_{tf} - \mu_I(f))^2 \sum_t (D_{tf} - \mu_D(f))^2}} \tag{3.9}$$

3.3.3. Parameter Optimization

In [62] we showed that the perceptual quality measure PSM from PEMO-Q has a high correlation with the subjective ratings of the overall quality. By varying the smoothing parameter of the STFT-based Ephraim-Malah algorithm (according to α in eq. 3.5) we could predict the optimal smoothing in terms of subjective overall quality. As a consequence, the parameter τ could be optimized by maximizing PSM.
In the case of the gammatone-filterbank based algorithm we have multiple parameters that are frequency dependent because of variable filter bandwidths and time resolution. The filterbank in the noise reduction system is similar to that used in PEMO-Q. This allows us to see the effects of frequency dependent algorithm parameters on each subband of the internal representation. If we combine the smoothing parameters $\tau_X(f)$ and $\tau_S(f)$ (eqns. 3.5,3.6) to a parameter vector

$$\mathrm{params}(f) = \{\tau_X(f), \tau_S(f), \ldots\} \tag{3.10}$$

then

$$\text{params}^{\text{opt}}(f) = \arg \max_{\text{params}(f)} \text{PSM}(f) \tag{3.11}$$

describes the optimal frequency dependent parameter vector. This can be used for automatic subband quality optimization of the proposed algorithm. However, unconstrained independent subband optimization can lead to a large variation of the optimal parameters across frequency bands. This happens if the variance of subband PSM-values for different settings is small, then, slight numerical changes of the input signal can cause a great change of "optimal" values. To overcome this problem we suggest to add the constraint that only small parameter changes between adjacent frequency bands are allowed and that the parameter values change monotonically with frequency.

The following section discusses the effects of the noise reduction system on the internal representation.

3.3.4. Effects of the Noise Reduction on the Internal Representations

Fig. 3.4 (a) shows a temporal section of the power envelope and (b) the related internal representation (IR) in subband 10 (569 Hz) for the clean speech signal (red dotted) and the noisy (speech-shaped noise, 5dB SNR) input signal (black solid). It can be seen in (b) that the most striking differences between clean speech and noisy signal are the stronger overshoots and undershoots in the clean speech signal IR whereas the behavior of the subband power envelope (a) is different: Here, the additive noise only influences the envelope in speech pauses and does not raise the peaks significantly. For stationary inputs, the (adaptation) feedback loops of the auditory model have a compressive effect (see [34]). Therefore, the noisy signal IR has less peaks than the clean speech signal IR, assuming that the noise is stationary compared to the speech signal. The task of the noise reduction scheme can be interpreted as reconstruction of the peaks of the speech signal IR. One drawback of spectral subtraction based noise reduction schemes is the occurrence of musical tones that can be identified in the IR as erroneous peaks (see Fig. 3.4 (c)). Here, the parameters of the noise reduction algorithm have been optimized to generate a processed signal IR (green dashed) that has the highest possible correlation for the given parameter space. The correlation between the reference IR (red dotted) and the processed signal IR (green dashed) is only slightly higher than the correlations of the IRs in (b), while the difference between the related audio signals is clearly audible. This shows that the results in single channel noise reduction systems are always a suboptimal trade-off between noise reduction and speech distortion. Even if the SNR is enhanced, the perceptual quality, predicted by the measure PSM, can hardly be improved. This implies that with the given parameter space of the algorithm it is impossible to get closer to the desired clean speech reference. Note that a perfect match (correlation = 1) between the IRs would predict that the processed signal is indiscriminable from the clean speech.

In summary, the clean speech signal IR cannot be reconstructed by single channel noise reduction schemes. This leads us to the assumption that a noisy signal at a higher SNR is better suited than a clean speech reference for perceptual optimization. The effect of the optimization with different reference signals on the IR is depicted in Fig. 3.5. It shows the IRs of the test signals (green dashed) and the references (red dotted) at higher

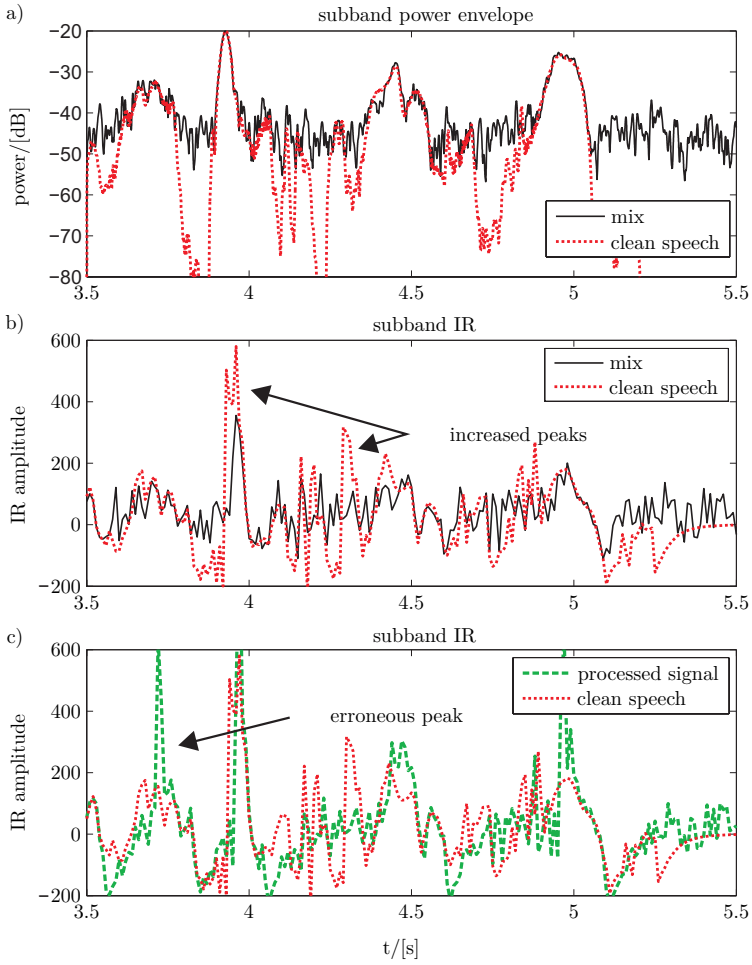


Figure 3.4.: Subband power envelope (a) and internal representations (b,c) for subband 10 (center frequency $f_c = 569$ Hz)

frequency bands ($f_c = 2119$ Hz). The noise reduction seems to be better suited for high frequency bands and therefore also the correlation between test signal and reference IR is higher compared to low frequency bands. We found that the optimization with the noisy reference signal sometimes leads to less artifacts (erroneous peaks in the test signal, see Fig. 3.5 (a) 1.6 sec.), because the noisy reference allows for residual noise after noise reduction (0.5-1 sec.). Using a noisy signal with an SNR of 25 dB (20 dB above the input signal) as a reference for the quality measure and incorporating the optimization constraints mentioned above, the perceptually optimal time constants of the proposed noise reduction algorithm are shown in table 3.2.

The subjective quality of the output signals was significantly better compared to the direct implementation with time constants $\tau_x, \tau_s = 0$. This approves the assumption that automatic parameter optimization with perceptual quality measures leads to a higher quality of the processed audio signal. However, compared to the STFT-based noise reduc-

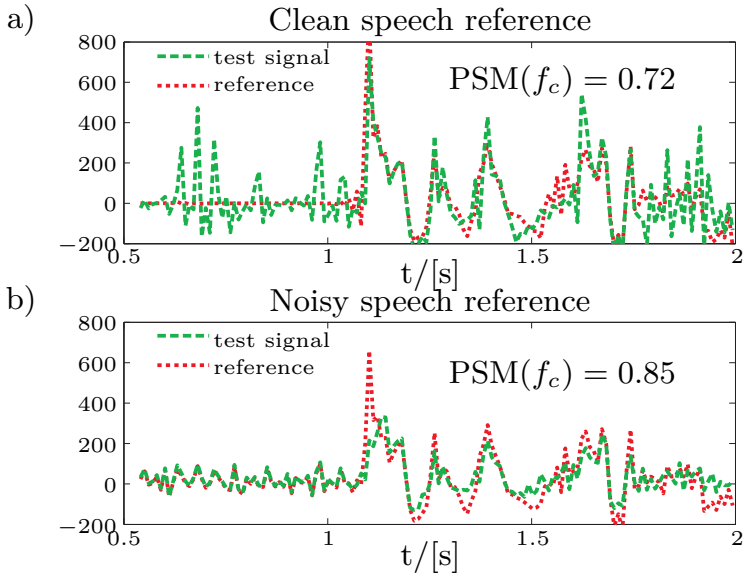


Figure 3.5.: Parameter optimization using a clean speech reference (a) and a noisy (25 dB SNR) speech reference (b)

Band	1	2	3	4	5	6	...	26	27
τ_s/[ms]	2.0	2.0	1.9	1.9	1.8	1.8	...	1.0	1.0
τ_x/[ms]	100.0	96.5	93.1	89.6	86.2	82.7	...	13.5	10.0

Table 3.2.: Perceptually optimal smoothing time constants derived from a subband-based parameter optimization of the auditory filterbank-based noise reduction algorithm.

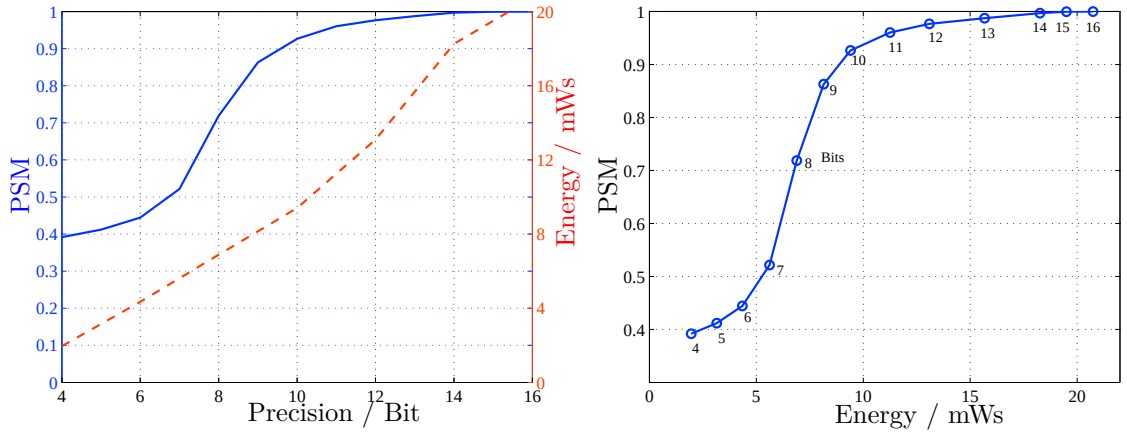
tion scheme the quality could not be improved, yet. Two reasons can be given for that: First, the bandwidths of the auditory filterbank for low frequencies are smaller than typical FFT-bandwidths which results in stronger envelope fluctuations in these bands. This means that the discrimination between speech and noise based on statistical properties of the envelope is more difficult and leads to more errors. Second, the proposed gain-table by [22] was optimized on the statistical properties of STFT-signals and does not hold for filterbank-signals with variable bandwidths.

3.4. Quality Assessment for Low-Power Applications

Hearing aids make high demands regarding the energy efficiency of the DSP circuit. Typically, the minimum requirements on operation time using a single battery are at least a few days. Thus, low-power optimization techniques are of high interest. Typically, low-power optimizations techniques work at a low level of abstraction and begin when the algorithms are already finalized. This reduces the effective degrees of freedom for power optimization which would be higher if the information about the expected power consumption of, e.g.,

a noise reduction scheme would already be available to the developer on a higher level of abstraction. One possibility to reduce power consumption is, e.g., to reduce the bit-rate of digital signals or to use adapted fixed point number systems that are appropriate for audio-signal processing applications. However, quantization effects are a consequence that generally have a non-linear influence on both the perceptual audio quality and the power consumption of the circuit. Furthermore, the quantization of algorithm parameters (e.g., filter coefficients) and data (e.g., time-varying spectral envelope) may have different requirements on accuracy. Consequently both, expected power consumption and perceptual audio quality should be available to the algorithm developer to facilitate decisions on a trade-off scale at least between these both competing properties of a low power algorithm. Hence, in a DFG-funded project AVSy, a mixed interdisciplinary team of computer scientists, electrical engineers and (psycho-)physicists explored ways to combine both measures in a single application. The subsequent section gives a brief summary of one of the main project outcomes.

3.4.1. Experiments and Results



(a) Audio signal quality and energy consumption over bitlength of filter coefficients. Solid line shows quality measure PSM (section 2.4.2) and dashed line shows energy consumption estimated by the power estimation tool ORINOCO [71]

(b) Audio signal quality over energy consumption

Figure 3.6.: Trade-off between power consumption and audio quality in a Matlab algorithm

Fig. 3.6 shows the results of a low-power optimization experiment with an arbitrary finite impulse response (FIR)-filter. The aim was to show the dependency of bitlength versus power consumption and bitlength versus signal quality on a high abstraction level, i.e., in a Matlab algorithm. It is realized using three toolboxes from the project partners that had been integrated in a Matlab simulation environment: (i) A library that simulates the

behavior of low-power circuits, (ii) a tool that can estimate the expected signal-dependent power consumption for the low-power circuit and (iii) the signal quality toolbox. The left panel, Fig. 3.6a, shows that the precision of the filter coefficients has a non-linear influence on both power consumption and signal quality. In the right panel, Fig. 3.6b, these results are used to show the dependance between power consumption and signal quality. Obviously, the highest objective quality enhancement per mWs energy consumption is expected between 7 and 9 Bit precision of the filter coefficient. For 11 to 13 Bit the quality enhancement becomes significantly lower and is negligible between over 14 Bit precision.With this information available, the developer would most probably decide to use a FIR filter with a precision around 12 Bit, while a decision based on the power consumption alone or only on the estimated signal quality would be difficult.

3.5. Conclusion

The results in section 3.2 showed that objective measures are able to predict subjective ratings in noise reduction schemes. In terms of noise reduction alone the SNRE measure was appropriate, but for objective assessment of perceived speech signal distortion or overall quality, perceptual measures such as PESQ and PSM (PEMO-Q) had higher correlations with subjective data and thus were better suited. Whereas PESQ was optimized for speech quality, PSM is a more general audio quality measure that is also applicable to, e.g., processed music and transients. The LAR measure which had shown high correlation with subjective quality measurement in [50, 59] in the current study only had high correlation with the overall quality rating for stationary white noise. For the overall correlation analysis done here, using different noise types and covering a broader range of signal qualities than in comparable studies, the high correlation of the LAR with subjective overall quality ratings could not be confirmed.
In section 3.3 a new method for subband based quality prediction and parameter optimization was proposed which was tested and analyzed on a monaural noise reduction algorithm. The direct conversion of the STFT-based algorithm due to [22] led to strong interferences and artifacts in the audio signal. These artifacts could be reduced by the proposed perceptual quality optimization scheme. With these settings the processed audio signal had a quality which was comparable to STFT-processed signals. However, the noise reduction scheme could not be improved compared to a constant bandwidth STFT-method by using an auditory filterbank. Looking at details of the internal representations in subbands, we were able to interpret the principle limitations of the monaural noise reduction scheme.
In section 3.4 parts of the results from the joint project AVSy were shown. The goal was to design an integrative simulation and evaluation tool for low-power algorithms which are applicable for hearing aids. The power optimization of those algorithm types is generally done at a stage of development where the algorithm design is already fixed and, moreover, only technical distortion measures are considered. Thus, it was desirable to have the information about power consumption and *perceptual* signal quality available at a higher abstraction level of the algorithm development. This was realized with the development of an integrative simulation and evaluation tool including power loss estimation and objective signal quality assessment. However, for complexity reasons concerning the

simulation of low-power circuits only exemplary results optimizing the bit resolution of FIR-filter coefficients could be shown here.

To conclude, from a selection of the best objective measures for each rating task together with representative noise types (section 3.2) a test-bench was defined and used for quality assessment of a typical low-power application (section 3.4) and for the parameter optimization of a novel noise reduction scheme (section 3.3). However, because of the high dimensionality of the parameter space, a sub-band measure based on PEMO-Q was derived which could successfully be used for parameter optimization. Although the single-channel noise reduction scheme was based on auditory filters, to match the frequency resolution of the algorithm to human perception, no significant benefit to traditional STFT-based methods was found. Hence, it is suggested that there are principle limitations of single-channel short time spectral attenuation (STSA) algorithms that cannot easily be overcome. Therefore, in the next section multi-channel algorithms are investigated and evaluated using the performance measures that have been identified to be significant for human perception of noise reduction.

4. Multi-Channel Noise Reduction Schemes with Binaural Output - Performance Evaluation and Optimization[1]

4.1. Introduction

Multi-channel spatial beamformers have been shown in the literature to be a useful element of hearing aids in order to suppress noise from *undesired* directions and to enhance the sound emanating from a *target* direction (see, e.g. [17, 27, 56]). The standard configuration for a single hearing aid on one side employs two or three microphones spaced closely to each other (maximum distance approx. two centimeters) and uses either a fixed or an adaptive beamforming algorithm ([20, 27, 37]) in order to typically enhance the signals emanating from the front of the wearers head. A higher functional array gain (i.e., improvement in signal-to-noise ratio assuming a constant microphone noise) at low frequencies can be achieved by applying superdirectivity [5, 11] which increases the susceptibility to self-noise of the microphones and model errors, and by increasing the physical distance between the left-most and rightmost microphone of the array. Hence, several solutions for hearing aid wearers with external arrays (for example mounted in a pair of glasses as broadside or endfire array according to Soede, [68, 69]) have successfully been introduced. An alternative approach to extend the physical dimensions is to utilize the acoustical input to both ears of the user and hence to mimic (to a certain degree) the acoustical aspects of the *cocktail party processing* normally performed by the human brain on the input from both ears ([13, 44, 82]). Since this requires a preferably wireless connection between the microphones at both sides of the head a number of papers considering this general setup has emerged only recently ([38, 47, 64, 65, 75, 85]). However, several problems connected to this *binaural array processing* have not been dealt with in the previous approaches in a systematic way. One important point is the binaural output mode: While the classical beamforming algorithms provide only one (monaural) output signal, *classical* cocktail-party processing schemes (such as, e.g. [43, 44]) provide a binaural output signal to both ears that enable the listener to still take advantage of the remaining binaural cues in the output signal (e.g., for localization and separation of acoustical objects and for spatial perception of the environment). Even if a binaural output is provided by a beamformer system, several options exist about the interaural relation between both output channels ([38, 47, 75, 82]) that may interact with the user's remaining binaural processing capability in a yet to be explored way. The current chapter therefore studies

[1] Parts of this chapter have been submitted as "Parameter Optimization for a Class of Binaural Multi-Channel Noise Reduction Schemes for Hearing Aids based on Perceptual Quality Measures" to IEEE Trans. on Audio, Speech and Language Processing, 2008

the effect of the binaural output mode in a systematic way.
Another important point is the model of the wave propagation: The diffraction properties of the sound propagating around the users head provide an acoustic decoupling across the microphones at both sides which might improve the array performance when properly considered in the beamformer design. A comparatively coarse model of this propagation has the advantage of being largely independent from the individual's exact geometric dimensions of the head and the pinna whereas an exact model (incl., e.g., measured head related transfer functions (HRTF)) may provide a better functional array gain. In order to find out if any of these extreme cases in head-shadow modeling or any compromise provides the best solution, a systematic evaluation was performed in this study.
Most beamformer approaches in the past have been designed and evaluated using a certain model of the ambient noise field to be suppressed (such as, e.g. isotropic noise under free field assumption, measured noise fields on a head or incoherent noise at the microphones, [5, 52]). In practical every day listening situations, however, these assumptions are not necessarily met. Additionally, certain noise field assumptions raise the array's susceptibility against uncorrelated or spatially white noise which has influence on the noise reduction performance and signal quality. Hence, the influence of different noise field characteristics is an important factor for a beamformer design and will be studied in a systematic way.
In literature, adaptive beamforming schemes have often been found to be superior in noise reduction performance to fixed beamformers [26]. While a fixed beamforming array characteristic has the advantage of being robust against estimation errors of the respective target and noise signal, adaptive beamformers are known to be less robust but have the advantage to adaptively steer spatial notches to the most disturbing noise source direction(s). Since it is unclear if the relative benefits from each solution outperform the respective potential disadvantages (in particular with respect to the more complex propagation model for head-worn arrays) a systematic evaluation was performed here.
In order to study the influence of those parameters listed above in a systematic way, a class of six-channel binaural beamformer systems is investigated here that operates on the output signal from a pair of binaurally-connected three-microphone behind-the-ear (BTE) hearing aids. An optimization of the respective parameters was performed by evaluating the performance of the algorithm with different objective measures and analyzing its robustness against mismatch of the real situation from the assumed acoustical situation.
As a most common physical performance measure the enhancement of the signal-to-noise ratio (SNRE) is considered here. The SNRE had shown high correlations with subjective ratings of background noise reduction for monaural noise reduction schemes [62]. However, as the amount of noise reduction often competes with speech distortion [35, 62] the SNRE does not accurately reflect human benefit in speech intelligibility or overall subjective perceptual quality. Hence a monaural objective quality measure based on an elaborate psychoacoustical processing model (PEMO-Q, [34]) has been used in the study to estimate the objective performance.
Psychoacoustical-model-based performance measures used in the past do not adequately take into account the respective benefit of binaural cues in the output signal that can be utilized by the human user of the binaural processing scheme. Hence, some discrepancies can be observed between the actual performance of binaural noise reduction schemes with humans and the predicted performance [75]. In order to account for this effect, a binaural

speech intelligibility prediction model BSIM [3] was used to assess the relative benefit that human listeners can achieve from the (remaining) binaural cues in the output signal.
The different binaural output schemes where tested in a subjective listening test with normal hearing listeners. The remainder of this chapter is organized as follows. In Section 4.2 the acoustical setup is introduced as a basis for the algorithm and evaluation. Section 4.3 describes the signal model, the beamformer algorithms and the binaural output schemes. The influence of the propagation model on the beamformer design is discussed in Section 4.3.4. A summary of signal-independent and signal-dependent performance measures is given in Section 4.4. The experiments and results on the perceptual optimization of the white noise gain constraint and on the binaural output quality follow in Section 4.5 together with a comparison of adaptive and fixed beamformers and robustness measurements. These results are discussed in Section 4.7 and the different binaural output modes are assessed with normal hearing listeners in a subjective listening test measuring the hearing effort in Section 4.6. Finally, Section 4.8 concludes the paper.

