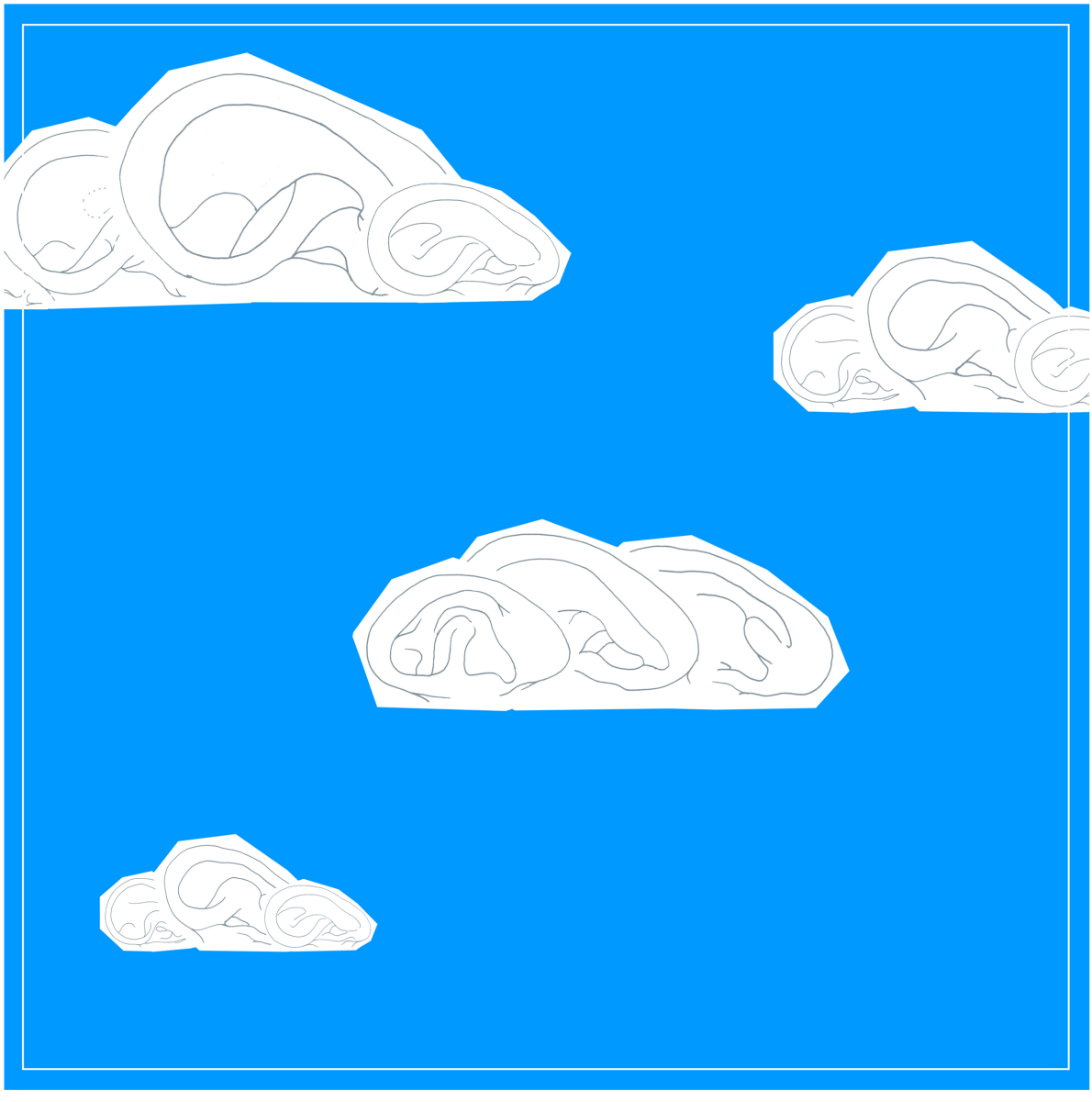

Binaural Model-Based Speech Intelligibility Enhancement and Assessment in Hearing Aids

Anton Schlesinger

Binaural model-based speech intelligibility enhancement and assessment in hearing aids

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen
op donderdag 12 januari 2012 om 10:00 uur

door

Anton SCHLESINGER

Diplom-Ingenieur, Technische Universität Ilmenau,
geboren te Dresden, Duitsland

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. A. Gisolf

Copromotor:

Dr. ir. M.M. Boone

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. A. Gisolf,	Technische Universiteit Delft, promotor
Dr. ir. M.M. Boone,	Technische Universiteit Delft, copromotor
Prof. Dr.-Ing. Dr. tech. h. c. J. Blauert,	Ruhr-Universität Bochum
Prof. dr. ir. J.M. Festen,	VU Amsterdam
Prof. dr. ir. L.J. Van Vliet,	Technische Universiteit Delft
Dr. ir. R. Heusdens,	Technische Universiteit Delft
Dr. rer. nat. V. Hohmann,	Universität Oldenburg

ISBN 987-94-6186-020-0

The research is financially supported by the Dutch Technology Foundation STW, which is part of the Netherlands Organization for Scientific Research (NWO) and partly funded by the Ministry of Economic Affairs, Agriculture and Innovation (project number DTF.7459).

Typesetting system: LaTeX.

Printed at RUB, Duitsland

Contents

1

Introduction

Abandoning an active lifestyle as a consequence of lacking the understanding of speech in noisy and complex environments is a depressing experience for many people and a severe social problem. Considering the European population, approximately one in five is suffering from a hearing loss (Shield, 2006). The annual monetary costs due to hearing loss are estimated to be 213 billion Euro for the European Union (Shield, 2006).

The problem of understanding speech in noise is well studied. The desensitization of the auditory perception as a result of different pathological causes and their combinations has generally direct consequences on speech intelligibility in silence and noise. Peripheral disorders, i.e. physiological illness in the middle ear and/or the cochlea are interdependent with higher neural stages of the auditory system.
Very consequential for the speech-in-noise problem is the degradation of the binaural processing, for the hearing impaired. Binaural processing is a central auditory process which takes a vital role in enriched and complex communication tasks. For instance, the normal hearing of a young person binaurally unmasks speech in noise, i.e. improves the signal-to-noise ratio (SNR) by about 10 dB if a continuous noise source with the long-term spectrum of speech rotates from frontal position, where the target speech is located, to the side. However, elderly people suffering from presbycusis—the majority of hearing impaired people is affected by this age-related widespread cochlea damage—experience only a difference of 2 to 3 dB in the same binaural comparison (Duquesnoy, 1983).
In addition, if the continuous noise source is substituted by a competing voice, young listeners with a healthy hearing generally gain another 3 to 4 dB advantage for lateral noise positions and even retain an advantage of 7 dB if the competing voice source coincides with the target voice in the frontal direction. In comparison, elderly people suffering from presbycusis show no benefit from gap listening due to their elevated hearing threshold and declined temporal acuity (Duquesnoy, 1983). In total, peripheral and concomitant central deficits of old people with presbycusis amount to an SNR difference of 5 to 15 dB with respect to young listeners with healthy hearing (Duquesnoy, 1983).
The severity of the problem is even more evident if one considers that one dB in SNR corresponds to 15 to 20 % of absolute speech intelligibility at the threshold of

understanding speech by 50 % in noise (Duquesnoy and Plomp, 1983).

Recent studies focus on the cognitive factors that are involved in the process of understanding speech in complex situations (see e.g., George et al., 2007; Pichora-Fuller, 2009). The findings widely indicate a primary dependence of speech intelligibility on auditory factors and a secondary dependence on cognitive factors, which, nevertheless, can be significant in active communication.
Another important finding is that speech intelligibility of old people relies more on cognition with respect to learned patterns than on an automatic bottom-up speech processing (Pichora-Fuller et al., 1995). These and related studies support an association between good hearing and cognitive health, and, therefore, clearly indicate the demand for early screenings and suitable clinical solutions to alleviate hearing impairment as well as to restore social well-being.

1.1 What a hearing aid should offer in noisy circumstances

The problem that the speech-in-noise problem poses can be judged from the benefit of todays hearing aids, as well as from their acceptance by those who are in need of an aid in noisy conditions.
At present there live about 55 million people in the European Union with different degrees of hearing loss. Until the year 2025, this number is expected to rise to approximately 100 million people (Shield, 2006, p. 32).

A first and general classification of the severity of a particular hearing loss is based on the threshold of audibility in silence. Therefore, the individual threshold is measured and related to the average limen of people with a healthy hearing. The audiogram is usually recorded with a pure tone test method. This method specifies a frequency-related hearing loss (HL) in dB, and serves as an important diagnostic. As a rule of thumb, people with a hearing loss higher than 35 dB are considered to have difficulties in understanding speech in silence (Plomp, 1986). To compensate their elevated thresholds, hearing aids are successfully prescribed, to transpose natural speech into the still available dynamic range, by signal amplification and compression.
Approximately 5 % of the EU inhabitants with a middle hearing loss ($>$ 35 dB HL) and 1 % of the EU inhabitants with a severe hearing loss ($>$ 60 dB HL) depend on hearing aids for understanding speech in silence.