4.2. Acoustical Setup

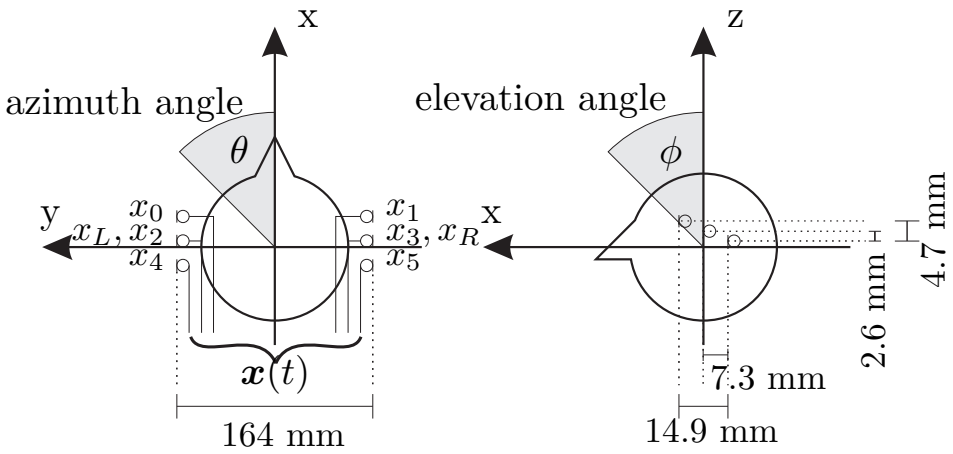


Figure 4.1.: Acoustical setup: Two linear microphone arrays are mounted bilaterally on a B&K HATS. Each array consists of 3 hearing aid microphones mounted in a hearing aid shell with a distance of 8 mm. The frontal direction is the x-axis which is equal to an azimuth angle $\theta = 0°$ and an elevation angle $\phi = 90°$.

Fig. 4.1 shows schematically the acoustical setup and the coordinate system used for defining microphone positions and sound source directions. 6-channel signals ($M = 6$) have been recorded from two 3-channel BTE hearing aid shells (Siemens Acuris) mounted on a Brüel & Kjær (B&K) head and torso simulator (HATS). The impulse responses (IRs) for all microphones have been measured with this setup in an anechoic room for azimuth directions of 0-180° in 5° steps at an elevation of 0° (horizonal plane). In the following these are referred to as 6-channel head related transfer functions (HRTFs) in the frequency domain that include head-shadow and diffraction effects, and the characteristics of the microphones. Similarly, HRTFs have been measured in an office environment (reverberation

time $\tau_{60} = 300$ ms).

Directional target speech and interfering noise signals were calculated by filtering source signals with these HRTFs. In addition, real-world environmental noise has been recorded in a cafeteria and in an office room. Furthermore, an artificial diffuse noise has been generated by filtering a speech-colored random noise with the anechoic HRTFs from all directions and summing up all filtered noise signals. This signal simulates a cylindrical 2D-isotropic noise field. From the database of 6-channel directional speech and noise signals various mixtures have been calculated for different signal-to-noise ratios (SNRs).

The target signal was varied in three acoustic scenes: Target speaker with anechoic HRTF from 30° (condition 1), from 0° (condition 2), and target speaker with office HRTF from 30° (condition 3). All signals were mixed with the superposition of recorded noise from the cafeteria and an interferer speech signal from −135°. The input signal-to-noise ratio (SNR) and performance values (see section 4.4.2) are given in table 4.1.

condition	reverberation τ_{60}	target signal	input SNR Left	input SNR Right	input PSM Left	input PSM Right	input SRT
1	< 5 ms	30°	4.3 dB	1.3 dB	0.6	0.34	-9.85 dB
2	< 5 ms	0°	2.4 dB	3.2 dB	0.52	0.45	-8.15 dB
3	300 ms	30°	3.2 dB	2.5 dB	0.60	0.45	-10.25 dB

Table 4.1.: Signal conditions and input values of the reference microphones

4.3. Algorithm

Fig. 4.2 shows the block diagram of the noise reduction scheme which will be described in the following. Note that the algorithm is not limited to the 6-channel setup used here but applies to any M-channel microphone array mounted near to a head. Throughout the paper, vectors and matrices are printed in boldface, scalars in italics. t denotes the time, ω the radian frequency and k the block-index. The superscripts ^{T}, * and ^{H} denote the transposition, the complex conjugation and the Hermitian transposition, respectively.

4.3.1. Signal Model

The multi-channel signal $\boldsymbol{x(t)} = [x_0(t), \ldots, x_{M-1}(t)]^T$ (Fig. 4.1, 4.2) is assumed to be a mix of the directional target signal $\boldsymbol{s}(t)$ and a noise signal $\boldsymbol{n}(t)$. In the frequency domain the signal model can be formulated as

$$\boldsymbol{X}(\omega, k) = \underbrace{\boldsymbol{d}_S(\omega) S(\omega, k)}_{\boldsymbol{S}(\omega,k)} + \boldsymbol{N}(\omega, k) \tag{4.1}$$

where the capital letters denote the time-frequency transformed signals of $\boldsymbol{x}, \boldsymbol{s}$, and $\boldsymbol{n}$ calculated by a STFT. The propagation vector $\boldsymbol{d}_S(\omega) = \boldsymbol{d}(\omega, \theta_S, \phi_S)$ is the vector of transfer functions between the source signal $S(\omega)$ and the signal vector $\boldsymbol{S}(\omega)$ observed at

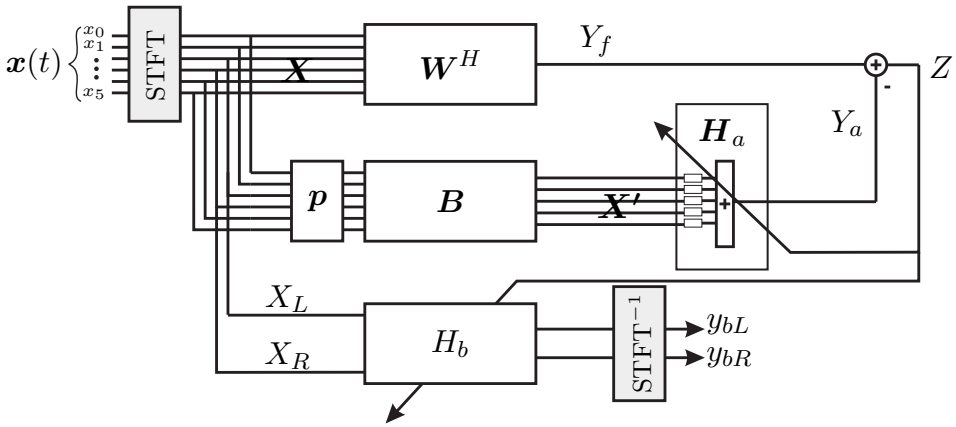


Figure 4.2.: Multi-channel beamformer system with binaural output. $\boldsymbol{W}^H$ is the fixed beamformer filter, $\boldsymbol{B}$ denotes the blocking matrix, $\boldsymbol{H}_a$ is the adaptive filter, and H_b is the filter that generates a binaural output from the reference microphone signals $X_2(= X_L)$ and $X_3(= X_R)$ at the left and right ear, $\boldsymbol{p}$ is the compensation vector that cancels the target signal in combination with $\boldsymbol{B}$.

the sensors. In general, the propagation vector for a signal coming from the azimuth angle θ and the elevation angle ϕ is

$$\boldsymbol{d}(\omega, \theta, \phi) = [d_0(\omega, \theta, \phi), \ldots, d_{M-1}(\omega, \theta, \phi)]^T \tag{4.2}$$

where the transfer function to a microphone $i = 0 \ldots M-1$ is

$$d_i(\omega, \theta, \phi) = a_i(\omega, \theta, \phi) e^{-j\varphi_i(\omega, \theta, \phi)} \tag{4.3}$$

where $a_i(\omega, \theta, \phi)$ denotes the amplitude spectrum and the group-delay can be calculated by $\tau_i(\omega, \theta, \phi) = \frac{\partial \varphi_i(\omega, \theta, \phi)}{\partial \omega}$.

4.3.2. Beamformer

A fixed filter-and-sum beamformer can be designed in the frequency domain to produce a monaural output that contains less noise energy than the multi-channel input signal $\boldsymbol{X}$ by

$$Y_f(\omega, k) = \sum_{i=0}^{M-1} W_i^*(\omega) X_i(\omega, k) = \boldsymbol{W}^H(\omega) \boldsymbol{X}(\omega, k). \tag{4.4}$$

The optimal filter $\boldsymbol{W}$ can be calculated by the well-known minimum variance distortionless response (MVDR) solution [5]:

$$\boldsymbol{W}(\omega, \theta, \phi) = \frac{{\boldsymbol{\Phi_{NN}}}^{-1}(\omega) \boldsymbol{d}(\omega, \theta, \phi)}{\boldsymbol{d}^H(\omega, \theta, \phi) {\boldsymbol{\Phi_{NN}}}^{-1}(\omega) \boldsymbol{d}(\omega, \theta, \phi)}, \tag{4.5}$$

where $\boldsymbol{\Phi_{NN}}^{-1}$ denotes the inverse noise correlation matrix which is discussed in 4.3.4. The fixed beamformer can be extended by an adaptive noise cancellation path which consists of a delay- (and amplitude-) compensation step, denoted by the delay compensation vector $\boldsymbol{p}$, followed by a blocking matrix $\boldsymbol{B}$ (producing the noise reference $\boldsymbol{X'}$) and an a multi-channel Wiener filter $\boldsymbol{H}_a$ that is adapted to cancel out noise components that $\boldsymbol{X'}$ and Y_f have in common. The (element-wise) Hadamard product of the delay compensation vector $\boldsymbol{p}$ and the propagation vector $\boldsymbol{d}$ should result in a zero-delay vector with amplitude 1:

$$\boldsymbol{p} \bullet \boldsymbol{d} = \mathbf{1} = [1, \ldots, 1]^T . \tag{4.6}$$

Thus, $\boldsymbol{p}$ is defined by

$$\boldsymbol{p}(\omega, \theta, \phi) = \left[\frac{{d_0}^*(\omega, \theta, \phi)}{|d_0(\omega, \theta, \phi)|^2}, \ldots, \frac{{d_{M-1}}^*(\omega, \theta, \phi)}{|d_{M-1}(\omega, \theta, \phi)|^2} \right]^T , \tag{4.7}$$

and the blocking matrix (which is a $[M-1 \times M]$ - subtraction matrix) is [5]

$$\boldsymbol{B} = \begin{bmatrix} 1 & -1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & 1 & -1 \end{bmatrix} . \tag{4.8}$$

The noise reference matrix $\boldsymbol{X'}$ at the output of the blocking matrix is

$$\boldsymbol{X'}(\omega, k) = \boldsymbol{B}(\boldsymbol{p}(\omega, \theta, \phi) \bullet \boldsymbol{X}(\omega, k)). \tag{4.9}$$

The multi-channel Wiener filter is designed with

$$\boldsymbol{H}_a(\omega) = {\boldsymbol{\Phi}_{\boldsymbol{X'X'}}}^{-1}(\omega) \boldsymbol{\Phi}_{\boldsymbol{X'Y_f}}(\omega) \tag{4.10}$$

where the PSD-matrix $\boldsymbol{\Phi_{X'X'}}$ and the cross-PSD row vector $\boldsymbol{\Phi_{X'Y_f}}$ denote expectation values defined by

$$\boldsymbol{\Phi_{X'X'}}(\omega) = \mathrm{E}\left\{ \boldsymbol{X'}(\omega) {\boldsymbol{X'}}^H(\omega) \right\}, \tag{4.11}$$

$$\boldsymbol{\Phi_{X'Y_f}}(\omega) = \mathrm{E}\left\{ \boldsymbol{X'}(\omega) Y_f^*(\omega) \right\}. \tag{4.12}$$

In practice, $\boldsymbol{\Phi_{X'X'}}$ and $\boldsymbol{\Phi_{X'Y_f}}$ are calculated by recursively averaging instantaneous short-time spectra:

$$\begin{aligned} \boldsymbol{\Phi_{X'X'}}(\omega, k) &= \alpha \boldsymbol{\Phi_{X'X'}}(\omega, k-1) + (1-\alpha) \boldsymbol{X'}(\omega, k) {\boldsymbol{X'}}^H(\omega, k), \boldsymbol{\Phi_{X'Y_f}}(\omega, k) & (4.13) \\ &= \alpha \boldsymbol{\Phi_{X'Y_f}}(\omega, k-1) + (1-\alpha) \boldsymbol{X'}(\omega, k) Y_f^*(\omega, k). & (4.14) \end{aligned}$$

Therefore, also the filter $\boldsymbol{H}_a$ is slowly varying over time and the noise estimate of the adaptive path, Y_a, is calculated by

$$Y_a(\omega, k) = \boldsymbol{H}_a^H(\omega, k) \boldsymbol{X'}(\omega, k) \tag{4.15}$$

which then can be subtracted from the fixed beamformer output so that we get the monaural output of the generalized sidelobe canceller (GSC):

$$Z(\omega, k) = Y_f(\omega, k) - Y_a(\omega, k). \tag{4.16}$$

In summary, we get the monaural outputs of the two beamformer types:

$$\text{fixed} : Z(\omega, k) = Y_f(\omega, k) = \boldsymbol{W}^H(\omega)\boldsymbol{X}(\omega, k), \tag{4.17}$$

$$\text{adaptive} : Z(\omega, k) = \boldsymbol{W}^H(\omega)\boldsymbol{X}(\omega, k) - \boldsymbol{H}_a^H(\omega)\boldsymbol{X'}(\omega, k). \tag{4.18}$$

Thus, the difference between fixed and adaptive beamformer consists of an additional noise subtraction path which can be added to the fixed beamformer. Note, that the original GSC [28] uses a standard delay-and-sum (D&S) beamformer in the fixed processing path, whereas we use an arbitrary superdirective design here, which is discussed below.

4.3.3. Binaural Output

The output can be extended to a binaural signal with left and right output signal Y_{bL} and Y_{bR}

$$\boldsymbol{Y_b}(\omega, k) = [Y_{bL}(\omega, k), Y_{bR}(\omega, k)]^T \tag{4.19}$$

with different strategies.

Target signal phase reconstruction

The simplest solution might be to reconstruct the phase and amplitude response of the target signal by multiplying the monaural output with the propagation coefficients d_L, d_R that relate to the reference microphones (denoted as x_L and x_R in Fig. 4.1) at the left and right hearing aid array, respectively:

$$Y_{bL}(\omega, k) = d_L(\omega, \theta, \phi)Z(\omega, k), \tag{4.20}$$

$$Y_{bR}(\omega, k) = d_R(\omega, \theta, \phi)Z(\omega, k). \tag{4.21}$$

However, this can only reconstruct the gross magnitude and phase characteristic of the target signal that is included in the assumed propagation model whereas the binaural information of the interfering noise signal is lost.

Binaural post-filter

A method to preserve the phase of both, signal and noise, can be realized according to [47] by applying a real-valued time-varying post-filter to the reference microphone signals

X_L, X_R:

$$H_b(\omega,k) = \frac{\left(|d_L(\omega,\theta,\phi)|^2 + |d_R(\omega,\theta,\phi)|^2\right)\Phi_{ZZ}(\omega,k)}{\Phi_{X_LX_L}(\omega,k) + \Phi_{X_RX_R}(\omega,k)} \tag{4.22}$$

$$Y_{bL}(\omega,k) = H_b(\omega,k)X_L(\omega) \tag{4.23}$$

$$Y_{bR}(\omega,k) = H_b(\omega,k)X_R(\omega). \tag{4.24}$$

$\Phi_{ZZ}, \Phi_{X_LX_L}$ and $\Phi_{X_RX_R}$ denote the power spectral density estimates for the signals Z, X_L, X_R, respectively. In practice, these can be estimated by recursively smoothing instantaneous signal powers. The binaural post-filter can be interpreted as a single-channel envelope Wiener filter applied to both reference channels X_L, X_R. Additional gain rules known from single channel noise reduction systems can be applied here.

Bilateral Beamformer

To investigate the behavior of two independently working unilateral beamformers $\boldsymbol{W_L}$ (left) and $\boldsymbol{W_R}$ (right), the system depicted in Fig. 4.2 can be split into two subarrays where $\boldsymbol{X_L} = [X_0, X_2, \dots X_{M-2}]$ denotes the signal matrix of the left subarray using the even-numbered microphones and $\boldsymbol{X_R} = [X_1, X_3, \dots X_{M-1}]$ denotes the signal of the right subarray using the odd-numbered microphones. $\boldsymbol{X'_L}, \boldsymbol{X'_R}$ are defined according to (4.9) but for shorter blocking matrices and delay compensation vectors $\boldsymbol{p_L}, \boldsymbol{p_R}$, respectively.

$$\begin{aligned} Y_{bL}(\omega,k) &= Z_L(\omega,k) \\ &= \boldsymbol{W_L}^H(\omega)\boldsymbol{X}_L(\omega,k) - \boldsymbol{H}_{a_L}^H(\omega)\boldsymbol{X'}_L(\omega,k) \end{aligned} \tag{4.25}$$

$$\begin{aligned} Y_{bR}(\omega,k) &= Z_R(\omega,k) \\ &= \boldsymbol{W_R}^H(\omega)\boldsymbol{X}_R(\omega,k) - \boldsymbol{H}_{a_R}^H(\omega)\boldsymbol{X'}_R(\omega,k) \end{aligned} \tag{4.26}$$

The subarrays do not need to be restricted to one side but can use any combination of microphones from both sides if a connection between the bilateral arrays exists. In the case of a complete bilaterally connected system every filter gets the complete M-channel information. However, in this case additional constraints have to be included into the beamformer design to partially reconstruct the binaural information of the target and noise signal. A detailed analysis of binaural systems of this type for two microphones can be found in [82] and for six microphones in [15].

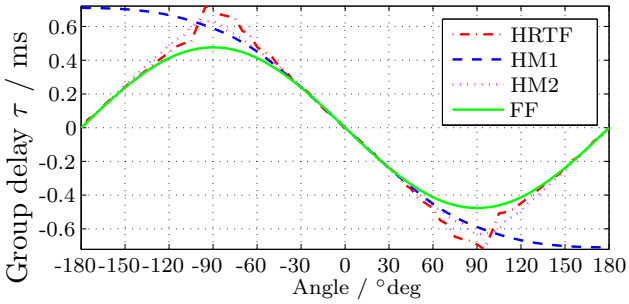
In summary, three different methods that produce a binaural output can be distinguished. In the following, the signal phase reconstruction method is denoted as (BIN_PR), the binaural post-filter as (BIN_PF), and the bilateral system using only the left (respectively, right) subarray is denoted as (BIN_BL). The monaural output Z is denoted as (MON).

4.3.4. Influence of Different Propagation Models on the Beamformer Design

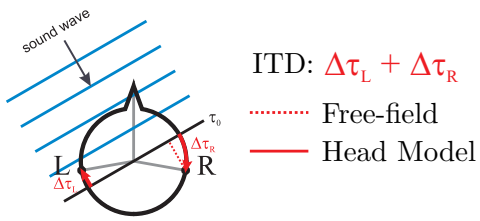
The fixed beamformer coefficients given by (4.5) ideally reduce a noise field[2] with the correlation matrix $\mathbf{\Phi}_{NN}$ under the constraint of an undistorted signal response in the desired look direction. The more exactly $\mathbf{\Phi}_{NN}$ is known, the higher is the noise reduction performance. The absence of distortion for the MVDR beamformer, however, is only given if the propagation model $\boldsymbol{d}$ used for the beamformer design and the true signal wave propagation vector $\boldsymbol{d}_S$ perfectly match. In general, the exact transfer functions $\boldsymbol{d}_S$ are unknown and several assumptions about the wave propagation must be made. The influences of the exactness of the propagation model on the beamformer performance are discussed below.

Propagation vector

All effects could be perfectly integrated into the beamformer design if the transfer functions $\boldsymbol{d}_S$ could be measured in the situation of interest, including the room response, the head-shadow and diffraction effects, and the microphone characteristics. However, as estimating the room response for a given target signal is not feasible under realistic conditions the second-best solution is measuring the anechoic transfer functions of the system including the head-influences and the microphone characteristics. These transfer functions can directly be used as a propagation vector $\boldsymbol{d}$ in (4.2) and will be referred to as HRTF in the following. If the anechoic HRTF is not available, the gross head-shadow and diffraction effects can be modeled by the wave propagation observed on a rigid sphere [8, 18]. For head-models, both, a_i and τ_i in (4.3) are angle and frequency dependent.



(a) Group delay for different propagation models.



(b) Diffraction at a sphere.

Figure 4.3.: Interaural time delay for measured HRTFs and head models.

In general, it is assumed that the target source is approximately in the horizontal plane, i.e., $\phi_S \approx 90°$. Therefore, the elevation angle ϕ_S will be disregarded in the following for the head-related wave propagation models used in this study. The first head model (HM1) by [8] is a simple and effective parametric model that estimates the characteristics of a sphere. The interaural time difference (ITD) cues are modeled by Woodworth and Schlosberg's frequency independent (ray-tracing) formula. The gross magnitude characteristics of the HRTF spectrum, namely the interaural level difference (ILD) cues, are covered by a

[2] a superposition of many unknown noise signals

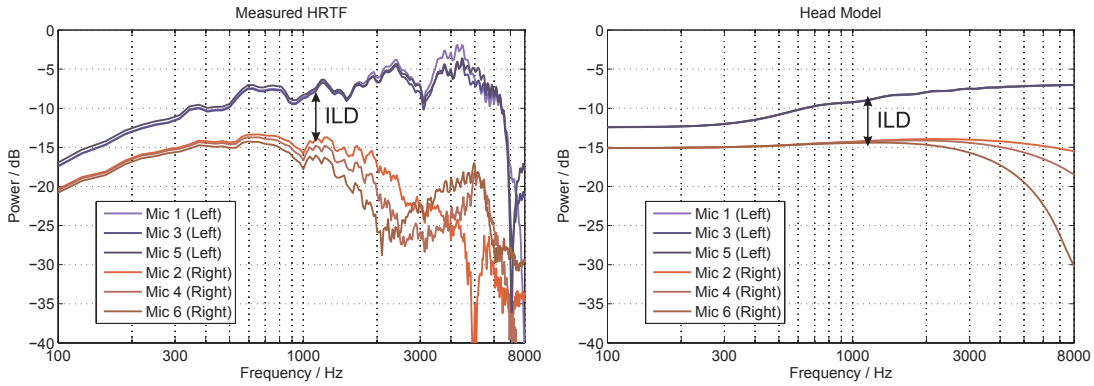


Figure 4.4.: Interaural level differences for measured HRTFs and head model (HM2).

single-pole, single-zero head-shadow filter which also accounts for an additional frequency dependent delay at low frequencies. For each microphone of the array an angle of a ray from the center of the sphere to the microphone θ_i, $i = 0 \dots M-1$, can be calculated. Choosing the angle to the desired sound source θ_S and some additional model parameters (e.g. sphere radius $r = 8.2$ cm, speed of sound, fitting parameters $\alpha_{min}, \theta_{min}$, see [8]), the transfer function is calculated by

$$\boldsymbol{d}(\omega, \theta_S) = [H_{HM1}(\omega, \theta_S, \theta_0, \text{params}), \dots, H_{HM1}(\omega, \theta_S, \theta_{M-1}, \text{params})]^T . \quad (4.27)$$

The spherical head model is calculated by

$$H_{HM1}(\omega, \theta_S, \theta_{mic}) = \frac{1 + j\frac{\alpha(\theta_S - \theta_{mic})}{2\omega_0}\omega}{1 + j\frac{\omega}{2\omega_0}} e^{-j\omega T_d(\theta_S - \theta_{mic})} \quad (4.28)$$

$$T_d(\theta) = \begin{cases} \frac{r}{c}\cos(\theta) & 0 \le |\theta| < \frac{\pi}{2} \\ \frac{r}{c}(|\theta| - \frac{\pi}{2}) & \frac{\pi}{2} \le |\theta| < \pi \end{cases} \quad (4.29)$$

$$\omega_0 = \frac{c}{r} \quad (4.30)$$

$$\alpha(\theta) = \left(1 + \frac{\alpha_{\min}}{2}\right) + \left(1 - \frac{\alpha_{\min}}{2}\right)\cos\left(\frac{\theta}{\theta_{\min}} 180^\circ\right) \quad (4.31)$$

$$\alpha_{\min} = 0.1 \qquad \theta_{\min} = 150^\circ \quad (4.32)$$

where ω_0 is the radian frequency corresponding to the speed of sound c and the radius r of the sphere, T_d is the travel time around the sphere from the angle of incidence θ to the angle $\theta_{\min}$ corresponding to the microphone position. The second head model (HM2) [18] additionally incorporates the distance of the source for modeling near-field effects and interference effects that introduce ripples in the response that are quite prominent on the shadowed side. It is numerically calculated by a recursive algorithm given in [18]. The propagation vector is built similar to HM1 (4.27). The far-field assumption implies that all microphones *see* the target sound wave arriving from the same angles (θ_S, ϕ_S) as a plane wave. Additionally assuming free-field (FF), i.e., no objects inside the sound wave path and a unity microphone response $a_i(\omega, \theta, \phi) = 1$, $\forall(\omega, \theta, \phi, i)$, the propagation coefficient

(4.3) simplifies to

$$\boldsymbol{d}(\omega, \theta_S, \phi_S) = \left[e^{-j\omega\tau_{00}(\theta_S,\phi_S)}, \ldots, e^{-j\omega\tau_{0M-1}(\theta_S,\phi_S)}\right]^T \tag{4.33}$$

where τ_{0i} is a constant group delay measured between a reference microphone 0 and microphone i. The group delay can easily be calculated based on the microphone array geometry where $\boldsymbol{l}_{0i}$ is the vector difference between a reference microphone 0 and the microphone i, c is the speed of sound, and $\boldsymbol{e}_r(\theta_S, \phi_S) = [\sin(\theta_S)\cos(\phi_S), \sin(\theta_S)\sin(\phi_S), \cos(\phi_S)]^T$ is the unit vector in target direction:

$$\tau_{0i}(\theta_S, \phi_S) \quad = \quad \frac{\boldsymbol{l_{0i}}^T \boldsymbol{e}_r(\theta_S, \phi_S)}{c}. \tag{4.34}$$

Thus, under the FF assumption the beamformer can be designed knowing the relative microphone positions and the direction of the target signal. The differences of the interaural time difference (ITD) for the propagation models are shown in Fig. 4.3 and the different interaural level difference (ILD) in Fig. 4.4.

Noise PSD-matrix

The *normalized* noise PSD-matrix $\boldsymbol{\Phi_{NN}}$ contains the information of how much the microphone signals are correlated expressed by their complex pairwise cross-power spectral densities for a measured or an assumed noise field. It is defined by

$$\boldsymbol{\Phi_{NN}}(\omega) = \frac{1}{\overline{\Phi}_{NN}(\omega)} \begin{pmatrix} \Phi_{N_0N_0}(\omega) & \ldots & \Phi_{N_0N_{M-1}}(\omega) \\ \Phi_{N_1N_0}(\omega) & \ldots & \Phi_{N_1N_{M-1}}(\omega) \\ \vdots & \ddots & \vdots \\ \Phi_{N_{M-1}N_0}(\omega) & \ldots & \Phi_{N_{M-1}N_{M-1}}(\omega) \end{pmatrix} \tag{4.35}$$

where the normalization factor $\overline{\Phi}_{NN}(\omega)$ forces the trace of $\boldsymbol{\Phi_{NN}}$ to equal M. The fact that it is inverted in the MVDR equation (4.5) can be interpreted as a decorrelation of the noise components included in $\boldsymbol{X}$. Thus, the subsequent summation of the microphone signals leads to an enhancement of correlated signal components compared to uncorrelated components. The simplest noise field model makes the assumption that the noise field is already uncorrelated, i.e. no further decorrelation is needed, and therefore it has a correlation matrix

$$\boldsymbol{\Phi_{NN}}(\omega) = \boldsymbol{\Phi_{NN}} = \boldsymbol{I} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} = \boldsymbol{\Phi_{NN}}^{-1}. \tag{4.36}$$

The optimal MVDR beamformer design for uncorrelated noise leads to a delay-and-sum (D&S)-beamformer (aka: *conventional* beamformer):

$$\boldsymbol{W}(\omega) = \frac{\boldsymbol{d}(\omega)}{\boldsymbol{d}^H(\omega)\boldsymbol{d}(\boldsymbol{\omega})} = \frac{\boldsymbol{d}(\omega)}{\sum_i a_i^2(\omega)} \tag{4.37}$$

$$= \frac{1}{M}\boldsymbol{d}(\omega),\ a_i(\omega) = 1\,\forall i \tag{4.38}$$

By summing up uncorrelated noise and correlated signal components the theoretical SNRE is $10\log_{10}(M)$ dB, i.e., ≈ 7.8 dB for $M = 6$ microphones. However, natural sound sources in general are correlated. Thus, a better noise field model can be used in (4.5) and a higher directivity can be achieved by designing a *superdirective* beamformer. It is important for the beamformer performance that the applied noise field model matches the observed noise PSD-matrix. The cross-spectral density $\Phi_{X_iX_k}$ of a signal Q arriving from azimuth angle θ as observed between the microphones i and k is given by

$$\Phi_{X_iX_k}(\omega,\theta) = \mathrm{E}\left\{Q(\omega)d_i(\omega,\theta)Q^*(\omega)d_k^*(\omega,\theta)\right\} \tag{4.39}$$

$$= \Phi_{QQ}(\omega)d_i(\omega,\theta)d_k^*(\omega,\theta) \tag{4.40}$$

$$= \Phi_{QQ}(\omega)a_i(\omega,\theta)a_k(\omega,\theta)e^{j\omega(\tau_i(\omega,\theta)-\tau_k(\omega,\theta))}. \tag{4.41}$$

The noise PSD-matrix of all noise sources can be calculated as the sum of individual noise cross power spectral densities arriving from different azimuth directions θ_v:

$$\Phi_{N_iN_k}(\omega) = \sum_{\theta_v}\Phi_{X_iX_k}(\omega,\theta_v). \tag{4.42}$$

If the directions of individual noise sources Q are unknown (which is mostly the case) the assumption of homogenously distributed sources is often made. Two typically used noise characteristics can be distinguished: A spherically isotropic or diffuse noise field (diff3D) is a good model for a reverberant room, and a cylindrical isotropic noise field (diff2D) for rooms with relatively low reflections from the ceiling and the floor. For the free-field case where the magnitude of the propagation vector a_i, a_k equal 1 and the delay only depends on the microphone distance l_{ik} and the angle of incidence θ, both noise fields can be derived by solving the integral of equal-power noise sources from all directions: If the propagation delay between two microphones i, k equals

$$\tau_i(\theta) - \tau_k(\theta) = \frac{l_{ik}}{c}\sin(\theta) \tag{4.43}$$

then the cross power spectral density of a noise source Q equals

$$\Phi_{X_iX_k}(\omega,\theta) = \Phi_{QQ}(\omega)e^{j\omega\frac{l_{ik}}{c}\sin(\theta)}. \tag{4.44}$$

Summing up an infinite number of noise sources in a plane with equal power $\Phi_{QQ} = 1$

from all directions we get

$$\Phi_{N_i N_k}(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_{X_i X_k}(\omega, \theta) d\theta \tag{4.45}$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j\omega \frac{l_{ik}}{c} \sin(\theta)} d\theta \tag{4.46}$$

$$= J_0 \left(\omega \frac{l_{ik}}{c} \right) \tag{4.47}$$

where J_0 is the zero-oder Bessel function of the first kind which describes the characteristic of a cylindrical isotropic noise. Beamformers using this noise model can easily be modified for an optimal front-to-back ratio by adjusting the limits of the integral [5]:

$$\Phi_{N_i N_k}(\omega) = \frac{1}{2(\pi - \delta)} \int_{\theta_0 - \pi + \delta}^{\theta_0 + \pi - \delta} e^{j\omega \frac{l_{ik}}{c} \sin(\theta)} d\theta \quad 0 \leq \delta \leq \pi \tag{4.48}$$

For spherically homogenous isotropic noise the integral over all azimuth and elevation angles leads to the well-known sinc-characteristic in free-field:

$$\Phi_{N_i N_k}(\omega) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \int_{-\pi/2}^{\pi/2} e^{j\omega \frac{l_{ik}}{c} \cos(\phi) \sin(\theta)} d\phi d\theta \tag{4.49}$$

$$= \frac{\sin(\omega \frac{l_{ik}}{c})}{\omega \frac{l_{ik}}{c}} = \operatorname{sinc}\left(\omega \frac{l_{ik}}{c} \right) \tag{4.50}$$

However, for head-related systems these solutions of the integrals are not valid due to the more general definition of the propagation vector $\boldsymbol{d}$. The head can be seen as spatial filter that reduces the correlation observed between the microphone signals. More precise noise field PSD matrices can be calculated by integrating (or summing) the propagation vectors over all directions using eq. (4.39)-(4.42). In summary, the different noise field models that were used are uncorrelated noise (uncorr), cylindrical isotropic noise (diff2D), spherical isotropic diffuse noise (diff3D), integrated HRTF or HM2 (intHRTF, intHM2) and long-term measured noise from real-world recordings (measured). In the literature the noise power cross-correlation matrix is often normalized so that we get the coherence matrix

$$\Gamma_{N_i N_k}(\omega) = \frac{\Phi_{N_i N_k}(\omega)}{\sqrt{\Phi_{N_i N_i}(\omega) \Phi_{N_k N_k}(\omega)}}. \tag{4.51}$$

In free-field, the coherence matrix $\boldsymbol{\Gamma_{NN}}$ and the noise correlation matrix $\boldsymbol{\Phi_{NN}}$ are equivalent and can be used likewise. However, as the definition of $\boldsymbol{\Phi_{NN}}$ is more general for any noise field characteristic this is used in the following.

4.3.5. Algorithm Combinations

The different propagation models, output types, and algorithm settings are summarized in Table 4.2. All combinations are possible and a subset of combinations was evaluated (see section 4.4). If not stated otherwise, the noise field model intHRTF is used in combination with the propagation model HRTF, intHM2 with HM2, and diff2D with HM1 and FF.

For stability reasons of the beamformer design the noise correlation matrices $\boldsymbol{\Phi_{NN}}$ have to be mixed with a certain amount of uncorrelated noise (see section 4.4.1). This WNG constraint has been evaluated and optimized in section 4.5.

Output Type	Wave Propagation Model $\boldsymbol{d}$	Noise field model $\boldsymbol{\Phi_{NN}}$	Beamformer type
BIN_PR	HRTF	uncorr	fixed
BIN_PF	HM2	diff2D	adaptive
BIN_BL	HM1	diff3D	
MON	FF	intHRTF	
		intHM2	
		measured	

Table 4.2.: List of possible algorithm combinations

4.4. Evaluation Methods

For microphone arrays *signal-independent* measures exist to evaluate the theoretically expected performance for different noise field characteristics. These measures allow a coarse estimate of the beamformer performance and are helpful for the numerical adjustment and optimization towards the desired system properties. In this study, modifications to existing measures that are suitable for head-worn systems are suggested in section 4.4.1. For a more elaborate performance analysis, simulations with realistic signals, such as real-world recordings on a prototype array-system have to be done. These *signal-dependent* performance measures are described in section 4.4.2.

4.4.1. Signal-independent Performance Measures and the Influences Of The Head

Array Gain

The array gain is a measure that shows the improvement of the SNR between the input signal of one sensor i and the output of the array. It is defined by

$$G_i(\omega) = \frac{\mathrm{SNR}_{\mathrm{out}}(\omega)}{\mathrm{SNR}_{\mathrm{in,i}}(\omega)} \tag{4.52}$$

$$\mathrm{SNR}_{\mathrm{out}}(\omega) = \frac{\Phi_{SS}(\omega)}{\overline{\Phi}_{NN}(\omega)} \frac{|\boldsymbol{W}^H(\omega)\boldsymbol{d_S}(\omega)|^2}{\boldsymbol{W}^H(\omega)\boldsymbol{\Phi_{NN}}(\omega)\boldsymbol{W}(\omega)} \tag{4.53}$$

$$\mathrm{SNR}_{\mathrm{in}}(\omega) = \frac{\Phi_{SS}(\omega)|d_i(\omega)|^2}{\Phi_{NNi}}. \tag{4.54}$$

If the input SNRs of all microphones ($\mathrm{SNR}_{\mathrm{in},0} \ldots \mathrm{SNR}_{\mathrm{in,M-1}}$) are the same the array gain for the fixed beamformer can be calculated by

$$G_i(\omega) = G(\omega) = \frac{|\boldsymbol{W}^H(\omega)\boldsymbol{d}_S(\omega)|^2}{\boldsymbol{W}^H(\omega)\boldsymbol{\Phi_{NN}}(\omega)\boldsymbol{W}(\omega)}. \tag{4.55}$$

In this form, the array gain is only valid for a free-field or a symmetric situation and should be modified if head-shadow and diffraction effects need to be considered. The nominator of (4.55) shows the amount of signal distortion for a signal with the measured propagation vector $\boldsymbol{d}_S$ whereas the denominator shows the ability to reduce a noise field which has the measured PSD-matrix $\boldsymbol{\Phi_{NN}}$ and may differ from the assumption made for the beamformer design. subsubsectionWhite noise gain The white noise gain (WNG) is a measure that shows the ability to reduce uncorrelated (i.e., spatially white) noise. Such noise can be associated to model errors, e.g., position, amplitude, phase errors, and self-noise of the microphones and is an important robustness measure for microphone arrays. If the WNG is small the beamformer is susceptible to uncorrelated noise (and model errors), i.e., such noise is increased rather than decreased. Thus, the WNG has to be limited to a minimum δ^2:

$$\mathrm{WNG}(\omega) = \frac{|\boldsymbol{W}^H(\omega)\boldsymbol{d}_S(\omega)|^2}{\boldsymbol{W}^H(\omega)\boldsymbol{W}(\omega)} \geq \delta^2. \tag{4.56}$$

One of the most popular robust approaches to achieve this is to apply a diagonal loading [5, 17, 37] :

$$\boldsymbol{W}(\omega,\theta,\phi) = \frac{(\boldsymbol{\Phi_{NN}}(\omega)+\mu(\omega)\boldsymbol{I})^{-1}\boldsymbol{d}(\omega,\theta,\phi)}{\boldsymbol{d}^H(\omega,\theta,\phi)(\boldsymbol{\Phi_{NN}}(\omega)+\mu(\omega)\boldsymbol{I})^{-1}\boldsymbol{d}(\omega,\theta,\phi)}. \tag{4.57}$$

However, the choice of $\mu(\omega)$ that limits the WNG to a minimum of δ^2 is not simple. It can be calculated, e.g., in a trial-and-error iterative process [17] or by a scaled projection algorithm [11] which was presented in [37]. In this study, a simple iterative trial-and-error method is used, and the relevance of this constraint is studied based on the perceptual performance measures described in 4.4.2.