Despite this success, the majority of hearing impaired people, i.e., approximately 17 % of the people in the EU with a mild hearing loss ($>$ 25 dB HL), experience generally no benefit from hearing aids. This is because their main problem is not understanding speech in silence, but understanding speech in noise. For instance, in case of Germany, the hearing aid manufacturer Audiological Technology Siemens estimate that only 10 % of the people with a mild hearing loss own a hearing aid (FCA, 2007).
The technological objectives a hearing aid should achieve to recover speech intel-

ligibility in noise, have been inferred by audiometry tests of speech in noise. An important diagnostic is the speech reception threshold (SRT) test in noise (Plomp and Mimpen, 1979). The method specifies the SNR threshold of 50 % speech intelligibility. Above an absolute sound pressure level (SPL) of 65 dB, the SRT is independent of the SPL. This allows for the definition of noise reduction solutions that offer a constant SNR gain. Related to extensive studies of hearing disorders and their prevalences, Plomp (1978) concluded that "every 4-5 dB of noise reduction halves the percentage of auditorily handicapped of any degree."
For a full compensation of the individual hearing loss in noise, the SRT difference between the hearing impaired and the normal hearing people has to be overcome. Referring to the numbers mentioned above, this corresponds to an averaged SNR improvement of 5 to 15 dB for the elderly with an age-dependent normative presbycusis pathology (Duquesnoy, 1983). It stands to reason that these are ambitious but necessary requirements for an SNR improvement. Individuals facing a profound hearing loss, as for example candidates for cochlear implants, will likely demand even more powerful noise suppression algorithms for regaining speech intelligibility in noise.

The SNR requirements must be interpreted with care, if one strives to meet them by algorithmic approaches. As such it turned out that the SNR can well be derived from speech intelligibility by speech audiometry. However, speech intelligibility can at best be loosely derived from SNR, if a signal undergoes a linear enhancement method. Higher processes of speech intelligibility on a microscopic signal level and on a macroscopic semantic level, are not expressed in the purely physical SNR measure. Besides, if a nonlinear algorithm is applied to enhance speech in noise, the SNR measure has been shown to fail as a predictor of speech intelligibility (see e.g., Loizou and Kim, 2011). Consequently, it is today increasingly accepted that psychoacoustical and physiological measures of speech intelligibility need to be applied, if speech enhancement approaches are to be assessed in an objective manner.

1.2 Today's hearing aid solutions

A recent three-year study on the benefit of current directional behind-the-ear (BTE) hearing aids across a wide range of types of hearing loss by Gnewikow et al. (2009), revealed the state of the art in terms of a speech intelligibility enhancement in noisy conditions. In an SRT test in a continuous diffuse noise field with a long-term speech spectrum, the hearing aids generated an SNR improvement of 2 to 3 dB in the directional mode relative to the omnidirectional mode. However, the overall success of wearing a hearing aid in noise conditions is higher. That is, the benefit of wearing a hearing aid in terms of speech intelligibility improved the SNR by 4 dB for people with a mild hearing loss and by 6 dB for people with a severe hearing loss. The assessment of the total gain in a percent-correct score test, using the directional processing, revealed an improvement of speech intelligibility of about 20 %, 30 %

and 15 % for people with a mild hearing loss, moderate hearing loss and severe hearing loss, respectively.
In terms of subjective ratings, people with a mild hearing loss consistently preferred in a user-preference questionnaire the directional hearing aid mode over omnidirectional mode. People with a moderate and severe hearing loss, however, consistently preferred listening in the omnidirectional mode.
It is suspected that the study may paint an optimistic picture of the overall benefit, for the reason that the hearing aid wearers were equipped with professionally fitted, top-notch hearing aids (Gnewikow et al., 2009).