Directivity Index

The directivity index is a performance measure for directional microphones that shows the difference between target signal suppression and the suppression of noise coming from all directions, i.e., isotropic diffuse noise:

$$\mathrm{DI}(\omega) = 10\log_{10}\left(\frac{|\boldsymbol{W}^H(\omega)\boldsymbol{d}_S(\omega)|^2}{\boldsymbol{W}^H(\omega)\boldsymbol{\Phi_{NN}}^{\mathrm{diffuse}}(\omega)\boldsymbol{W}(\omega)}\right). \tag{4.58}$$

Note that the head can be seen as spatial filter that changes the correlation between the sensors. Thus, the correlation matrix of a diffuse noise field in free-field is different from the correlation matrix measured on a head-worn sensor-array. The head-related diffuse noise field can be estimated by integration of HRTFs from all directions (see section 4.3.4). To have a scalar performance value, the frequency dependent directivity index (DI) can be weighted by a band importance function γ_k for speech perception taken from the articulation index [27, 70]. Thus, the sum over all bands k is

$$\mathrm{DI_{AI}} = \sum_k \gamma_k \mathrm{DI}(\omega_k). \tag{4.59}$$

Spatial Directivity Pattern

The *spatial directivity pattern* or *beam pattern* is the response of the beamformer filter $\boldsymbol{W}$ to a signal arriving from the direction θ, ϕ with the wave propagation $\boldsymbol{d}(\omega, \theta, \phi)$:

$$|H_{DP}(\omega, \theta, \phi)|^2 = 10 \log_{10} |\boldsymbol{W}(\omega)^H \boldsymbol{d}(\omega, \theta, \phi)|^2. \tag{4.60}$$

For the objective evaluation of head-worn arrays, $\boldsymbol{d}$ should contain the true measured wave propagation, which is not necessarily the same $\boldsymbol{d}$ as used for the beamformer design. Another well-known definition of the beam pattern used in [5] is based on the array gain (4.55). Here, the noise correlation matrix $\boldsymbol{\Phi_{NN}}$ is replaced by the correlation matrix of a signal source in direction θ, ϕ with the true wave propagation $\boldsymbol{d}(\omega, \theta, \phi)$:

$$\boldsymbol{\Phi_{DD}}(\omega, \theta, \phi) = \boldsymbol{d}(\omega, \theta, \phi)\boldsymbol{d}^H(\omega, \theta, \phi), \tag{4.61}$$

leading to

$$|H_{BP}(\omega, \theta, \phi)|^2 = -10 \log_{10} \left(\frac{|\boldsymbol{W}^H(\omega)\boldsymbol{d}_s(\omega)|^2}{\boldsymbol{W}^H(\omega)\boldsymbol{\Phi_{DD}}(\omega, \theta, \phi)\boldsymbol{W}(\omega)} \right). \tag{4.62}$$

However, the second definition does not allow us to see target signal distortions caused by an imperfect filter $\boldsymbol{W}$.

4.4.2. Signal-dependent Performance Measures

Signal-dependent performance measures allow for a more precise performance analysis especially if calculated on real-world recordings of typical acoustical scenes. For the performance measures used here, the separated desired signal and the noise signals have been processed with the same time-varying filters that have been calculated based on the mixture. This method, sometimes referred to as a *shadow filter* or *master/slave processing* method, is only appropriate in simulation environments. Given the target and the noise signals processed separately, different signal based performance measures such as the SNRE as well as perceptual quality measures can be calculated accurately.

Signal-to-Noise Ratio Enhancement (SNRE)

The SNR-Enhancement (SNRE) is the difference of the SNR at the output of the beamformer and a reference input-SNR, both measured in dB. For binaural systems the SNRE is calculated between the left (right) output of the binaural system and the left (right) input at the reference microphone, respectively. Although there exist many modifications to this measure, e.g., by using short-time (segmental) SNRE estimates or incorporating speech importance band weighting, the linear broadband SNRE is still an appropriate measure that had shown high correlations with subjective data on the assessment of background noise reduction [62]. The SNR was calculated as the mean power of the broadband speech component on a dB scale (excluding speech pauses, i.e. signal segments with levels of -60 dB re full scale and below) minus the broadband noise power in dB. For head-worn

systems bilateral performance evaluation is relevant because a better-ear effect would be ignored by simply taking the mean SNRE.

Perceptual Similarity Measure (PSM)

The quality measure PSM from PEMO-Q [34] estimates the perceptual similarity between the processed signal and the clean speech source signal. It has shown high correlations between objective and subjective data and has been used for quality assessment of noise reduction schemes in [62, 64, 65]. PSM increases with increasing (input) SNR. As we are also interested in the quality enhancement introduced by the algorithm, we use the deduced measure ΔPSM that is calculated as the difference between the perceptual similarity measure (PSM) of the output and of the unprocessed input signal.

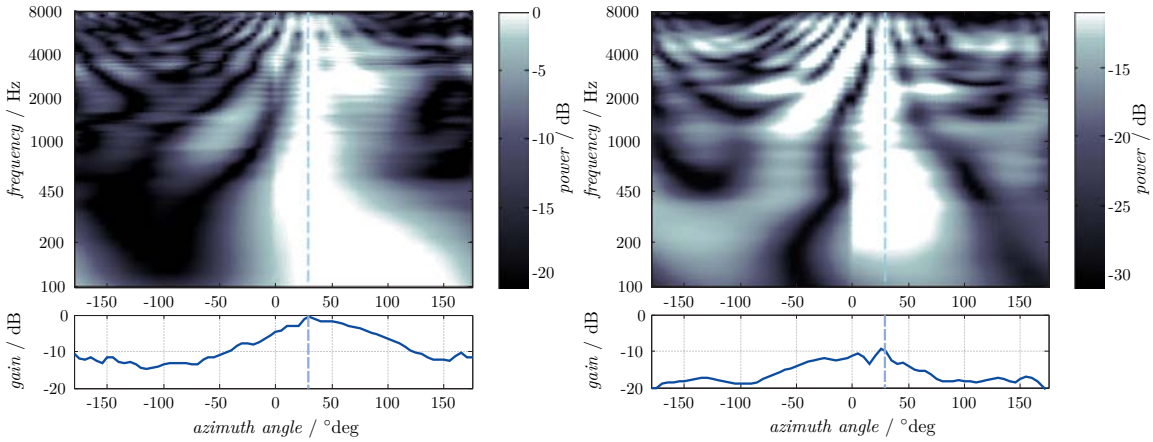
Binaural Speech Intelligibility Measure (BSIM)

The Binaural Speech Intelligibility Measure (BSIM) aims to predict the speech reception threshold (SRT) which is defined as the signal-to-noise ratio (SNR) at 50% speech intelligibility. It is calculated based on a psychoacoustic model from four input signals, the speech signal and the noise signal at the left and right ear. The model which is based on the equalization-cancellation (EC) processing by Durlach and the speech intelligibility index (SII) is described in [3]. It has shown high correlations between the individually measured speech reception threshold (SRT) for normal hearing and hearing impaired subjects and objective predictions made by BSIM. If BSIM (i.e., the estimated SRT) is lower for the output of a noise reduction scheme than for the input signal this means that the speech intelligibility has increased due to the algorithm. However, as speech intelligibility, and thus BSIM, is a nonlinear function of the SNR and other signal features such as the preservation of binaural cues, we use the difference between input and output BSIM, namely the ΔBSIM, as an indirect objective measure for the increase of intelligibility.

4.5. Experiments and Results

4.5.1. Spatial Directivity Pattern

Fig. 4.5a shows the spatial directivity pattern for the monaural output (MON) of a beamformer designed with optimal information about the HRTF for a target signal direction of 30° as observed on a head-worn microphone array. Generally, directivity patterns are plotted on a linear frequency scale. Here, we use a logarithmic scale to better emphasize attenuations that are relevant for speech perception. The best attenuation can be seen on the averted side and on the rear hemisphere. The main lobe is broad enough to be robust against target signal movement within ±10° degree over a broad frequency range. Fig. 4.5b shows the directivity pattern of a beamformer designed for free-field applications but observed on a head-worn microphone array (all other settings as for Fig. 4.5a). The distortionless response constraint of the MVDR-design is violated as the desired target direction is distorted. Furthermore, the side lobes caused by spatial aliasing are relatively strong. DI_{AI} values for the two designs are 6.3 dB and 5.0 dB, respectively.



(a) HRTF, $DI_{\mathrm{AI}} = 6.3$ dB (b) Free-field, $DI_{\mathrm{AI}} = 5.0$ dB

Figure 4.5.: Directivity pattern observed on a head-worn microphone array. a) Optimal superdirective design using head-related impulse responses (HRTF) and b) design using the free-field assumption (FF).

4.5.2. Perceptual Optimization of the White Noise Gain Limitation

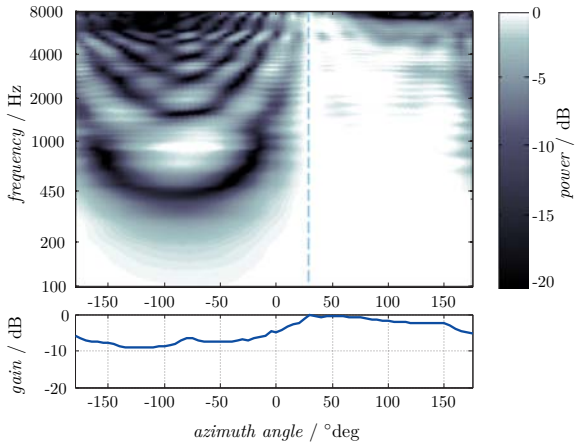


Figure 4.6.: Directivity pattern of a conventional beamformer which is optimal for the suppression of uncorrelated white noise. $DI_{\mathrm{AI}} = 3.2$ dB

A common method to raise the beamformer's robustness against various inaccuracies is to limit the WNG to a minimum δ^2 (see section 4.4.1). The WNG limitation, on the other hand, reduces the directivity and thus the noise reduction performance. The highest robustness is achieved for conventional beamformers (Fig. 4.6) with a low spatial selectivity and noise reduction performance. However, the robustness problems of *superdirective* designs (Fig. 4.5) mainly occur at low frequencies (long wavelength relative to microphone distance) where the correlation of the observed signals is high. In this frequency area, a high amplification of the superdirective beamformer is needed to guarantee a distortionless

response for the desired signal. As an unwanted secondary effect, also uncorrelated noise is increased which may lead to a degradation of the sound quality. This trade-off between directivity and robustness was evaluated with the perceptual performance measures to find the optimal setting for a prototype hearing aid array under realistic acoustic conditions. For the simulation of model errors and internal microphone noise, white uncorrelated noise was added to the microphone signals at a digital signal level of −55 dB re full scale, i.e., 30 dB below the averaged rms-level of the target speech signal observed at the microphone array. Then the beamformer coefficients were constrained iteratively based on (4.57) starting with $\mathbf{\Phi_{NN}} \triangleq$ intHRTF (intHM2, respectively) and increasing $\mu(\omega)$ for each frequency ω independently. The iteration was stopped when the frequency dependent WNG (4.56) did not exceed a predefined value of δ^2. After the processing with a fixed beamformer and the binaural post-filter (BIN_PF) the signal quality was evaluated with the performance measures PSM (averaged between left and right output) and BSIM for conditions 1 and 2. Fig. 4.7 shows BSIM and PSM as a function of the minimum white noise gain δ^2. For the HRTF (left panel) the optimal WNG limitation was found to be in the range between −24 and −17 dB. This value was consistent for both conditions and for both perceptual measures. For HM2 (right panel) the optimal values lay only slightly higher, between −22 and −15 dB. Models with lower exactness may need a stronger restriction of the WNG. In practice, the frequency dependent WNG needed to be constrained up to a frequency of 3000 Hz. Between $0 - 1000$ Hz the resulting optimized $\mu(\omega)$ was decreasing from −30 to −60 dB, between $1000 - 3000$ Hz it was fluctuating and slightly increasing to −55 dB, and for frequencies higher than 3000 Hz the resulting $\mu(\omega)$ was decreasing to −80 dB and below indicating that for these frequencies the correlation is low and no WNG constraint is needed.

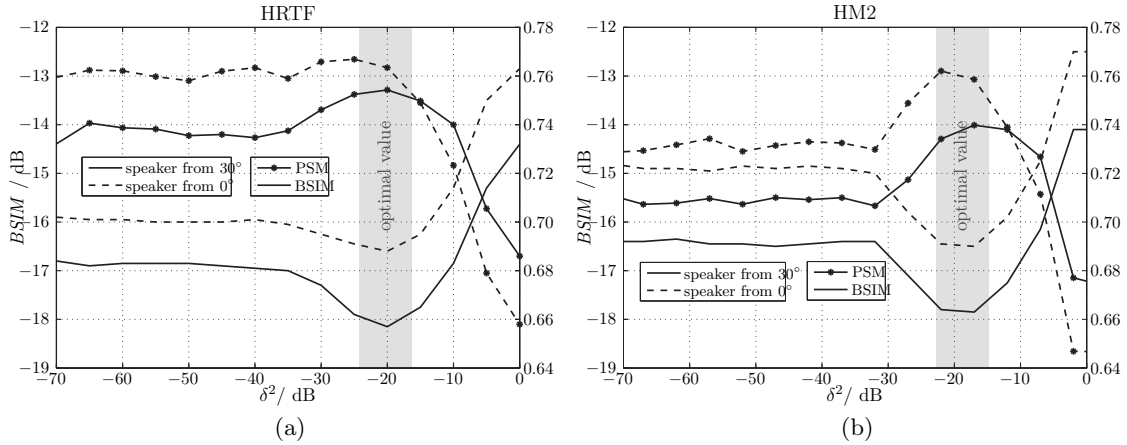


Figure 4.7.: Perceptual measures BSIM and PSM as a function of the minimum white noise gain δ^2 for the HRTF (left panel) and HM2 (right panel) assumptions for conditions 1 and 2.

4.5.3. Binaural Output Quality

	ΔBSIM	ΔPSM Left	ΔPSM Right	ΔPSM mean	SNRE Left	SNRE Right	SNRE mean
BIN_PF	7.5 dB	0.19	0.31	0.25	6.7 dB	6.1 dB	6.4 dB
BIN_PR	5.0 dB	0.14	0.32	0.23	6.1 dB	7.7 dB	6.9 dB
BIN_BL	5.2 dB	0.11	0.21	0.16	4.3 dB	4.5 dB	4.4 dB

Table 4.3.: Overall performance for beamformer processing with binaural output.

The performance values for the different binaural strategies (BIN_PF, BIN_PR, BIN_-BL) averaged over all conditions and propagation models are shown in Table 4.3. The binaural post-filter (BIN_PF) had the highest ΔBSIM and ΔPSM values and thus had the highest speech intelligibility and perceptual sound quality improvement. The binaural target signal phase reconstruction filter (BIN_PR) had a slightly higher SNRE, a similar ΔPSM mean but a lower ΔBSIM than BIN_PF as the phase of the background noise was not reconstructed and binaural information was lost. The bilateral beamformers (BIN_BL) had lower ΔPSM and lower SNRE values compared to the other binaural output systems which was caused by the smaller array size (3-microphones instead of 6). Interestingly, binaural information was partially preserved which led to a slightly higher ΔBSIM than for BIN_PR.

4.5.4. Performance Analysis of Adaptive and Fixed Beamformers

	ΔBSIM		ΔPSM		SNRE	
	fixed	adaptive	fixed	adaptive	fixed	adaptive
HRTF	8.5 dB	8.5 dB	0.28	0.27	7,1 dB	6.8 dB
HM2	8.1 dB	7.6 dB	0.27	0.27	7.0 dB	6.5 dB
HM1	7.9 dB	7.4 dB	0.25	0.26	6.4 dB	6.0 dB
FF	7.3 dB	5.1 dB	0.23	0.19	5.8 dB	5.4 dB

Table 4.4.: Performance of fixed and adaptive Beamformers

Table 4.4 shows the performance values for the fixed and the adaptive beamformer in combination with the binaural post filter (BIN_PF) which were averaged over all conditions for each propagation model. The HRTF model had a good performance for both, fixed and adaptive beamformers. The averaged ΔBSIM was $8,5$ dB for both, adaptive and fixed beamformer, as the adaptive beamformer was only better under optimal conditions (conditions 1, 2). The averaged ΔPSM (additionally averaged over left and right ΔPSM) was almost equal for adaptive and fixed beamformers if head influences were included. The head models HM1 and HM2 had a good average performance for the fixed beamformer, and only slightly lower ΔBSIM values for the adaptive beamformer. The free-field model (FF) showed a significantly lower performance than the head models in all situations and in particular for the adaptive beamformers.

4.5.5. Robustness Against Steering Errors

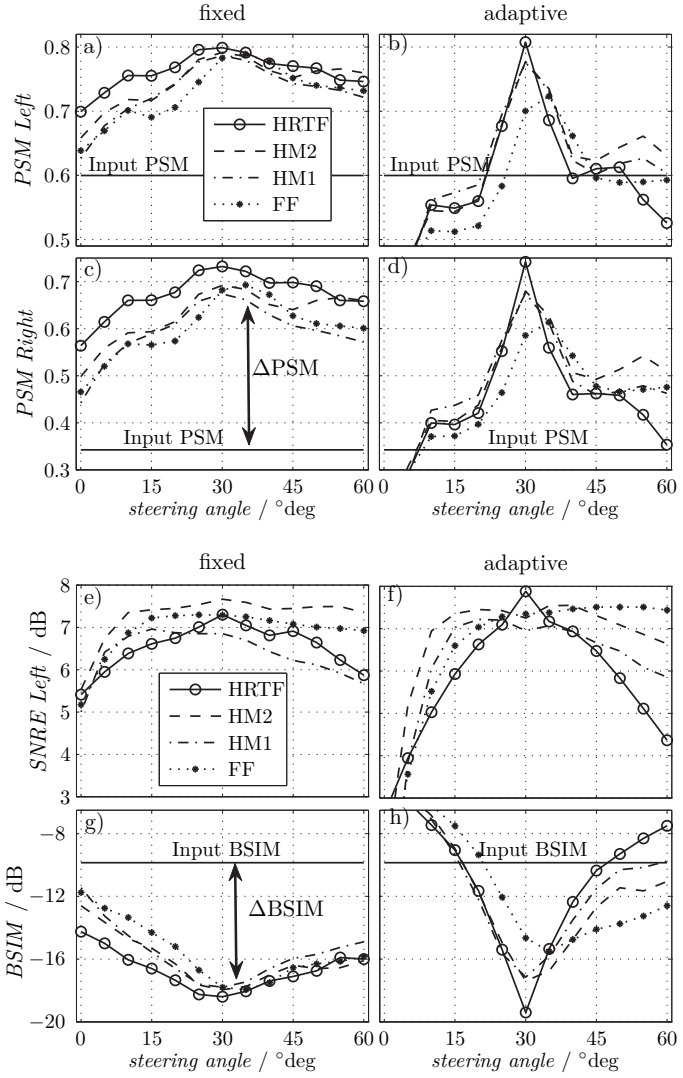
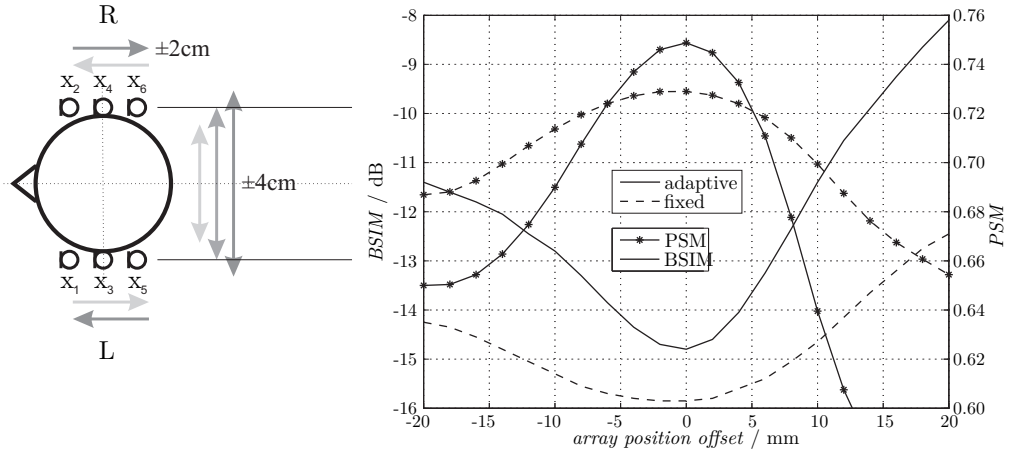


Figure 4.8.: PSM measure at left and right ear (panel a–d) and SNRE and BSIM (panel e–h) for fixed and adaptive designs as a function of the steering angle. The design target direction was 30°, the respective values at the input of the beamformer (without processing) are indicated with solid lines. Different curves denote different propagation models.

For head-worn microphone arrays it is usually assumed that the look-direction is fixed at zero degrees, and that the user always turns his or her head towards the desired signal. Thus, we are interested how the performance reduces if the signal is not exactly coming from the desired direction which can be attributed to steering errors or head-movements. Fig. 4.8 shows the robustness results for fixed and adaptive beamformers in combination

with the binaural postfilter (BIN_PF) using different propagation models in signal condition 1. The target speech signal arrives from 30°, so the best performance values should be expected if the beamformer is steered to this direction. However, for free-field beamformers (FF, ···•···) the *optimum* steering direction is dragged to greater azimuth angles because head-shadow and diffraction effects are neglected. The left panels (a–d) show the performance values for the measure PSM Left and Right. The lowest PSM values are measured for the FF propagation model, the highest values for the measured HRTFs. The PSM values for the two headmodels HM1, HM2 lie in between. The quality enhancement compared to the PSM of the input signal (ΔPSM, distance between the PSM curves and the Input PSM) was relatively robust for the fixed beamformers (panel a,c). For the adaptive beamformers (panel b,d) the PSM curves have a sharper peak at the 30° target signal direction, but for steering errors greater ± 10° the predicted signal quality falls below the input signal's quality. Interestingly, this rapid quality reduction can not be seen with the SNRE measure (see panel f for the left head-side): The enhancement of the signal-to-noise ratio does *not* reduce to 0 dB (i.e. the input SNR) within ± 10° because the measure doesn't include all the effects of the target signal distortion. Furthermore, in panel f) the ranking of the propagation models as predicted by the SNRE is questionable. Thus, the SNRE might not be appropriate for a precise performance analysis. The BSIM integrates the binaural information, therefore only one measure is needed for the evaluation of the binaural noise-reduction system. A lower speech reception threshold (SRT) can lead to a better speech intelligibility (see discussion on quality evaluation). The adaptive beamformer based on HRTF had a 2dB lower BSIM than the fixed beamformer (panel g,h), but this higher performance was reduced or even turned negative in case of imperfect steering or an imperfect propagation model.

4.5.6. Robustness Against Positioning Errors and Head Model Variation



(a) Model variation

(b) Variation of subarray position

Figure 4.9.: Robustness against variation of array position and model parameters for HM2.

In practical applications the exact positions of BTE-hearing aids may vary and the influence of the position offset on the noise reduction performance is of interest. Parametric head models are useful, because the individual transfer functions to the hearing aid microphones are usually unavailable. Thus, Fig. 4.9a schematically shows the displacements of the hearing aids that have been tested using HM2. The microphone spacing within a hearing aid shell was fixed. Fig. 4.9b shows the influence of a position offset in look-direction on the performance measures PSM and BSIM for BIN_PF in combination with HM2 in noise condition 1 for fixed and adaptive beamformers, respectively. Obviously, the fixed beamformers are robust for a shift of ±5 mm (i.e, one hearing aid was shifted 5 mm forward, the other one backward). Adaptive beamformers were slightly more susceptible to displacements, resulting in steeper PSM and BSIM curves. Additionally, different head-sizes have been analyzed which are not shown here. Measures show robustness against this type of deviation within ±1.5 cm head diameter.

4.6. Subjective Listening Test

4.6.1. Method

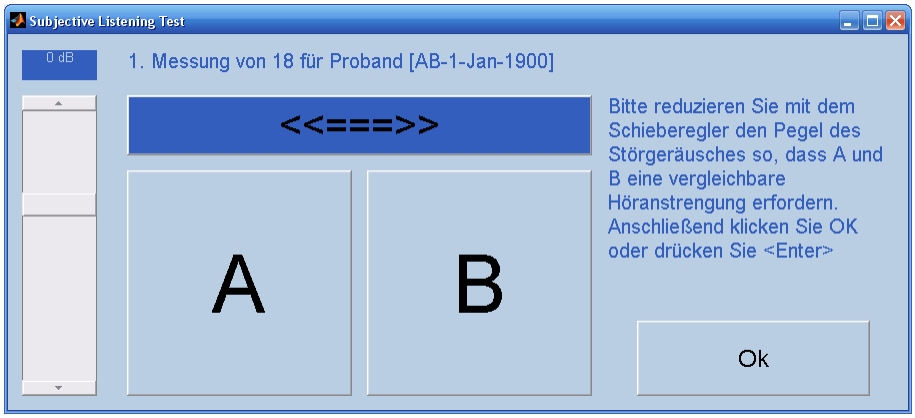


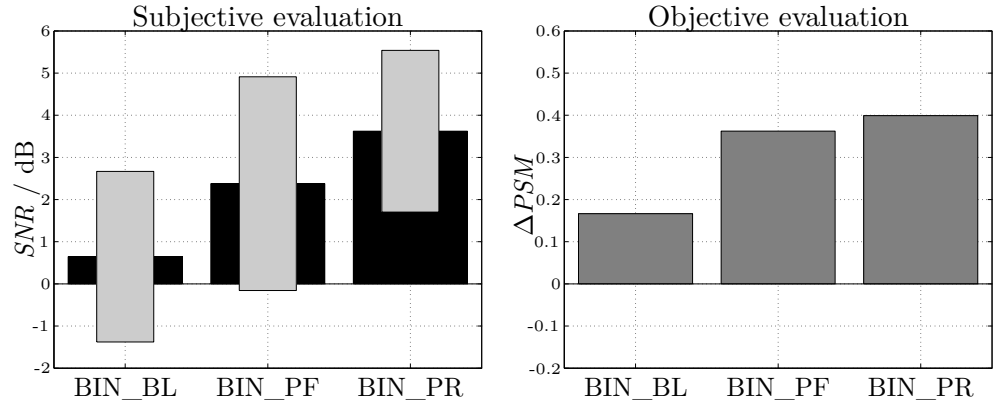
Figure 4.10.: Measuring tool: Two signals A and B are played alternately

Paired comparison tests were carried out with 10 normal hearing subjects. A pair of two signals, an unprocessed noisy speech signal A and a signal B processed by one out of three different binaural beamformer algorithms were presented to the listeners. The task was to adjust the level of the noise component in signal A by a slider so that signals A and B needed the same subjective listening effort. 18 signals where presented in a random order consisting of three successive sentences spoken by female and male german speakers (taken from the OLSA sentence test corpus, [80]) which were mixed with recorded babble noise from a cafeteria and artificial speech-spectrum-like diffuse noise. The output signals of the binaural beamformer algorithms described in 4.3 were calculated for input signals of −4 dB, −2 dB, and 2 dB SNR. Subjects were advised to make their decisions on the hearing effort in two steps. First, they should attend to the background noise and adjust the slider so that signal A had the same perceptual noise level than signal B. In a second step, this

value should be corrected so that the signals would need the same listening effort, keeping in mind that a severe distortion of the speech component in the noisy signal might reduce the ease of listening and increase the listening effort in difficult acoustical situations. The results of three trials of the listening test were averaged per subject and test signal.

4.6.2. Results

The results of the overall average for each binaural output type, bilateral beamformer (BIN_BL), binaural postfilter (BIN_PF), and binaural phase reconstruction (BIN_PR) are shown in figure 4.11a. The values show the average benefit in listening effort expressed through the ΔSNR between processed output signals and unprocessed input signals of the beamformer schemes. Negative ΔSNR values indicate that the processed signal needs a higher listening effort than the unprocessed input signal (which is undesirable) whereas positive values show a reduced listening effort that is achieved by the algorithm. As the ΔSNR values emerge from subjective assessment they may differ from the physically measurable SNRE that the algorithm may show compared to the input signal. With respect to a high standard deviation, the binaural phase reconstruction (BIN_PR) seems to achieve the lowest listening effort for normal hearing subjects although the binaural information of the background noise is not preserved with this binaural method. The order of preference could be predicted by the objective quality measure ΔPSM shown in Figure 4.11b which is discussed in section 4.7.6. A two-way repeated measures ANOVA



(a) Averaged subjective listening effort for the three binaural output types of the beamformer. Overall means and standard deviations for all subjects and conditions

(b) Results obtained by the objective quality measure PSM on the better ear side

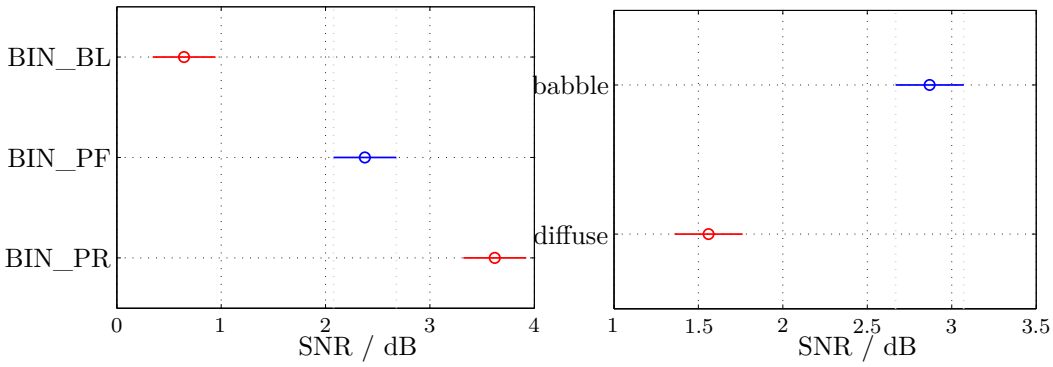
Figure 4.11.: Comparison of objective and subjective Assessment

(binaural output type [3] × background noise type [2]) was performed to check if the results where *significant* in a statistical sense and with parameters led to a significant subjective difference. The results (see Table 4.5) showed a highly significant main effect of binaural output type on the measured subjective listening effort. Also statistically significant is the

effect on the noise type, i.e., the subjective difference between diffuse and babble noise. However, no interaction was found between noise type and binaural output type. Thus, there was no algorithm that performed better only on a specific noise type than others. A low interaction between subjects and binaural output type was found which indicated that there may be groups that preferred different binaural output types. For the significant data post-hoc tests were applied to test the amount of binaural output preference (Fig. 4.12a) and the difference regarding the noise type (Fig. 4.12b). Figure 4.12a shows

Source	Prob>F
Subject	0.0141
Bin_Algo	0.0007
Noise_Type	0.0018
Subject*Bin_Algo	0.0002
Subject*Noise_Type	0.078
Bin_Algo*Noise_Type	0.3746

Table 4.5.: Results from the analysis of variance (ANOVA) test



(a) Listening effort for the binaural output types

(b) Listening effort for the noise types

that all binaural types are significantly different from the others and that all algorithms reduce the listening effort on average compared to the input signal. For BIN_BL the reduction of listening effort was below 1 dB, for BIN_PF about 2.5 dB, and for BIN_PR about 3.5 dB. This indicates that the listening effort, e.g., the output signal processed by BIN_PF has the same subjective listening effort than the input signal to the beamformer system increased by 2.5 dB in SNR. In other words, the algorithm had a *subjective* SNR enhancement of 2.5 dB.

4.7. Discussion

4.7.1. Influences of the Head and Head Models

The head influence was found to be quite small in case a unilateral array is attached to the head. It becomes particularly relevant in the case of binaural arrays, i.e., if the head is in between the coupled microphones. In summary, three aspects of the head-effect can be distinguished:
(i) The frequency-dependent noise field correlation characteristic is changed in that case so that the microphone signals can already be considered as uncorrelated at lower frequencies [17]. A mismatch between model and true noise field correlation lowers the maximum directivity that would be possible. The first beamformer algorithms that were considering head influences only modified the noise correlation matrix so that they were valid for head-related sound fields of isotropic noise. However, the practical advantage and performance increase depends on the actual sound field.
(ii) For lateral target signals the accuracy of the propagation vector $\boldsymbol{d}$ becomes relevant to avoid signal distortion. The influence of the correct choice of the propagation vector increases with increasing number of microphones and higher spatial selectivity. However, most of the microphone array systems assume that the target signal is in front (0°) of the hearing aid user. For these symmetric situations the influence of the head might be lower. On the other hand, a fixed orientation to the front direction might become inconvenient for the hearing aid user and thus may be replaced by self-steering beamformers [65] in the future. For the more general case of a lateral target direction, the head influences for the target direction are quite prominent and should not be neglected in the beamformer design.
(iii) For the binaural perception of lateral target signals head-shadow and diffraction effects play an important role because these factors influence the interaural time difference (ITD) and interaural level difference (ILD). As these binaural cues are different for the individual human head, a reconstruction of the cues after beamforming is an ambitious task. As the microphone position of the BTE-hearing aid differs from the ear vent position the inclusion of pinnae effects in the binaural reconstruction filter might be useful as well.
It is known that adaptive beamformers are principally more susceptible to model errors than fixed beamformers. Our evaluations have shown, that for a typical cocktail-party situation adaptive beamformers perform worse than fixed superdirective beamformers. Only for an optimum head model (HRTF) and an optimum choice of the propagation vector $\boldsymbol{d}$ an advantage is visible. That means that model deviations prevent the adaptive system from estimating the noise field properly, because of the interference of the estimation process with model errors.

4.7.2. Adaptive Versus Fixed Beamfomers

The results in section 4.5.4, 4.5.5, and 4.5.6 have shown that the adaptive beamformer is more susceptible to parameter mismatch (angle, head-size etc.) and deviations between the propagation model (FF, HM1, HM2, HRTF) and the true wave propagation (anechoic or office transfer functions) than the fixed beamformer. Thus model errors and parameter mismatch lead to a distortion of the target speech signal which directly reduces the

speech intelligibility. The overall-quality is also reduced but this influence can be seen particularly for strong model deviations (FF). On the other hand, even for exact models and correct steering to the target signal the performance improvement gained by adding the adaptive noise canceler path seems to be low under realistic signal conditions. Several reasons can be identified why other studies might show a greater increase of noise reduction by adapting to the noise field: (i) most adaptive beamformers are based on the standard GSC approach [28] that uses a *conventional*, low directivity beamformer in the fixed path while we use a *superdirective* beamformer. Thus, the additional noise reduction applied by the adaptive beamformer is lower. (ii) The performance strongly depends on the signal condition. In situations with only few directional interferers, low reverberation and low diffuse environmental noise a performance gain of the adaptive beamformer might be visible more clearly. Here, we used a typical cocktail-party situation with diffuse babble noise and additional directional interferers in a reverberant environment. (iii) A priori knowledge about speech pauses could be used to increase the performance. However, our test-signals contain only few short speech pauses and we believe that a reliable detection of speech pauses in babble noise with additional interfering speech is difficult.

4.7.3. Binaural Signal Reconstruction

Generally, beamformers based on the MVDR equation (4.5) provide a monaural output since all available microphone channels are summated. Note, that the target phase reconstruction filter (4.20, 4.21) is equivalent to the design of two separate MVDR beamformers for each side with the constraint of reconstructing the desired signal as observed at the left (respectively: right) side. Thus the optimal beamformer coefficients for the left (respectively the right) side, $\boldsymbol{W_L}$ ($\boldsymbol{W_R}$), are found by minimizing the expected noise field (see [5])

$$\min_{\boldsymbol{W_L}} \boldsymbol{W_L^H \Phi_{NN} W_L}, \tag{4.63}$$

and changing the constraint of an undistorted target signal from

$$\boldsymbol{W_L}^H \boldsymbol{d} = 1 \text{ to:} \quad \boldsymbol{W_L}^H \boldsymbol{d} = d_L, \tag{4.64}$$

where d_L denotes the transfer function to the left ear reference microphone. However, if the the noise phase should also be partially preserved, additional constraints are needed that go at the expense of noise reduction. This kind of trade-off between interaural cue preservation and noise reduction has been analyzed for multi-channel wiener filters in [38]. A similar finding applies to the MVDR beamformer: Its superdirectivity distorts the noise phase (and thus the noise localization cues) for the sake of a higher noise reduction. The bilateral beamformers suffer from the same effect. In particular, if the left or the right beamformer uses signals from the opposite side, additional constraints for binaural cue preservation would be needed. However, the binaural post-filter seems to be a possible way out. Here, the phases of both, desired signal and noise are kept because the filter is real-valued. On the other hand, it relies on a good speech-estimate in the beamformer output $Z(k)$. If the speech component in $Z(k)$ is already distorted due to an inexact propagation

model or parameter mismatch, this effect will be increased by the post-filter. To reduce artifacts that can be caused by the rapidly fluctuating envelope filter, the involved power spectral estimates need to be recursively smoothed. Here, a time constant of 30 ms was found by objective perceptual optimization. Additionally, modifications of the binaural post-filter could be suggested including statistical gain rules known from single-channel envelope filters.

4.7.4. Objective Perceptual Measures

ΔPSM is a suitable measure for predicting the overall perceptual quality of monaural signals. For the assessment of binaural noise reduction schemes ΔPSM has been evaluated on the each side separately. In an asymmetric situation the performance measures for the left and right head-side are quite different. For the perceptual signal quality it is unclear how PSM Left and PSM Right are integrated to a binaural *quality*. Further research into models of binaural sound quality are therefore indicated.
The BSIM measure that integrates binaural information, however, might not be able to identify the principle differences between the increase of speech intelligibility and increase of perceptual quality. $\Delta BSIM$ shows the decrease of the Speech Reception Threshold compared to the SRT of the input signal. For a linear system this measure would be equivalent to the increase of *effective* SNR (in terms of speech intelligibility) due to the algorithm. In other words $\Delta BSIM$ shows the *head-room* which is left for understanding speech. Based on the results presented both here and in the literature it can be assumed that this measure is highly correlated to a SRT-change that is achieved by the noise reduction algorithm as long as this difference is moderate.

4.7.5. Realistic Signal Conditions

Frequently, algorithms developed and evaluated under laboratory conditions do not show the expected performance under real life conditions, even if they were evaluated based on established objective measurements. Moreover, it is impossible to cover the whole range of signal conditions the algorithms are intended for in a test environment. Nevertheless, the following three aspects of performance analysis which are proposed here (and were considered in the current study) may be important to improve laboratory evaluations and their significance for real life conditions: (i) The biggest problems for hearing aid users occur for conversations in cocktail-party-like situations where the interference is babble noise which has the same average spectrum as the signal of interest. Therefore noise signals should not be stationary and should have a similar spectrum than the target signals (which was fulfilled here by, e.g., using babble noise from a cafeteria recording). (ii) Natural environmental signals are generally correlated between the microphones to a certain amount with a frequency dependent correlation function. Thus, broadband uncorrelated signals may not be appropriate test signals for natural sound sources. However, they can be used to test the susceptibility of the microphone sensors against self-noise or non-systematic random model errors (see Section 4.5.2). Furthermore, additional directional interferers may occur in reality that show a similar correlation characteristic than the signal of inter-

est. Hence, if an algorithm is tested, also directional interferers should be considered. (iii) The test signals should cover the presence of reverberation and multi-path propagation of noise and target signal. Recorded signals and real-world impulse responses are preferred as they are more realistic than simulations with, e.g., the image-source method. To check the robustness of the test-algorithm against these effects a comparison to an anechoic situation is appropriate.

If these aspects are considered in a laboratory test, algorithms will probably show a comparable performance under real-life conditions.

4.7.6. Subjective Listening Tests

The results suggest the conclusion that for normal hearing subjects the speech signal distortion is the most important factor in terms of subjective listening effort whereas the amount of noise reduction as well as the correctness of the binaural signal representation seems to play a minor role. This effect can be ascribed to the enormous efficiency of the healthy human auditory system in object segregation in noisy environments that can hardly be outperformed by technical noise reduction systems. Earlier experiments showed that the objective measure PSM from PEMO-Q has the highest correlation with subjective ratings if the *overall quality* is the decisive factor for the subjective listening effort. Consequently, this objective measure can correctly predict the order of precedence shown in Fig. 4.11b if applied to the *better-ear side*. However, the *speech intelligibility* predicted by the objective measure BSIM seems not to be the important factor for listening effort. This is plausible, because all signals were presented at SNR values clearly above the speech intelligibility threshold for normal hearing subjects (*typically* in an SNR range where real-world noise reduction systems perform best).
On the other hand, the statistical analysis indicated that there may be groups of subjects that benefit from the binaural cue preservation in (BIN_PF) in terms of listening effort. More extensive subjective listening test are needed to account for this assumption.

4.8. Conclusions

This study illustrated that a performance evaluation of multi-channel noise reduction schemes under realistic conditions is only reliable using perceptual models of the auditory system and realistic signal conditions. With the objective perceptual measures suggested here, the influence of various algorithm parameters and settings on the performance and the robustness against model errors and deviations have been analyzed. This information was used to optimize a binaural beamformer system on a trade-off scale between directivity and susceptibility to uncorrelated noise (such as self-noise of the microphones and model inaccuracies).
For head-worn binaural hearing aid systems the influence of the propagation model that was used for the beamformer design was analyzed.It was shown that systems with a high spatial selectivity such as adaptive beamformers generally have higher requirements on the exactness of the propagation model. If the deviations between the presumptions about

the wave propagation and reality are too large the design constraints of the MVDR beamformer are violated and the performance of the noise reduction system is significantly reduced. Consequently, at least coarse head models should be included for head-worn systems. The inclusion of measured (anechoic) HRTFs, however, may only have theoretical advantages that disappear in realistic echoic environments.
The binaural speech intelligibility measure (BSIM) provided an integrative measure of binaural unmasking and could identify differences in the estimated speech-reception threshold (SRT) if binaural information was distorted. The binaural post-filter technique had a good performance and could preserve binaural information on the target signal, the background noise and directional interferers. This information could be helpful for the hearing aid user for spatial object segregation. However, a definite evaluation on how the preserved binaural information is beneficial to the individual hearing impaired user has to be done based on subjective listening tests.
The integrative approach from algorithm design to user benefit in binaural hearing aids presented in this study may also be appropriate for other classes of hearing aid and man-machine communication algorithms.

The listening tests showed that the speech signal distortion is probably the most important factor in terms of subjective listening effort for normal hearing subjects. This raises the suggestion that for these listeners the perceptual similarity measure (PSM) of the better ear is the best prediction measure and that binaural information (which is not included in PSM) may play a minor role for their listening effort. However, it can be expected that hearing impaired subjects in general have other preferences and may accept a - for their hearing ability *inaudible*- speech distortion to benefit from a higher noise reduction or a binaural cue preservation. This of course needs to be analyzed in more extensive listening tests with hearing impaired subjects.