Nevertheless, the study reveals clearly the limitations of today's hearing aid solutions in noise, and seems to agree well with other studies (Hamacher et al., 2008). The reasons for the limited effectiveness of hearing aids are threefold.
First, small monaural hearing aids, i.e. BTE or in-the-ear (ITE) hearing aids, which are largely sold and advertised with soft factors like "invisibly small and comfortable", are physically inadequate to incorporate an effective spatial sampling scheme, nor do they have the computational processing power to run strong and robust speech enhancement algorithms.
Secondly, a stigma is attached to hearing loss. Hearing loss is associated with age and dementia. As a consequence, many people with mild hearing loss are reluctant to face and treat their handicap.
Thirdly, small and more comfortable high-end hearing aids are expensive. People usually postpone purchasing and extensive audiological testing until their basic communication is severely affected.
As a consequence, active rehabilitation is often missed and non-auditory factors, as cognitive and mental health, are affected (Pichora-Fuller, 2009). Furthermore, a consideration of the hearing aid market situation reveals an apparent amplification effect of these three factors.
A patient referred for counselling faces an intransparent market with limited technical solutions that are generally differentiated by cosmetic and comfort features. The market is dominated by a small number of hearing aid manufacturers, which access a common patent-pool and, in Germany for instance, sell their products via an exclusive dealer network that is bound by contract to a designated product line (FCA, 2007). The product and market policy of feeding the stigma of hearing loss by advertising invisible products, which implicitly excludes speech intelligibility improvement in noisy situations, on the one hand, and controlling distribution and innovation by excluding new concepts and other competitors, on the other, generates and secures a total revenue of one billion Euro per year for an exclusive set of companies, only in Germany (Handelsblatt, 2010). The current development might be well justified in economical terms, but it does not correspond to the needs of at least three quarter of the hearing impaired that have difficulties understanding speech in noise.

1.3 Current research lines

The enhancement of speech intelligibility is a difficult problem. After many decades of pioneering research it can be summarized that primarily algorithms that exploit spatial diversity by spatial sampling, provide a solution to the problem (see e.g., Hamacher et al., 2008). These algorithms are known as multichannel filters. Their unifying feature is to enhance the target speech, either by a direct enhancement of the target signal, or implicitly by suppressing the noise.

Popular multichannel filters are the well-known beamformers. Until now the beamforming filters pose the most robust and practically efficient solution to the speech-in-noise problem. There are different types of beamforming filters. A powerful type is the minimum variance distortionless response (MVDR) beamformer that allows for a high and frequency-independent improvement of the SNR. The generalized sidelobe canceler (GSC) framework, which is an adaptive method to calculate the optimal filters instantaneously, represents a further advancement of beamformers. The method is superior in coherent noise conditions, but interference suppression in more complex and diffuse conditions is generally reduced to the gain that is provided by the underlying fixed processing scheme (Greenberg and Zurek, 2001). Recent implementations extend the GSC processing over two ears or bilaterally head-worn arrays (see e.g., Hamacher et al., 2008).

Another class of multichannel filters aims at decomposing the input into independent signals. This class is known as the blind source separation (BSS) approach (Kocinski et al., 2011). The methods can be highly efficient in the suppression of coherent noise sources, but are at the same time constrained by the underlying mathematics. To overcome this constraint, algorithmic approaches have been developed for the underdetermined case, i.e. when there are fewer microphones than mixed sources, which show much potential (Zheng et al., 2009).

The third well-known class of multichannel filters is the multichannel Wiener filter. An efficient version of this filter, which makes few assumptions about the noise field, is based on the binaural auditory principles of sound perception, that is, the computational mimicry of the auditory scene analysis (i.e., CASA). Similar to the model hearing process, the performance of binaural CASA filters is for the main part given by the binaural interaction process and the head shadow effect. After basic attempts showed signs of success, around the late eighties and early nineties of the last century (see e.g., Gaik and Lindemann, 1986; Kollmeier and Koch, 1994), and a decade of slower progress, the field lately got a new impulse by the introduction of statistical models that simulate parts of the auditory schema-driven top-down processing, thereby increasing the robustness of CASA algorithms in complex noise fields considerably (Harding et al., 2005; Nix and Hohmann, 2006).