5. Combined Source Tracking and Noise Reduction for Application in Hearing Aids[1]

5.1. Introduction

Multi-microphone noise reduction schemes are promising solutions for hearing aids as they are capable of exploiting the spatial distribution of the interfering signals. Thus, they generally lead to less signal distortion and better noise reduction performance for nonstationary signals compared to single-channel noise reduction algorithms.
Binaural connections between left and right hearing aids have been investigated in chapter 4 and in recent publications [16, 47, 65] and the first hearing aids that transfer program and algorithm settings using wireless links are available on the market. It can be expected that in near future also full-band audio information will be transmitted, provided that a significant performance gain can be achieved. For binaural head-worn beamformer systems head shadow and diffraction effects become important, in particular for algorithms with a high spatial selectivity (section 4.3.4, [65]). Up to now, beamformer systems for hearing aids made the assumption that the relative target direction is at the front. However, this assumption might become unsatisfying for the hearing aid user if the signal of interest is coming from the side or is even moving (due, e.g. to head movements of the wearer). In order to overcome this problem, algorithms are necessary that track the location of the *desired* sound source and adapt the spatial noise reduction algorithm accordingly.
In this chapter different direction of arrival (DOA) estimation techniques are suggested based on the generalized cross correlation (GCC) approach by Knapp and Carter [41] and the spatial response pattern (SRP)-phase transform (PHAT) extension by DiBiase [14]. They are used in combination with a beamformer that automatically steers to the most prominent source. The importance of a proper model of wave propagation is investigated for a head-worn DOA-beamformer system. Furthermore, the performance of the system is evaluated in terms of estimation errors and signal-quality by means of objective perceptual measures that are based on models of the auditory system (see chapter 2). With these measures, the influences of inevitably occurring estimation errors can be quantified on a perceptual scale. Based on these results, the optimum compromise between algorithmic complexity and benefit can be derived.

[1] Parts of this chapter have been published as "Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays", in proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008 [65]. and as "Combined Source Tracking and Noise Reduction for Application in Hearing Aids", in proceedings of ITG-Fachtagung Sprachkommunikation, 2008 [66]

5.2. Signal Model, Recorded Signals, And Binaural Multi-Channel Noise Reduction

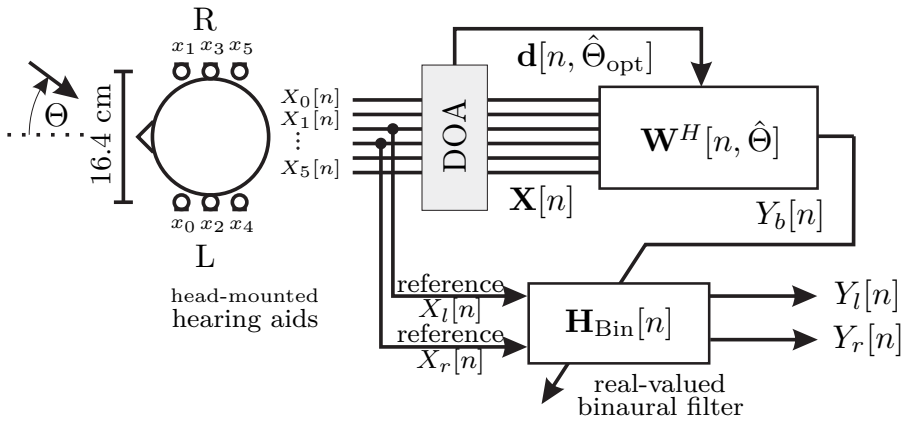


Figure 5.1.: Signal model and beamformer setup[2].

The noise reduction scheme used in this contribution is depicted in Fig. 5.1. With two 3-channel BTE hearing aid shells mounted on a Brüel & Kjær (B&K) head and torso simulator (HATS), 6-channel HRTFs were recorded in an anechoic room and in an office environment (reverberation time $\tau_{60} = 300\,\text{ms}$) for azimuth directions of the frontal hemisphere from $-90°$ to $90°$ in $5°$ steps. A moving target signal was generated by filtering a speech signal with time-varying office HRTFs and anechoic HRTFs that change due to a pre-defined virtual azimuth path (Fig. 5.2). For the office HRTFs a partitioned convolution algorithm was used. Within the $5°$ steps, the HRTFs were linearly interpolated. Real-world environmental noise was also recorded in a cafeteria (including babble, rattling dishes, and ambient noise) and in an office room (ventilation and ambient noise from outdoors through an opened window). Additionally, an artificial diffuse noise has been generated by summing up a speech-colored random noise that was filtered with HRTFs from all directions to simulate a cylindrical 2D-isotropic noise field. The moving speech signal was mixed with the noise signals at different SNRs.

	Position		
Mic. no.	x in mm	y in mm	z in mm
(1) LF	14.9	0	4.7
(2) RF	14.9	-164	4.7
(3) LM	7.3	0	2.6
(4) RM	7.3	-164	2.6
(5) LB	0	0	0
(6) RB	0	-164	0

Table 5.1.: Microphone positions in mm (compare Figure 5.1).

[2] *Notation:* Vectors and matrices are printed in boldface while scalars are printed in italic. k is the discrete time index and n the discrete frequency index. The superscripts ^{T}, *, and ^{H} denote the transposition, the complex conjugation and the Hermitian transposition, respectively.

In Table 5.1 and 5.2 the absolute positions of the microphones used in the hearing aid setup and the inter-microphone spacings are given, respectively. These show the applicability using different microphone pairs for the DOA estimation, discussed below.

	Distance in mm to microphone no.					
Mic. no.	(1) LF	(2) RF	(3) LM	(4) RM	(5) LB	(6) RB
(1) LF	-	164.0	7.9	164.2	15.6	164.7
(2) RF	164.0	-	164.2	7.9	164.7	15.6
(3) LM	7.9	164.2	-	164.0	7.7	164.2
(4) RM	164.2	7.9	164.0	-	164.2	7.7
(5) LB	15.6	164.7	7.7	164.2	-	164.0
(6) RB	164.7	15.6	164.1	7.7	164.0	-

Table 5.2.: Distances between microphones in mm. (L:left, R:right, F:front, M:middle, B:back).

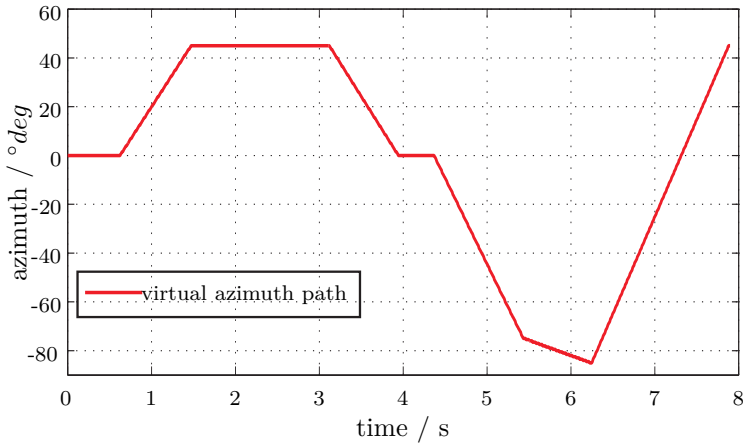


Figure 5.2.: Virtual azimuth path of a moving speech source used for time-variant convolution with anechoic and office HRTFs

In Fig. 5.1, $X_i[n]$ denotes the audio signal transformed into the frequency domain by use of the STFT, where $i = 0..5$ is the microphone channel index. A DOA detection algorithm estimates the target signal's azimuth angle $\hat{\Theta}$ which is used to steer the beamformer to this direction by means of the propagation vector $\mathbf{d}[n, \hat{\Theta}]$. The beamformer $\mathbf{W}[n, \hat{\Theta}]$ generates a single channel output $Y_b[n]$ via the well known Minimum Variance Distortionless Response (MVDR) approach [5]:

$$\mathbf{W}[n, \Theta] = \frac{\mathbf{\Gamma}_{NN}^{-1}[n]\mathbf{d}[n, \Theta]}{\mathbf{d}^H[n, \Theta]\mathbf{\Gamma}_{NN}^{-1}[n]\mathbf{d}[n, \Theta]}. \tag{5.1}$$

$$\mathbf{d}[n,\Theta] = [d_0[n,\Theta], d_1[n,\Theta], \ldots, d_{M-1}[n,\Theta]]^T \quad (5.2)$$

$$d_i[n,\Theta] = |d_i[n,\Theta]| e^{-j2\pi n \frac{f_s}{N}\tau_i[N,\Theta]},\ i = 0..M-1 \quad (5.3)$$

The fixed noise-field characteristic is coded in the coherence matrix $\mathbf{\Gamma_{NN}}[n]$ which additionally influences the beamformer properties directivity and susceptibility to white noise, and therefore has to be constrained (see section 4.5.2 and [5, 64]). Both, $\mathbf{d}[n,\Theta]$ and $\mathbf{\Gamma_{NN}}[n]$ depend on to the assumed wave propagation model which may differ from the true (and generally unknown) wave propagation from the source to the microphones. We distinguish four models: free-field (FF), two head models (HM1 [8], HM2 [18]) and the measured anechoic transfer functions from the source to the head-mounted hearing aid microphone array (HRTF).

The simplest approach is to use a free-field / far-field assumption (FF), i.e., the sound propagation is modeled as a plane wave without interfering objects in the propagation path. For FF, $\mathbf{d}[n,\Theta]$ has unity magnitude, $|d_i[n,\Theta]| = 1\ \forall(i,n,\Theta)$ and constant group delay $\tau[n,\Theta] = \tau[\Theta]$ that can be calculated from the inter-microphone distance and the angle of incidence.

For head-worn arrays it is beneficial to include knowledge about head shadow and diffraction effects [25, 64], especially for lateral target signal sources. Thus, the head models already introduced by Duda et al. [8, 18] are applied which are effective parametric models that are based on the characteristics of a sphere. In HM1, the ITD cues are modeled by Woodworth and Schlosberg's frequency independent ray-tracing formula. The gross magnitude characteristics of the HRTF spectrum, namely the ILD cues, are covered by a first order IIR head shadow filter which also accounts for an additional frequency dependent delay at low frequencies [8]. In HM2, near-field effects and interference effects that introduce ripples in the frequency response which are quite prominent on the shadowed side are incorporated as described in [18]. For both head models (HM1, HM2) the frequency dependent group delay $\tau[n,\Theta]$ and magnitude have to be calculated for each microphone and angle of incidence due to eq. (4.27)ff., [8, 18].

For HRTF, the propagation vector $\mathbf{d}[n,\Theta]$ equals the measured anechoic 6-channel HRTF for the angle of incidence Θ. $\Gamma_{NN}[n]$ can be estimated for a cylindrical isotropic diffuse noise field by integrating the propagation vectors over all directions Θ. For FF, this solution can be calculated via the Bessel function of the first kind of order zero (4.45) . For the white noise gain constraints and further details see section 4.3.4 and [5].

The binaural output is calculated by a real-valued time-varying post-filter based on [47] that is controlled by the monaural beamformer output Y_b:

$$H_{\text{Bin}}[n] = \frac{\left(|d_l[n,\Theta]|^2 + |d_r[n,\Theta]|^2\right)\Phi_{Y_bY_b}[n]}{\Phi_{X_lX_l}[n] + \Phi_{X_rX_r}[n]} \quad (5.4)$$

$$Y_l[n] = H_{\text{Bin}}[n]X_l[n] \quad (5.5)$$

$$Y_r[n] = H_{\text{Bin}}[n]X_r[n] \quad (5.6)$$

Here $X_l[n], X_r[n]$ (see Fig. 5.1) denote the reference input signals and $d_l[n], d_r[n]$ the propagation coefficients for the estimated signal direction $\hat{\Theta}_{\text{opt}}$, at the left and right reference microphone, respectively. $\Phi_{Y_bY_b}[n]$, $\Phi_{X_lX_l}[n]$ and $\Phi_{X_rX_r}[n]$ are the power spectral density

estimates for the signals $Y_b[n], X_l[n], X_r[n]$, respectively. As depicted in Fig. 5.1 we chose channel 3 and 4 as reference channels for the left and right site. For a detailed analysis of the binaural output see section 4.3.3 and [64].

5.3. Performance Of Direction Of Arrival Estimators

5.3.1. Generalized Cross Correlation Phase Transform (GCC-PHAT)

Direction of arrival estimation is done by estimating the signal delay between microphone pair $x_l[k]$ and $x_r[k]$ via the GCC-PHAT (Generalized Cross Correlation-Phase Transform) [41] which has been proven to give reliable estimates for various environments. The time delay corresponding to the estimated direction of arrival can be determined by

$$\hat{\tau} = \arg\max_k R_{x_l x_r}[k] \tag{5.7}$$

with the generalized cross correlation [41]

$$R_{x_l x_r}[k] = \frac{1}{N} \sum_{n=0}^{N-1} \Psi[n] X_l[n] X_r^*[n] e^{j\frac{2\pi}{N}nk}. \tag{5.8}$$

Typical signal delays that occur between the left and right microphones are about $8.3\mu s/1°$deg in the range of $\pm 30°$ deg. For a sampling rate of 16 kHz these are 7.5° deg per sample. Thus, an appropriate oversampling of the generalized cross-correlation $R_{x_l x_r}[k]$ is suggested.

The time-delay of arrival due to diffraction is longer for lateral signals then expected in the free-field case. Therefore the time-delay corresponds to other angles of incidence for the head models than for the free-field. Fig. 5.3 depicts deviations that occur due to a wrong delay-to-azimuth mapping. Fig. 5.3(a) shows the time delay of arrival between the microphones $x_l[k]$ and $x_r[k]$ against the azimuth angle for different propagation models. Between $\pm 30°$ the dependency is almost linear and only small deviations between the propagation models exist. For more lateral angles the differences increase due to the increased traveling time of the sound signals around the human head. In Fig. 5.3(b) the deviation of the estimated angle for the propagation model and true angle as determined from the measured HRTF is depicted. Note that for the free-field model (FF) delays beyond ± 0.5 ms are assigned to $\pm 90°$. Therefore, the azimuth error decreases for values beyond these maximum delays. The gray and black bars show the corresponding values in (a) and (b). It can be seen that the head models give a better approximation of the true time delay than FF assumptions. Although the group delays for the head models are frequency dependent [8], these effects are omitted in the GCC approach (eq. 5.7-5.8). In practice, the direction of arrival for the GCC-PHAT method is determined in three steps. First, $R_{x_l x_r}[k]$ is calculated at equidistant time samples k. Since in practice, time differences of arrival (TDOA) between two microphones are short and the interesting area is covered by only a few samples, the crosscorrelation $R_{x_l x_r}[k]$ is interpolated by an oversampled IFFT. Second, the time-delay which corresponds to the highest correlation value is found by a maximum search $\arg\max_k$. As we are interested in the correlation

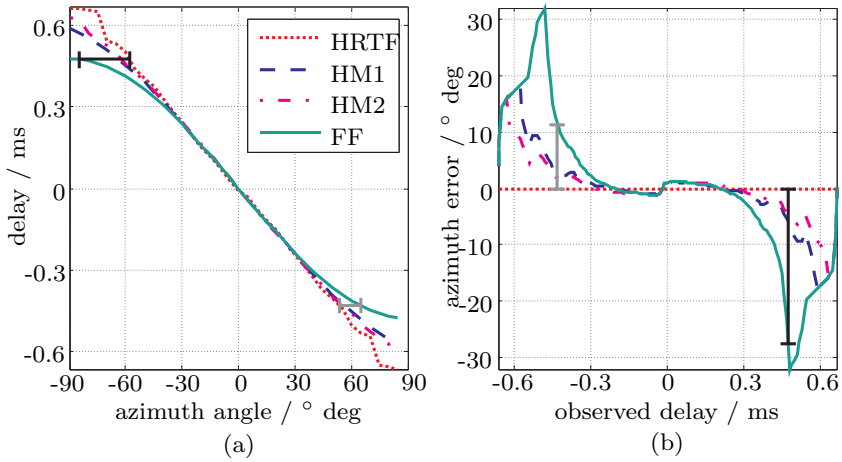


Figure 5.3.: a) "Interaural" delay as a function of azimuth for different propagation models. b) Azimuth error for different time delays τ_d and propagation models in relation to the "correct" anechoic HRTF model.

function on an equidistant azimuth angle scale, namely, the direction of arrival, in a final third step we have to re-map the time delay $\hat{\tau}$ to the azimuth angle with a non-linear mapping function which can become quite complex for the head-related case.

5.3.2. The Spatial Response Pattern (SRP-PHAT) extension

This three-step estimation method is suboptimal, because for a satisfying resolution of lateral azimuth angles between $|\Theta| = 30° \ldots 90°$, the oversampling needs to be high whereas for angles between $[-30° \ldots 30°]$ the DFT resolution is sufficient (for a microphone pair in broadside direction). Thus, by directly applying time-delays that are equidistant on the azimuth-scale only the interesting parameter space needs to be calculated. However, the trade-off between the computationally efficient IFFT usually used for the GCC-approach (eq. 5.8) and the lower dimension and higher precision (including group-delay dispersion) of the azimuth-scaled response pattern (eq. 5.9) in the SRP-approach depends on the conditions of use.

$$R_{x_l x_r}[\Theta] = \frac{1}{N} \sum_{n=0}^{N-1} \Psi[n] X_l[n] X_r^*[n] e^{j\frac{2\pi}{N} n \tau_{lr}[\Theta,n]} \tag{5.9}$$

Here $e^{j\frac{2\pi}{N} n \tau_{lr}[\Theta,n]}$ is the phase component of the inter-microphone transfer function between microphone l and r for a source signal impinging from the direction Θ. Note, that τ_{lr} may also be frequency dependent accounting for dispersion effects observed for head-worn arrays.

The DOA estimation method in (5.9) has been described by DiBiase in [14] as the spatial response phase transform (SRP-PHAT). DiBiase analyzed the performance of the

SRP-PHAT using all microphone combinations l and r.

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{i=1}^{M} \sum_{j=1}^{M} R_{x_i x_j}[\Theta] \tag{5.10}$$

Although redundancies of the correlation between microphone pairs $[l, r]$,$[r, l]$ and auto-correlations $[l, l]$ where included in the estimate, DiBiase found no detrimental effect using all combinations. However, for the microphone array used here, it was expected that some microphone pairs might not yield any information about the direction of arrival because of a very low inter-microphone distance.

As it can be seen from Table 5.2 the distances between some of the microphone pairs (e.g. microphone 1 and 3) are very small. Thus, they may be too small for a DOA estimation by means of GCC-PHAT or SRP-PHAT since real-world noise fields often are diffuse. Diffuse noise fields are highly correlated up to a certain edge frequency which is inversely proportional to the inter-microphone distances. Furthermore, the spatial positions of the microphone pair consisting of microphone 1 and 2 and the microphone pair consisting of microphone 3 and 4, e.g., are quite similar. Thus, a combination of these microphone pairs may provide only little more information about the desired signal. For the experiments a subset of all possible microphone pairs was used and different combinations were evaluated compared to a single microphone pair.

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{p=1}^{P} R_{x_{p,1} x_{p,2}}[\Theta] \tag{5.11}$$

Here p is number of the actual microphone pair and P is the number of microphone pairs used. For the investigated hearing aid system, the DOA could theoretically be estimated from $P_{\max} = \frac{(M-1)\cdot M}{2}$ pairs where $M = 6$ is the number of microphones. However, due to the constraints discussed above, a number of $P \leq 3$ is realistic.

5.3.3. Source Tracking Constraints

Figure 5.4 shows the correlation pattern seen for a moving signal under office ambient noise conditions at and SNR of 8 dB. In the algorithms presented here, only a single moving source in the frontal hemisphere $(-90 \ldots 90)^\circ$ is being tracked. The maximum tracking speed of the DOA estimator is limited to $125^\circ/s$ as described in [25] to avoid sudden peaks in the DOA estimate that lead to severe disturbances of the subsequent beamformer. A simple speech activity detector based on the magnitude of $R_{x_l x_r}$ is applied by updating the DOA estimate only if $R_{x_l x_r}$ is greater than a threshold ξ. During speech pauses (occurring in Fig. 5.4at around 3s and 5.5s) the tracking algorithm continues the update of the azimuth estimate based on the gradient of the last estimates. The gradient was calculated by a simple regression over the last 10 estimates (approximately 80 ms memory). However for the application in a hearing aid it might be useful to apply more sophisticated tracking algorithms (including multiple source tracking) that increase the

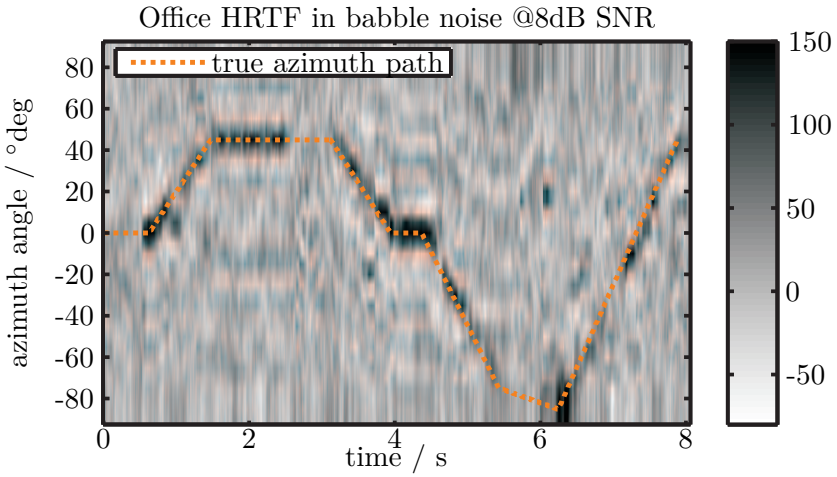


Figure 5.4.: Correlation Pattern $R_{x_l x_r}[\Theta]$ eq. (5.9) of SRP-PHAT over time. At 8 dB the maximum tracking is relatively reliable. However, outlier and correlation patterns induced by interfering (correlated) background noise degrade the estimate.

robustness of the estimate while at the same time allowing for a quick change of direction due to a moving speaker.

5.3.4. DOA Estimation Error

In Figure 5.5 the mean DOA estimation error $\bar{e}_\Theta = \frac{1}{|\mathcal{A}|} \sum_{\mathcal{A}} \Theta - \hat{\Theta}$ is shown dependent on the input SNR for the GCC-Phat method (subplots a) and d)), the SRP-PHAT method using one microphone pair (subplots b) and e)) and the SRP-PHAT method using two microphone pairs (subplots c) and f)). Here, Θ and $\hat{\Theta}$ are the true and the estimated direction of arrival, respectively. $\mathcal{A}$ is the set of frames where speech is present and $|\mathcal{A}|$ its cardinality. The left subplots (a)-c)) show the performance of the algorithms for an anechoic situation (no reverberation) in diffuse noise conditions while the right subplots (d)-f)) show the performance of the DOA estimators in a reverberant environment ($\tau_{60} \approx 300$ms) and babble noise conditions. Different propagation models (free-field (FF), head models (HM1, HM2) and measured HRTFs) were evaluated. It can be seen that, in general, the algorithms perform best for measured HRTFs and worst if no diffraction and shadowing effects are incorporated into the design (FF). Using head models is a good approximation for the measured HRTF which is unknown in practical systems.
Comparing the GCC-PHAT and SRP-PHAT curves reveals that the SRP-PHAT algorithm performs slightly better. However, the averaging over multiple microphone pairs did not lead to the expected performance improvement (particularly not for real-world conditions including reverberation and high environmental noise) that was reported in literature. It was found that a small variance decrease of the correlation matrix $R_{x_{p,1} x_{p,2}}$ could be seen for an ideal diffuse noise field. Looking at real-world recorded babble noise this small effect disappeared because the noise had a stronger correlation which was seen by all microphone

pairs simultaneously.

In summary it can be stated that the SRP-PHAT algorithm using only one microphone pair (5.9) showed the best performance for the given microphone setup.

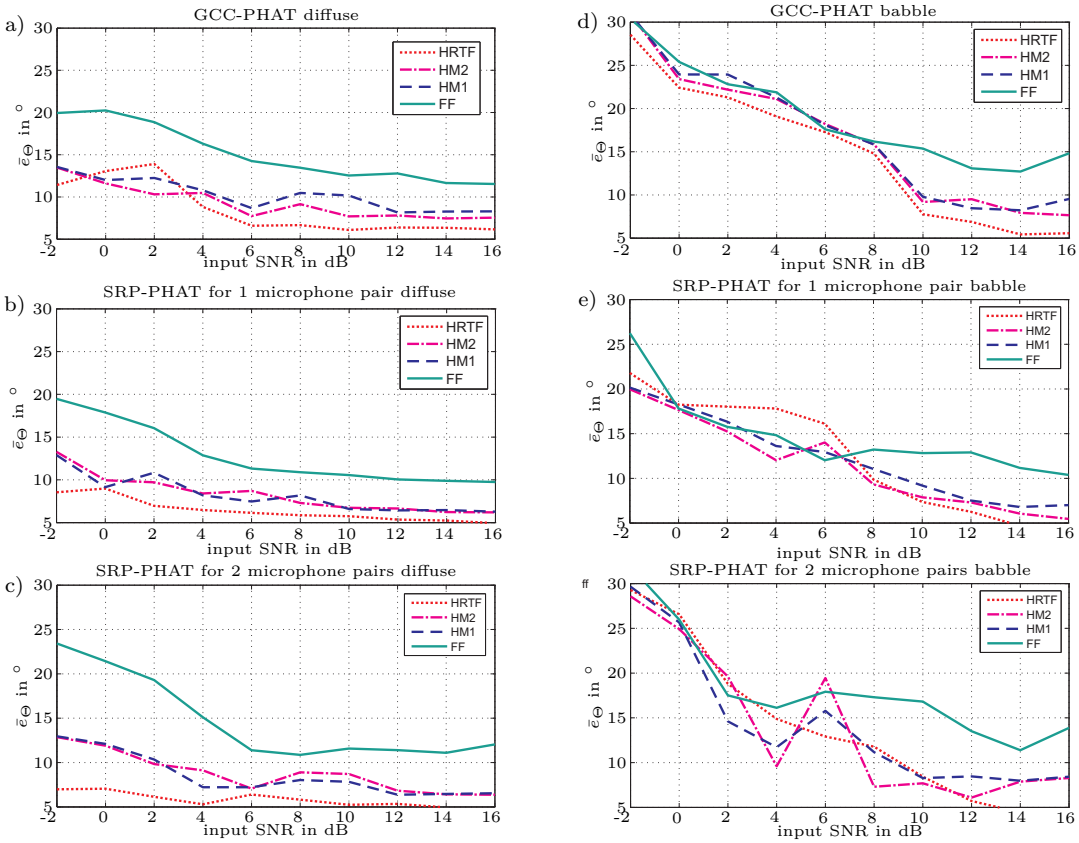


Figure 5.5.: Mean DOA estimation error over input SNR for diffuse noise conditions without reverberation (upper part, a)-c)) and in reverberant environment ($\tau_{60} \approx 300$ms) and babble noise conditions (lower part, d)-f)) for GCC-PHAT (a), d)) and SRP-PHAT for 1 microphone pair (b), e)) and 2 microphone pairs (c), f)).

5.4. Objective Quality Assessment for the Complete Noise Reduction System

It has been shown in Fig. 5.3 that the assumption of an imperfect propagation model leads to systematic errors in the estimation of the signal-source direction. As we are interested in the influence of these estimation errors on the performance in combination with a spatial noise reduction algorithm and the resulting signal quality for realistic scenarios, we propose three performance measures. They all estimate the benefit a subject will receive

from a binaural noise-reduction scheme (described in 5.2) that utilizes any of the DOA estimators including head models described so far.
SNRE: The SNR-Enhancement (SNRE) is the difference of the SNR at the output of the beamformer and a reference input-SNR, both measured in dB. For binaural systems the SNRE is calculated between the left (right) output of the binaural post-filter and the left (right) input at the reference microphone, respectively; by simply taking the mean SNRE a better-ear effect would be ignored.
PSM / ΔPSM: The quality measure PSM from PEMO-Q [34] estimates the perceptual similarity between the processed signal and the clean speech source signal. It has shown high correlations between objective and subjective data and has been used for quality assessment of noise reduction schemes in [62–64]. PSM increases with increasing (input) SNR. As we are interested in the quality enhancement introduced by the algorithm, we use the deduced measure ΔPSM that is calculated as the difference between the perceptual similarity measure (PSM) of the output and of the unprocessed input signal.
Binaural Speech Intelligibility Measure BSIM / ΔBSIM: The speech reception threshold (SRT) is defined as the signal-to-noise ratio (SNR) at 50% speech intelligibility. In [3] a binaural model of speech intelligibility based on the equalization-cancelation (EC) processing by Durlach had been defined which is able to predict the SRT with high accuracy. This objective measure, described in section 2.4.5, is denoted as BSIM in the following. If BSIM of the output of a noise reduction scheme is lower than for the input signal this means that the speech intelligibility has increased due to the algorithm. However, as the speech intelligibility (and BSIM) are nonlinears function of the SNR and other signal features such as the preservation of binaural cues, we use the difference between output and input BSIM, namely the ΔBSIM, as an indirect measure for the increase of intelligibility. BSIM as described in [3, 64] assumes a spatially stationary source configuration. To be applicable to moving sources it had to be extended to a block-wise measure with subsequent averaging across blocks.

5.5. Objective Perceptual Quality Results for the Combined System

Figure 5.6 shows the performance of the combined SRP-PHAT-steered noise reduction system evaluated in a reverberant office environment by three objective measures: the signal-to-noise ratio enhancement SNRE, the perceptual similarity measure ΔPSM from PEMO-Q (section 2.4.2,[34, 65]) and the binaural speech intelligibility measure ΔBSIM [3, 65], see section 2.4.5.

All three measures show relative enhancements compared to the unprocessed signal and are plottet over the SNR of the input signal (SNR_{in}). With increasing SNR_{in} the relative performance enhancement decreases which is a fact common to all noise reduction systems as for infinite SNR_{in} no further improvement is possible. The ideal system (solid black line with markers) has *a priori* information about the direction of arrival and uses the measured anechoic HRTF as a propagation model. Therefore, it should set the upper performance limit. The non-steered (0° fixed) system (light dashed gray line) has no information on the direction of arrival. It is fixed to the 0° look direction (just like traditional directional

hearing aids) and uses the measured anechoic HRTF as a propagation model. It marks the decision criteria, from where a combined self-steering beamformer system has a higher performance than a traditional non-steered system.

The broadband SNRE shows the amount of noise reduction in a technical sense. Although it is an established measure which is correlated with the perceived amount of noise reduction (section 3.2) it has some severe deficiencies as signal distortions may not be seen properly. Thus inconsistencies may occur, e.g., in Figure 5.6 a) the estimated system is better than the optimal system. In subplot b), ΔPSM shows the increase of the estimated perceptual quality compared to the input signal. Both, SNRE and PSM are monaural measures and therefore are evaluated for left and right output signal, respectively. The difference between the input and output speech reception threshold is estimated by ΔBSIM in Figure 5.6 c). This measure integrates binaural information that might be used by the listener for localization and object segregation. The ΔBSIM plot shows that the performances of the self-steered systems with included head-models converge to the ideal system at about $8 - 10$ dB input SNR. A ΔBSIM value of -4 dB means, e.g., that the expected speech reception threshold (i.e., 50% speech intelligibility) of the binaural output is 4 dB lower than the speech reception threshold estimated for the input signal. Thus ΔBSIM can be interpreted as the amount of additional *head-room* of speech intelligibility achieved by the binaural noise reduction scheme.

The performance of the beamformer designed for free-field was much lower than for the head-model based designs. This is partly due to the white noise gain constraint which is an important factor that influences the amount of noise reduction and that had been optimized for the head-worn array. However, it has already been shown in [64, 65]and in section 4 that free-field beamformers are suboptimal for head-worn arrays.

5.6. Discussion

The study showed that theoretical advantages (such as, e.g., using 6 microphones instead of 2) vanish as soon as realistic signals are involved. This is primarily because realistic noise fields are correlated to some extent and thus apply additional interfering correlation to the spatial response pattern, which can not be averaged out by the combination of multiple microphone pairs. Hence, microphone pairing for the present behind-the-ear (BTE) hearing aids was not beneficial.

With the SRP-PHAT method sound propagation models can be more complex compared to the GCC-PHAT, as dispersive group-delays can be integrated in the calculation of the spatial response pattern. Additionally, the SRP-PHAT method has a higher flexibility to apply a direction of arrival search only within a spatial search area of interest or to use variable angular resolution. This leads to a reduced parameter space for the subsequent maximum tracking. On the other hand, the IFFT that can be used for the GCC-PHAT is more efficient in terms of computational complexity. Under the conditions used here, the SRP-PHAT showed slightly better DOA estimation results, hence it was used for the evaluation of the combined DOA-beamformer system. However, computational complexity was not investigated in this study.

For the evaluation of the combined binaural system, the objective measure ΔBSIM

has the advantage that it has only a single performance value whereas the monaural measures ΔPSM and SNRE need to be evaluated for both sides separately. Hence, using the monaural measures for algorithm optimization the developer has to decide which side is more relevant for the overall signal quality. Compared to the SNRE measure, ΔBSIM leads to more robust results indicated by a system ranking as expected (i.e., optimal system can not be outperformed).

The maximum performance gain measured with ΔBSIM between the traditional non-steered and optimally steered system was approximately 1.5dB. For the head model based systems the performance lay in between the non-steered system and the optimally steered DOA-beamformer system for the total analyzed input SNR range from -2 dB to 16 dB. For an input SNR $>$ 8dB the performance difference between the optimal and the estimated head-model based systems was negligible. Hence, an automatic steering was beneficial for the signal conditions analyzed here.

5.7. Conclusion

The results show that the SRP-PHAT method using only a subset of microphone pairs leads to a slightly higher performance than the GCC-PHAT method. The advantage of using the SRP-PHAT is a higher precision and flexibility in the sampling of the parameter space with equidistant azimuth angles and no need for a TDOA-mapping. For the microphone array used here, a combination of multiple microphone pairs did not lead to a consistent improvement compared to a single microphone pair. This can be explained by the low inter-microphone distances and the similarity of the correlation patterns for those microphone pairs that are applicable to DOA estimation. The combination of SRP-PHAT based direction of arrival estimation and a constrained superdirective beamformer showed that the objectively estimated signal quality and the estimated speech intelligibility were improved compared to traditional non-steered systems, if at least a rudimentary head model was included in the DOA estimation algorithm and in the binaural noise reduction scheme for head-worn microphone arrays.

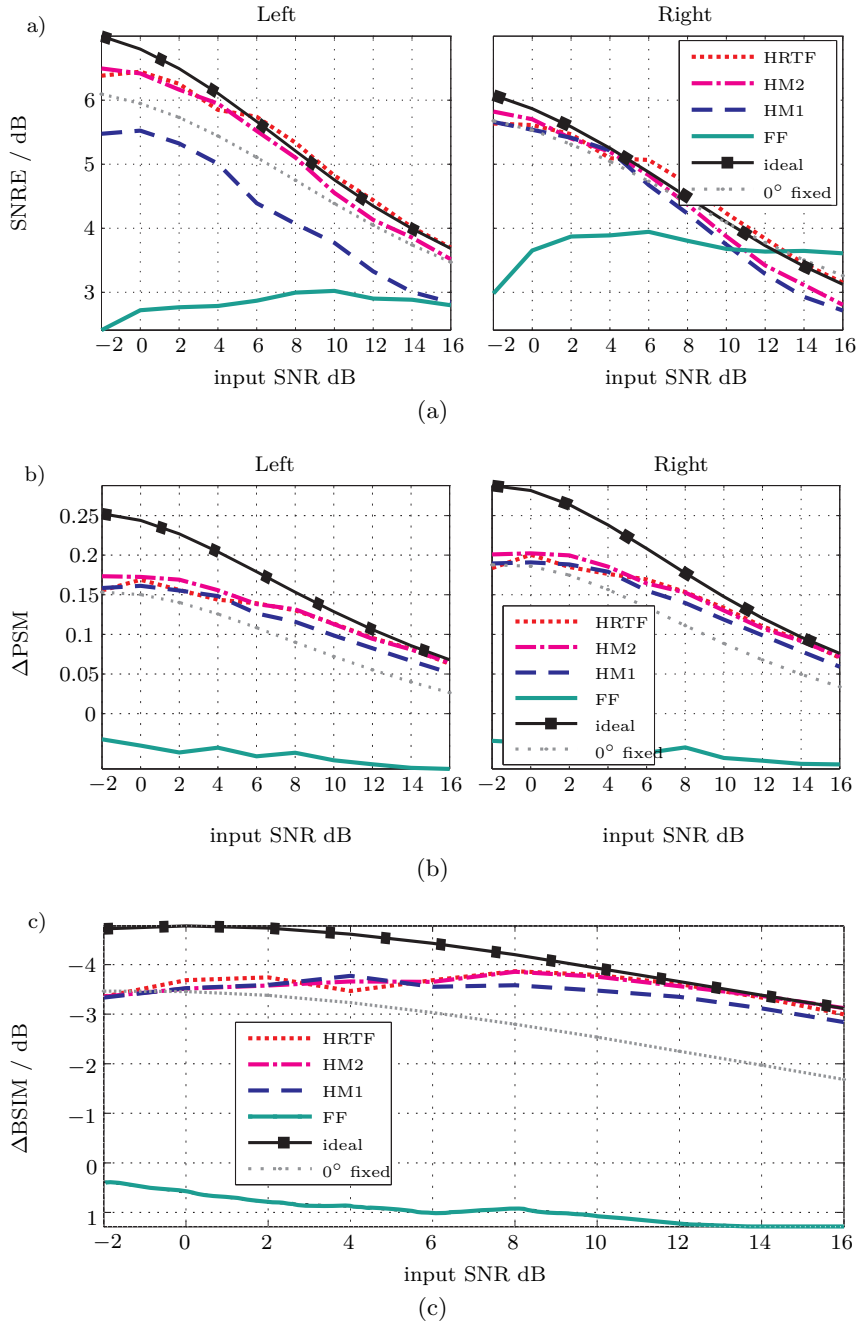


Figure 5.6.: Performance evaluation with the objective measures SNRE, ΔPSM and ΔBSIM for a speech signal in a reverberant office environment ($\tau_{60} = 300$ ms) mixed with babble noise at different input SNR's and processed by the SRP-PHAT-steered binaural noise reduction scheme. Data are presented such that an improvement in performance is pointing upwards.

6. Conclusions and Further Research

The current work contributes novel monaural and binaural noise reduction algorithms for applications in hearing aids and modern communication systems, that can improve speech communication in adverse conditions. The proposed algorithms account comprehensively for real-life conditions that are characterized by non-stationary noise, interfering speakers, reverberation, diffraction effects for head-worn devices as well as movements of the source signals and head movements. While other studies usually evaluate performance using only a subset of these conditions, this study provides meaningful evaluation methods with recorded signals from real listening environments and investigates the major influencing factors in combination.

The development and parameter optimization of single and multi-channel noise reduction schemes is supported by objective evaluation measures, that are based on recent knowledge of the human auditory system, including novel binaural measures. These psychoacoustic measures and other widely-used performance measures have been compared to subjective data in the aforesaid listening conditions and the measures with the highest significance on perceptual signal quality have been identified. Hence, this work also provides a toolbox for performance evaluation that is useful for the development of noise reductions schemes in general.

The study on monaural noise reduction schemes (chapter 3) clearly showed the limitations that single-channel noise reduction algorithms have when used in realistic fluctuating non-stationary noise. From the results shown in this work it can be concluded that in most cases this algorithm class leads to unexpected degradation of the desired signal, and hence cannot be recommended for use in case of strong non-stationary noise backgrounds.

On the other hand, multi-channel noise reduction algorithms were shown to provide a significant signal enhancement in adverse conditions if certain design criteria are met (chapters 4 and 5). Binaural outputs that preserve the spatial configuration of all signals are feasible without compromising noise reduction. For these adaptive binaural beamformers head-influences need to be incorporated into the algorithm design. Fortunately, parametric head-models have shown convincing results, so that individual head related transfer functions are not needed, even if they outperform the *generic* head-model based HRTFs in certain conditions. The latter finding facilitates the use of these algorithms in commercial audio devices and hearing instruments.

To account for head movements and moving target signals, self-steering binaural beamformer algorithms have been proposed (chapter 5) and evaluated in realistic conditions. For a single moving speech signal in a noisy and reverberant environment, these combined direction-of-arrival (DOA) estimation and beamformer schemes have shown higher performance compared to non-steered systems. These results encourage the use of these self-steering beamformers as a future application in binaural hearing systems.

Because the human auditory system is perhaps the most complicated and less thoroughly

understood human sense, only a few research problems could successfully be tackled here. In terms of noise reduction the *definition of noise* is extremely difficult and subjective - a fact on which any noise reduction algorithm must fail somehow. Algorithms that perform a Computational Auditory Scene Analysis (CASA) and leave the decision about the desired *audio objects* to the user are highly desirable. However, CASA based algorithms are computationally demanding and currently not applicable for real-time processing in hearing aids. Moreover, not all effects of the auditory perception are understood and much less is the resynthesis of *internal representations* to an audible signal. Hence, signal enhancement and noise reduction in hearing aids and mobile communication devices presently remains a domain on empirical and more *technically* based approaches that are optimized with measures based on auditory models. However, it is conceivable that these approaches will gradually be transformed into more theory-driven approaches if our understanding on how the human brain processes these complex situations is improved and translated into application-oriented computational models.

To conclude, this work provides a comprehensive survey of objective performance assessment and development of noise reduction schemes for application in hearing aids and modern communication systems and provides a solid scientific basis for existing and new noise-reduction algorithms in terms of evaluation and applicability.