In general these three classes of multichannel algorithms give the means to improve speech intelligibility in a signal-based manner. Prevalent and standard objective listening tests, which are generally designed to exclude any possible cognitive exploita-

tion of contextual effects, demonstrate the suitability of these algorithms. Despite that, when allowing more realistic interactions between lower and higher processes of the brain, it recently turned out that approaches that generally prevent fatigue and discomfort by even a coarse suppression of noise in speech gaps, that is, approaches that increase the ease of listening, enhance speech intelligibility in a manner that is beneficial to higher cognitive processes. For instance, dual task experiments of speech reception and cognition provide evidence that top-down and bottom-up processes are complementary means for solving the speech-in-noise problem (Humes, 2002). In accordance with this understanding it became possible to demonstrate that single channel speech enhancement algorithms, which generally fail in enhancing speech intelligibility objectively (Hu and Loizou, 2007), yet provide a benefit by reducing the cognitive load and can even increase the speech recognition rate in semantically meaningful circumstances (Sarampalis et al., 2009). Nevertheless, if the objective is a considerable improvement of speech intelligibility, results show that any successful system will have to operate in domains that enable the best possible instantaneous decomposition of the complex texture of real-world sound scenes. To date, as mentioned above, this possibility is only provided by spatial sampling and processing schemes.

1.4 Contents of this thesis

The above-mentioned research approaches can be classified as speech enhancement algorithms suitable for the suppression of diffuse noise fields and algorithms that are suited for the suppression of coherent noise interference. The combination of these classes of algorithms has been pursued in several works, see e.g., Martin (2001), Hamacher et al. (2008) and Rohdenburg (2008), and was laid down in a fundamental account by Simmer et al. (2001) on the factorization of the minimum mean square error (MMSE) solution into an MVDR-beamformer and a single-channel Wiener post-filter.
The motivation of the present study is stemming from the same intent. Based on the legacy of speech enhancement with beamforming techniques in Delft, known, e.g. by the works of Soede et al. (1993), Merks (2000) and the country-wide market launch of the MVDR-based hearing glasses of the manufacturer Varibel Innovations BV (Boone, 2006), the present work proposes to combine bilaterally applied beamforming front-ends with binaural CASA post-filters, for the purpose of a higher overall speech intelligibility gain in noise.

Conceptually this work puts generalizability of the approaches and results before refinement of a particular technical solution. This approach suggests itself in a field where there is great heterogeneity of speech enhancement and evaluation methods. In addition, the aim is to assess the binaural CASA approaches under realistic conditions. Therefore only commercially available bilaterally applied hearing aids are used as a front-end to the binaural CASA post-processors. For the same reason, real-

world recordings of sound scenes are applied to test the signal processing schemes with their complex physical nature.

Even though combined processing schemes are analyzed, the assessment throughout this work is limited to the intelligibility improvement given by a set of binaural CASA post-processors. These are the coherence-based algorithm of Allen et al. (1977), the binaural waveform algorithm of Gaik and Lindemann (1986) and the binaural envelope algorithm of Kollmeier and Koch (1994).
The influence of the beamforming front-end will be analyzed in terms of a statistical analysis of binaural waveform and envelope cues in different noise conditions, and throughout the assessment of the binaural post-filters.

The present review is based on a comprehensive study that incorporates the functions of different CASA-processors, the psychophysical nature of binaural cues in noise, as well as the model-based assessment speech intelligibility. The parts of this holistic approach are interconnected via an evolutionary optimization method and, partly, via a pattern-based classification approach. The latter mimics top-down processes of the auditory scene analysis and allows for an optimal adaption of the post-processor to the beamforming front-end and the sound scene in terms of the binaural classification.

Chapter 2 to 4 introduce the general signal processing approach, a statistical analysis of binaural cues, their optimal activation and the intelligibility assessment of binaurally and nonlinearly processed speech. Chapter 5 deals with the optimization of the post-processors and the assessment of these throughout a wide range of acoustic scenes. In the remainder of this introduction the contents of the following chapters are described in greater detail.

Chapter 2 introduces the signal model of the spatial sampling scheme and the MMSE factorization into an MVDR beamformer and a single-channel Wiener post-filter. This serial processing scheme forms the general framework of the speech enhancement approaches of this thesis. Following an analysis of theoretical and practical MVDR beamformer solutions, the conventional Wiener post-filter will be introduced and contrasted with the widely applied concept of ideal binary masks in CASA noise suppression. In order to gain an understanding of the energy dispersion of different signal mixtures in the time-frequency domain, an elementary statistical analysis of the SNR distribution will be given.
Nonlinear speech enhancement by a varying time-frequency processing is a faulty process in real-world applications, that generally leads to a quality-impeding artefact known as musical noise. An efficient method for the suppression of this artefact is the cepstral smoothing technique (Breithaupt and Martin, 2008). The method will be introduced in Chapter 2.3.4 for a later optimization and application in Chapter 5.3.