A. Tables

A.1. Table of Articulation Index

The Articulation Index (AI) was first described by French and Steinberg (1947) [24] as a way to express the amount of average speech information that is available to patients with various amount of hearing loss. It is usually described as a number between 0 and 1.0 or as a percentage, 0% to 100%. The AI can be calculated by dividing the average speech signal into several bands and obtaining an importance weighting for each band. Based on the amount of information that is audible to a patient in each band and the importance of that band for speech intelligibility, the AI can be computed.

Band Number	Center Frequency (Hz)	Weight	Band Number	Center Frequency (Hz)	Weight
1	50	0.003	14	1148	0.032
2	120	0.003	15	1288	0.034
3	190	0.003	16	1442	0.035
4	260	0.007	17	1610	0.037
5	330	0.010	18	1794	0.036
6	400	0.016	19	1993	0.036
7	470	0.016	20	2221	0.033
8	540	0.017	21	2446	0.030
9	617	0.017	22	2701	0.029
10	703	0.022	23	2978	0.027
11	798	0.027	24	3276	0.026
12	904	0.028	25	3597	0.026
13	1020	0.030			

Table A.1.: Table of articulation index in 1/3 octave filters (source: [59])

A.2. Critical Bandwidth / Equivalent Rectangular Bandwidth (ERB)

The equivalent rectangular bandwidth (ERB) is a psychoacoustic measure to approximate the *critical bandwidth* of auditory filters in human hearing [53]. It uses the simplification of modeling filters as rectangular band-pass filters. An ERB passes the same amount of energy as the auditory filter it corresponds to and shows how it changes with input frequency. The ERB scale is very similar to the Bark-scale [86] (and also the mel-scale) their differences are due to different psychoacoustic measuring methods and standardizations

(e.g., "notched-noise", "within-band masking" or "pitch-perception" methods). The ERB is calculated by the following equation:

$$\mathrm{ERB}(f) = l + \frac{f}{q}, \qquad \text{with} \quad l = 24.7, \quad q = 9.265. \tag{A.1}$$

To model the frequency resolution on the basilar membrane a gammatone filterbank [32, 54] can be built using band-spacing of 1 ERB. In the following, the 14th band of this auditory filterbank is related to 1000 Hz. To have integer band numbers b an auxiliary constant c_{1kHz} is defined by:

$$c_{1kHz} = q \cdot log\left(1 + \frac{1000}{l \cdot q}\right) - 14 = 1.5725. \tag{A.2}$$

The center frequency f_c of the bth auditory filter now is:

$$f_c(b) = \left(\exp\left(\frac{b + c_{1kHz}}{q}\right) - 1\right) \cdot l \cdot q \tag{A.3}$$

and the equivalent rectangular bandwidth of this auditory filter b is

$$\mathrm{ERB}(b) = l \cdot \exp\left(\frac{b + c_{1kHz}}{q}\right) \tag{A.4}$$

For a given center-frequency f_c in Hz the auditory filter ("ERB-number") b is defined by:

$$b(f_c) = q \cdot log\left(1 + \frac{f_c}{l \cdot q}\right) - c_{1kHz}. \tag{A.5}$$

For reasons of efficiency or the availability of linear frequency short-time spectral estimates the frequency grouping of the auditory filter is often approximated by summing up (non-overlapping, linear) frequency bands around a center-frequency on the ERB (or Bark) scale. An example is given in eq. (2.7).

B. Acronyms

BTE	behind-the-ear
DFT	discrete Fourier transform
DOA	direction of arrival
GCC	generalized cross correlation
GSC	generalized sidelobe canceller
HRTF	head related transfer function
ILD	interaural level difference
ITD	interaural time difference
MMSE	minimum mean squared error
MVDR	minimum variance distortionless response
PSM	perceptual similarity measure
SNR	signal-to-noise ratio
WNG	white noise gain
WSS	weighted spectral slope
SNRE	signal-to-noise ratio enhancement
STFT	short time Fourier transform
STSA	short time spectral attenuation
SRT	speech reception threshold
D&S	delay-and-sum
SII	speech intelligibility index
FFT	fast Fourier transform
FIR	finite impulse response
SRP	spatial response pattern

PHAT phase transform

MWF multi-channel Wiener filter

MOS mean opinion score

Bibliography

[1] ANSI, "Methods for the calculation of the speech intelligibility index," American National Standards Institute, American National Standards Institute, New York, Tech. Rep. ANSI S3.5-1997, 1997.

[2] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE International Conference on ICASSP '79. Acoustics, Speech, and Signal Processing*, vol. 4, 1979, pp. 208–211.

[3] R. Beutelmann and T. Brand, "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America (JASA)*, vol. 120, no. 1, pp. 331–342, 2006.

[4] J. Bitzer, "Mehrkanalige Geräuschunterdrückungssysteme - eine vergleichende Analyse," Ph.D. dissertation, Universität Bremen, 2001.

[5] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, ch. 2, pp. 19–38.

[6] J. Bitzer, U. Simmer, I. Holube, and T. Schaer, "Some experiments on short-time spectral attenuation (STSA) algorithms and speech intelligibility," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2005.

[7] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[8] P. C. Brown and R. O. Duda, "A structural model for binaural sound synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 476–488, Sep 1998.

[9] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[10] J. Collura, "Speech enhancement and coding in harsh acoustic noise environments," in *Proc. IEEE Workshop on Speech Coding*, 20–23 June 1999, pp. 162–164.

[11] H. Cox, R. M. Zeskind, and T. Kodu, "Practical supergain," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 3, pp. 393–398, Jun. 1986.

[12] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. i. detection and masking with narrow-band carriers." *Journal of the Acoustical Society of America (JASA)*, vol. 102, no. 5 Pt 1, pp. 2892–2905, Nov 1997.

[13] J. Desloge, W. Rabinowitz, and P. Zurek, "Microphone-array hearing aids with binaural output .i. fixed-processing systems," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 6, pp. 529–542, Nov 1997.

[14] J. DiBiase, "A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays," Ph.D. dissertation, Brown University, Providence,Rhode Island, USA, May 2000.

[15] S. Doclo, T. J. Klasen, T. Van den Bogaert, J. Wouters, and M. Moonen, "Theoretical analysis of binaural cue preservation using multi-channel wiener filtering and interaural transfer functions," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2006.

[16] S. Doclo, T. van den Bogaert, J. Wouters, and M. Moonen, "Comparison of reduced-bandwidth mwf-based noise reduction algorithms for binaural hearing aids," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 223–226.

[17] M. Dörbecker, "Mehrkanalige Signalverarbeitung zur Verbesserung akustisch gestörter Sprachsignale am Beispiel elektronischer Hörhilfen," Ph.D. dissertation, Rheinisch-Westfälische Technische Hochschule Aachen, 1998.

[18] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *Journal of the Acoustical Society of America (JASA)*, vol. 104, no. 5, pp. 3048–3058, 1998.

[19] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *Journal of the Acoustical Society of America (JASA)*, vol. 35, no. 8, pp. 1206–1218, 1963.

[20] G. W. Elko, "Superdirectional microphone arrays," in *Acoustic signal processing for telecommunication*, S. L. Gay and J. Benesty, Eds. Hingham, MA, USA: Kluwer Academic Publishers, 2000, pp. 181–238.

[21] G. Elko and A.-T. N. Pong, "A simple adaptive first-order differential microphone," in *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 15–18 Oct. 1995, pp. 169–172.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean–square error log spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[23] S. Fischer and K. U. Simmer, "Beamforming microphone array for speech acquisition in noisy environments," *Speech Communication*, vol. 20, pp. 215–227, 1996.

[24] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[25] S. Goetze, T. Rohdenburg, V. Hohmann, B. Kollmeier, and K.-D. Kammeyer, "Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays," *Intelligent Signal Processing and Communication Systems, 2007. ISPACS 2007. International Symposium on*, pp. 84–87, 28 2007-Dec. 1 2007.

[26] J. Greenberg and P. Zurek, "Microphone-array hearing aids," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. Springer, 2001, ch. 11, pp. 229–249.

[27] ——, "Evaluation of an adaptive beamforming method for hearing aids," *Journal of the Acoustical Society of America (JASA)*, vol. 91, no. 3, pp. 1662–1676, mar 1992.

[28] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Antennas Propagation*, vol. 30, pp. 27–34, 1982.

[29] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Sydney, Australia, Dec. 1998.

[30] M. Hansen and B. Kollmeier, "Objective modelling of speech quality with a psychoacoustically validated auditory model," *Journal of the Audio Engeneering Society (JAES)*, vol. 48, pp. 395–409, 2000.

[31] M. Hansen, "Assessment and prediction of speech transmission quality with an auditory processing model," Ph.D. dissertation, Universität Oldenburg, Oldenburg, Jun. 1998.

[32] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta acustica / Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

[33] R. Huber, "Objective assessment of audio quality using an auditory processing model," Ph.D. dissertation, University of Oldenburg, 2003, http://docserver.bis.uni-oldenburg.de/publikationen/dissertation/2004/hubobj03/hubobj03.html.

[34] R. Huber and B. Kollmeier, "Pemo-Q -A new Method for Objective Audio Quality Assessment using a Model of Auditory Perception." *IEEE Trans. on Audio, Speech and Language Processing*, 2006, special Issue on Objective Quality Assessment of Speech and Audio.

[35] ITU-T, "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," ITU, Series P: Telephone Transmission Quality Recommendation P.835, Nov. 2003.

[36] K. D. Kammeyer and K. Kroschel, *Digitale Signalverarbeitung: Filterung und Spektralanalyse*, 3rd ed. Stuttgart: Teubner, 1998.

[37] J. M. Kates and M. R. Weiss, "A comparison of hearing-aid array-processing techniques," *Journal of the Acoustical Society of America (JASA)*, vol. 99, no. 5, pp. 3138–3148, May 1996.

[38] T. J. Klasen, S. Doclo, T. V. den Bogart, M. Moonen, and J. Wouters, "Binaural multi-channel wiener filtering for hearing aids: Preserving interaural time and level differences," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. V. Toulouse, Frankreich: IEEE, May 2006, pp. 145–148.

[39] T. J. Klasen, T. V. den Bogaert, M. Moonen, and J. Wouters, "Binaural noise reduction algorithms for hearing aids that preserve interaural time delay cues," *IEEE Trans. on Speech and Audio Processing*, vol. 55, no. 4, pp. 1579–1585, April 2007.

[40] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. IEEE International Conference on ICASSP '82. Acoustics, Speech, and Signal Processing*, vol. 7, May 1982, pp. 1278–1281.

[41] C. H. Knapp and G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[42] B. Kollmeier, "Meßmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache," Habilitationsschrift, Universität Göttingen, 1990.

[43] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction." *J Acoust Soc Am*, vol. 95, no. 3, pp. 1593–1602, Mar 1994.

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

[45] P. C. Loizou, *Speech Enhancement: Theory and Practice.* Taylor & Francis, 2007.

[46] P. C. Loizou, A. Lobo, and Y. Hu, "Subspace algorithms for noise reduction in cochlear implants." *J Acoust Soc Am*, vol. 118, no. 5, pp. 2791–2793, Nov 2005.

[47] T. Lotter and P. Vary, "Dual-channel speech enhancement by superdirective beamforming," *Journal on Advances in Signal Processing (EURASIP)*, vol. 2006, pp. Article ID 63 297, 14 pages, 2006.

[48] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[49] ——, "Spectral subtraction based on minimum statistics," in *Proc. EURASIP European Signal Processing Conference (EUSIPCO)*, Edingburgh, Großbritannien, Sep. 1994, pp. 1182–1185.

[50] M. Marzinzik, "Noise reduction schemes for digital hearing aids and their use for hearing impaired," Ph.D. dissertation, University of Oldenburg, Nov. 2000.

[51] J. Meyer, "Beamforming for a circular microphone array mounted on spherically shaped objects," *Journal of the Acoustical Society of America (JASA)*, vol. 109, no. 1, pp. 185–193, Jan. 2001.

[52] ——, "Microphone array for hearing aids taking into account the scattering of the head," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2001.

[53] B. C. Moore and B. R. Glasberg, "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns." *Hear Res*, vol. 28, no. 2-3, pp. 209–225, 1987.

[54] R. D. Patterson, J. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Paper presented at a meeting of the IOC Speech Group on Auditory Modelling at RSRE, Dec. 14–15 1987.

[55] J. Peissig and B. Kollmeier, "Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners." *J Acoust Soc Am*, vol. 101, no. 3, pp. 1660–1670, Mar 1997.

[56] P. M. Peterson, "Using linearly-constrained adaptive beamforming to reduce interference in hearing aids from competing talkers in reverberant rooms," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1987, pp. 2364–2367.

[57] R. Plomp, "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired." *J Speech Hear Res*, vol. 29, no. 2, pp. 146–154, Jun 1986.

[58] D. Püschel, "Prinzipien der zeitlichen Analyse beim Hören," Ph.D. dissertation, Universität Göttingen, 1988.

[59] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality.* Englewood Cliffs: Prentice Hall, 1988.

[60] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 2, 7–11 May 2001, pp. 749–752.

[61] A. Rix, J. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective assessment of speech and audio quality&echnology and applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1890–1901, Nov. 2006.

[62] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective perceptual quality measures for the evaluation of noise reduction schemes," in *9th International Workshop on Acoustic Echo and Noise Control*, Eindhoven, 2005, pp. 169–172.

[63] ——, "Subband-based parameter optimization in noise reduction schemes by means of objective perceptual quality measures," in *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC).* Paris: Télécom Paris, SEPTEMBER 12-14 2006.

[64] ——, "Robustness analysis of binaural hearing aid beamformer algorithms by means of objective perceptual quality measures," *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pp. 315–318, Oct. 2007.

[65] T. Rohdenburg, S. Goetze, V. Hohmann, K.-D. Kammeyer, and B. Kollmeier, "Objective perceptual quality assessment for self-steering binaural hearing aid microphone arrays," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 2449–2452, 31 2008-April 4 2008.

[66] ——, "Combined source tracking and noise reduction for application in hearing aids," in *8. ITG-Fachtagung Sprachkommunikation*, no. 8. VDE VERLAG GMBH, Oct. 2008.

[67] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays*, Brandstein and Ward, Eds. Springer, 2001, ch. 3, pp. 39–60.

[68] W. Soede, A. Berkhout, and F. A. Bilsen, "Development of a directional hearing instrument based on array technology," *Journal of the Acoustical Society of America (JASA)*, vol. 94, pp. 785–798, 1993.

[69] W. Soede, F. A. Bilsen, and A. Berkhout, "Aseessment of a directional microphone array for hearing impaired listeners," *Journal of the Acoustical Society of America (JASA)*, vol. 94, pp. 799–808, 1993.

[70] R. Stadler and W. Rabinowitz, "On the potential of fixed arrays for hearing aids," *Journal of the Acoustical Society of America (JASA)*, vol. 94, no. 3, pp. 1332–1342, Sep. 1993.

[71] A. Stammermann, L. Kruse, E. Schmidt, A. Pratsch, M. Schulte, A. Schulz, and W. Nebel, "Orinoco: Verlustleistungsanalyse und optimierung auf der algorithmischen abstraktionsebene," Entwurf integrierter Schaltungen / 10. E.I.S.-Workshop. - Berlin, 2001.

[72] S. S. Stevens, "On the psychological law," *Psychological Review*, vol. 64, no. 3, pp. 153–181, 1957.

[73] J. Tribolet, p. Noll, B. McDermott, and R. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1978, pp. 586–590.

[74] T. Van de Bogaert, J. Wouters, T. Klasen, and M. Moonen, "Distortion of interaural time cues by directional noise reduction systems in modern digital hearing aids," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 16-19 Oct. 2005, pp. 57–60.

[75] T. Van den Bogaert, J. Wouters, S. Doclo, and M. Moonen, "Binaural cue preservation for hearing aids using an interaural transfer function multichannel wiener filter," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2007.

[76] P. Vary, *Signal Processing*. Elsevier, 1985, vol. 8, ch. Noise Suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits -, pp. 387–400.

[77] P. Vary and R. Martin, *Digital Speech Transmission*. John Wiley & Sons Ltd., 2006.

[78] ——, *Digital Speech Transmission: Enhancement, Coding, and Error Concealment*, P. Vary and R. Martin, Eds. Wiley, 2007, vol. 121, no. 1.

[79] K. Wagener, "Factors influencing sentence intelligibility in noise," Dissertation, University of Oldenburg, 2003.

[80] K. Wagener, T. Brand, and B. Kollmeier, "Entwicklung und Evaluation eines Satztests für die deutsche Sprache iii: Evaluation des Oldenburger Satztests," *Zeitschrift für Audiologie/Audiological Acoustics*, vol. 38, no. 3, pp. 86–95, 1999.

[81] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 30, no. 4, pp. 679–681, Aug 1982.

[82] D. Welker, J. Greenberg, J. Desloge, and P. Zurek, "Microphone-array hearing aids with binaural output. ii. a two-microphone adaptive system," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 6, pp. 543–551, Nov. 1997.

[83] T. Wesker, B. Meyer, K. Wagener, B. Kollmeier, and A. Mertins, "Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines," in *Interspeech*, 2005.

[84] T. Wittkop, "Two-channel noise reduction algorithms motivated by models of binaural interaction," Dissertation, University of Oldenburg, 2001.

[85] T. Wittkop and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Commun.*, vol. 39, no. 1-2, pp. 111–138, 2003.

[86] E. Zwicker, *Psychoakustik*. Springer-Verlag, 1982.

Erklärung

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

Oldenburg, den 4. Oktober 2008

.....
Thomas Rohdenburg

Danksagung

Diese Arbeit entstand während meiner Zeit als wissenschaftlicher Mitarbeiter an der Carl-von-Ossietzky Universität, Oldenburg. An dieser Stelle möchte ich mich herzlich bei denjenigen Personen bedanken, die mich unterstützt und begleitet haben.

Mein erster Dank gilt Prof. Dr. Dr. Birger Kollmeier, der mit seinem großen Interesse, zahlreichen Ideen und Anregungen meine Arbeit unterstützt und bereichert hat. Durch ihn hat sich die Arbeitsgruppe Medizinische Physik zu einer sehr vielfältigen und interdisziplinären Arbeitsgruppe entwickelt, von deren hervorragenden Arbeitsklima ich sehr profitieren konnte und in der ich mit meinen Fragen auf viele offene Ohren getroffen bin.

Ganz besonders möchte ich mich bei PD Dr. Volker Hohmann bedanken, der meine Arbeit intensiv betreut hat und viele Ideen und Ansätze beigesteuert hat. Gerade wenn ich mich zu tief im Detail verloren hatte, konnte er mich besonders motivieren und mir mit seinem umfangreichen Wissen den Blick für das Wesentliche schärfen.

Prof. Dr.-Ing. Karl-Dirk Kammeyer und Dr. Volker Hohmann danke ich für die freundliche Übernahme des Korreferats. Prof. Dr. Volker Mellert danke ich für die kurzfristige Übernahme des Beisitzes im Disputationsausschuss.

Dr. Rainer Huber und Dr. Rainer Beutelmann danke ich für den Support im Bereich Qualitätsmaße. Den Kollegen an der Fachhochschule Oldenburg, ganz besonders Prof. Dr.-Ing. Jörg Bitzer und Dr.-Ing. Uwe Simmer, sowie an der Universität Bremen Herrn Dipl.-Ing. Stefan Goetze, danke ich für wertvolle Tipps und die tolle Zusammenarbeit in Sachen Beamforming und Noise Reduction.

Neben allen Kollegen, die mir Rat und Tat zur Seite standen, möchte ich noch Susanne Garre, Ingrid Wusowski, Frank Grunau und Anita Gorges erwähnen. Sie sorgten dafür, dass ich in der MEDI wenig Sorgen mit Organisation und Administration hatte. Vielen Dank!

Meinen Bürokollegen und Freunden Dr. Jörg Damaschke und Dr. Stephan Ewert möchte ich für zahlreiche fachliche und nichtfachliche Unterhaltungen und ein tolles Arbeitsklima danken.

Nicht zuletzt möchte ich meinen Eltern für die Ermöglichung meines Studiums und die große Unterstützung auf dem Weg zur Doktorarbeit mit vielen, nicht selbstverständlichen Dingen und Hilfen danken.

Lebenslauf

Dipl.-Ing. Thomas Rohdenburg

geboren am 28. Mai 1975
in Bremen

Staatsangehörigkeit: deutsch

September 2003 bis August 2008	Universität Oldenburg, Promotion in der Arbeitsgruppe “Medizinische Physik”.
Mai 2003	Diplom, Titel der Diplomarbeit: “Klassifikation von Audiosignalen”
Oktober 1997 bis September 2003	Universität Bremen, Studiengang Elektrotechnik und Informationstechnik, Studienvertiefungsrichtung Informationstechnik
Oktober 1996 bis September 1997	Grundwehrdienst
Juni 1995	Abitur
Juni 1993 bis Mai 1995	Gymnasium Sekundarstufe II in Bremen