Following this general introduction of the here applied non-adaptive and adaptive speech enhancement methods, three binaural CASA processors, which share the separation of speech and noise by spatial cues, are conceptually introduced.

Various designs of binaural CASA algorithms for speech enhancement exist. Many of them originate from the binaural algorithm of Gaik and Lindemann (1986). This speech processor accomplishes a bilateral frequency decomposition and subsequently calculates the interaural phase and level differences (IPD and ILD, respectively) of the acoustic waveform, to employ these parameters as a directional classifier in a magnitude weighted separation process.
A second group of binaural CASA algorithms adopts the concept of the multi-channel spatial coherence algorithm of Allen et al. (1977). Based on primitive grouping, this algorithm exploits the binaural waveform coherence to suppress diffuse sound.
A third well-known binaural CASA algorithm filters the signal in a conjoint centre and modulation frequency domain and was developed by Kollmeier and Koch (1994). Therein the separation process is based on the binaural level and phase differences of the envelope signal in the range of the fundamental frequency of speech. As the envelope of the signal is considered to be more robust towards noise than the acoustic waveform, this algorithm triggered much hope for an efficient speech enhancement in highly adverse conditions, at the time of its development.
All these algorithms offer a binaural output signal, which is known to add to the audiological benefit due to a cue-supported hearing.

In view of recent advancements in the field, the present study undertakes to update and review these three binaural CASA processors, and additionally combines these with a set of binaural front-ends. These front-ends are an artificial head (Institute for Technical Acoustics (ITA) head of the RWTH Aachen), a BTE hearing aid (GN ReSound type Canta 470-D) with and without directional processing, and the hearing glasses (HG) in two directivity modes (Varibel Innovations BV). Both hearing aids are mounted upon the Aachen head mannequin.

Chapter 3, the first part, deals with a statistical analysis of binaural waveform and envelope cues in noise at the output of binaural front-ends. In the second part, a pattern-based classification method for binaural waveform and envelope cues is presented.

So far, CASA algorithms are generally applied without a thorough understanding of the signal power dispersion of multiple sources in different feature spaces and the manner in which binaural parameters change in noise. The current study works towards better understanding by providing the statistics of binaural parameters of the envelope and the fine-structure of waveforms in noise. For this purpose, the binaural output of an artificial head is compared with the binaural output of highly directional hearing aids. The binaural parameters are calculated on a short-time base with a discrete Fourier transform (DFT) framework and an averaging of the DFT-based power spectral densities over auditory filters. Given this psycho-physical insight, the aim is to answer the question why binaural cue-based source separation succeeds in some circumstances and fails in others.

An important consequence of the statistical analysis of binaural cues in noise is the manner in which binaural cues are employed in the noise suppression process. As

will be shown, the distribution of binaural parameters shows a strong dependency on the strength and the spatial dispersion of the interference. To account for this dependence, the application of binaural cues needs to be pattern-driven, that is, comparable to the manner in which top-down processes of the auditory scene analysis activate different cues. Harding et al. (2005) introduced the principles of a pattern-driven binaural source separation by employing a Bayesian classifier. We adopted this approach for the calculation of weighting functions in the algorithms of Gaik and Lindemann (1986) and of Kollmeier and Koch (1994). In contrast, the algorithm of Allen et al. (1977) will be based on the standard primitive grouping scheme, using the non-directional magnitude squared coherence at zero lag, as a noise classifier.

Chapter 4 turns to the problem of speech intelligibility assessment of binaurally and nonlinearly processed speech. To that end, a speech-based version of the speech transmission index (STI) is extended by a binaural stage that incorporates the binaural interaction and head-shadow effect. As will be shown in a subjective test series, the model-based assessment method explained a great part of the binaural advantage for linearly processed speech. However, the intelligibility of nonlinearly processed speech cannot be predicted with this, or other purely bottom-up approaches of speech intelligibility.

Therefore, the second part of this chapter aims at progressing towards an instrumental measure that is capable of estimating the influence of nonlinearity on intelligibility. A method that roughly incorporates the changing information content in short-time frames of speech, is the short-time level weighted speech intelligibility index (SII) of Kates and Arehart (2005b). The measure is abbreviated with I3, which refers to three ranges of short-time speech frames that contribute differently to the overall speech intelligibility. Thereby the metric accounts for the fact that consonants are generally more meaningful to speech intelligibility than vowel sections. Likewise, transients and transitions of speech are of higher weight to speech intelligibility. These sections are well separated from the vowel sections by the speech power level in small frames (Yoo et al., 2007). Nonetheless, as the subjective assessment will show, the I3 measure constitutes a suboptimal solution in the assessment of nonlinearily processed speech. Therefore, the short-time SII measures will be combined with a series of information theoretic quantities, e.g. Shannon's entropy, to label transitional parts in speech.
Although subjective tests will show that these measures are able to detect transitional parts in speech, they are generally outperformed by the I3 measure, the STOI measure (Taal et al., 2010) and an optimized level-based SII version. Originating from these results on objective speech intelligibility assessment, a short-time and critical-band and Better Ear I3 method will be developed to account for the dominant factor in binaural speech intelligibility, the head-shadow effect.

Chapter 5 pursues the optimization and finally the assessment of binaural CASA processors. Using the Better Ear I3 measure of binaural speech intelligibility as an objective function, algorithmic parameter sets of the binaural speech processors will be optimized with a genetic algorithm across groups of acoustic scenes. As a result of the replicated model hearing process, the optimization realizes a most favourable balance of, algorithmically accessible, cues and binding features. The holistic approach of model-based speech enhancement and assessment yields an optimal set of parameters in this framework and consequently realizes the optimal algorithmic benefit in a particular scene. Since the three binaural processors of this work are based on different binaural processes, the evolutionary optimization constitutes an innovative approach to the qualification of certain binaural cues and binding features in varying acoustics.

Equipped with optimized parameter sets for certain algorithmic front-end/back-end combinations for particular scenes, the benefit of binaural speech processors will be assessed throughout a wide range of SNR conditions, target-masking angles, multi-masking conditions, several diffuse real-world backgrounds and artificial reverberation. Thereby the analysis will not only reveal the benefit that can be expected in matched conditions, i.e. scenes the algorithm was optimized for, it also gives information on the generalizability of binaural CASA algorithms in unmatched conditions.
Lastly, a genetic optimization will be applied to the cepstral smoothing technique. As a means to maintain the intelligibility benefit of CASA processors while maintaining a quality enhancement in terms of the suppression of musical noise, the cepstral smoothing technique is subsequently integrated in the processing chain with the algorithm of Gaik and Lindemann (1986) and objectively assessed with respect to binaural speech intelligibility and a binaural quality measure.

The study is concluded with a summary and an outlook in Chapter 6.

2

The Minimum Mean Square Error solution factorized into a beamformer and a post-filter

A familiar acoustical experience in populous surroundings is the ambiance of multiple speakers and a diffuse background of reverberating sound power. The stronger a talking partner is energetically masked by competing voices and reverberation, the lower speech intelligibility gets.
If a target speaker is to be enhanced, array technology and beamforming filters offer excellent means to exploit the spatial diversity of a given sound scene. The principle of the popular and simple delay-and-sum beamformer is to add the microphone signals of the target speaker coherently through a correction based on the inter-microphone delay times. In doing so, the signals from all other directions are added incoherently. As a result, the processing causes an enhancement of speech intelligibility by increasing the SNR.
A more powerful variant of beamforming is well-known as the minimum variance distortionless response (MVDR) beamforming, a directivity optimized spatial filtering. The filter coefficients of the MVDR method are calculated by minimizing the output power of the beamformer with the constraint of unity gain in the target direction. As the MVDR filters are only optimal in the noise field they are optimized for, the approach constitutes a maximum likelihood solution.
Under the assumption that the target signal and noise are uncorrelated, a universal and optimal reconstruction can theoretically be obtained by the multi-channel Wiener filter, which realizes a Minimum Mean Square Error (MMSE) solution. To approximate this theoretical approach, Simmer et al. (2001) showed that the MMSE solution can be factorized into an MVDR beamformer and a single-channel Wiener post-filter. As mentioned in the introduction, this combined processing scheme will be the strategy of the present work to enhance speech intelligibility in noisy surroundings.

With regard to the factorized MMSE solution, there are different approaches to distribute the tasks of coherent and incoherent noise suppression over the beamforming

front-end and the post-processing back-end. Some systems combine the Generalized Sidelobe Canceller (GSC), which is an adaptive beamforming framework for the suppression of coherent noise, with a single-channel post-filter for the suppression of diffuse noise (see e.g., Simmer et al., 2001). These approaches, however, have demonstrated to have two major disadvantages in real-world application. First, the GSC tends to be unstable in diffuse and variant noise conditions and is constrained by its mathematical solution when the amount of coherent sources exceeds the number of microphones (Greenberg and Zurek, 2001). Secondly, single-channel post-filters need to define an estimate of the noise with a classification algorithm that distinguishes between speech and noise, well-known as voice activity detection (VAD) algorithms. Voice activity detection in a single-channel speech signal is generally known to be an inaccurate estimation process that has shown to introduce an error that precludes speech intelligibility improvement (Rohdenburg, 2008; Loizou and Kim, 2011).
In multi-channel setups, the attenuation of diffuse sound with a post-filter can alternatively be based on the Magnitude Squared Coherence (MSC) function between the microphones, an approach that was first applied to the problem of speech enhancement by Allen et al. (1977). If binaural hearing aids are considered, head-based adaptive MVDR systems (GSC frameworks) with coherence-based post-processors were developed by Lotter and Vary (2006) and Rohdenburg (2008).
A different allocation of tasks to suppress incoherent and coherent sources is obtained through the combination of the general fixed MVDR beamformer and a scene-adaptive directional post-processor. While the former is optimized to suppress an ideal diffuse noise field, the latter performs the suppression of coherent noise sources. Variants of this approach can be found in the work of Seltzer et al. (2007), who proposed the combination of an array and a post-filter, which suppresses incoherent and coherent noise based on statistical modeling, or in the approach of Lockwood et al. (2004), in which the output of bilaterally applied cardioid microphones[1] was post-processed with the algorithm of Kollmeier et al. (1993). Hence, provided that fixed MVDR beamformers (or generally superdirective beamforming solutions) are bilaterally applied as a front-end, an approximation of the MMSE solution can be obtained by the sequential application of binaural speech processors for the suppression of lateral noise sources. Binaural CASA algorithms have shown to make only few assumptions about the noise field and, moreover, to belong to an exclusive set of varying magnitude-based filters that are able to gain significant improvement of the intelligibility (see e.g., Wittkop et al., 1997).
Moreover, the front-back ambiguity of CASA algorithms, which can be considered as a modeled natural artifact, is supposed to be reduced by the superposition of a directional beampattern (Kollmeier and Koch, 1994). Therefore, a benefit for the processing of the binaural post-filter is expected due to the directivity of the front-end.

[1] Directional microphones using an acoustical network to render a hypercardioid directivity pattern (i.e. a first order gradient solution) are analogue realizations of superdirective arrays (Merks, 2000).

In this chapter, the combined MVDR front-end and binaural CASA-based back-end system will be introduced. Subsequently to a theoretical introduction of the methods, practical solutions of the beamforming front-ends are analyzed. Post-filters in general and their binaural CASA realizations are studied in the second part of this chapter.

2.1 Signal model of the MMSE approximation approach

This section introduces the general approach of the combined processing scheme that is studied in this work. As introduced, the MMSE solution is factorized into an MVDR beamformer and a binaural CASA-based post-filter. Figure 2.1 **A** shows the

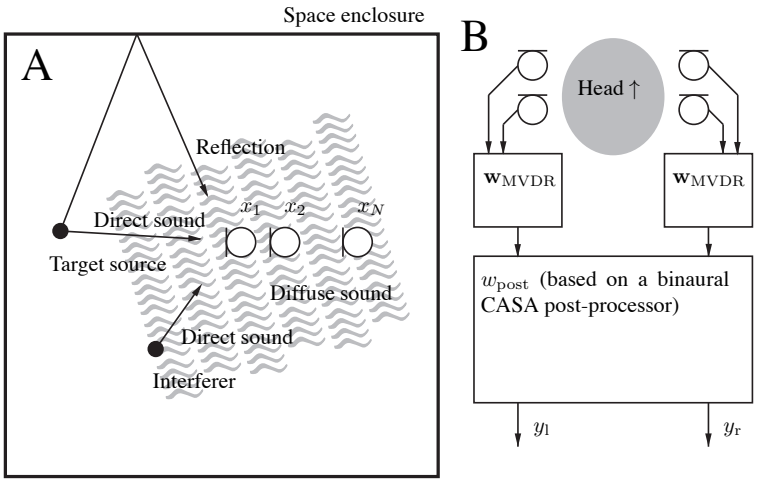

Figure 2.1: A *gives a schematic sketch of an additive signal model in a sound-field with certain room characteristics and an array in endfire setup with respect to the target source. Sketch* B *shows the processing scheme of this work, which comprises bilaterally applied MVDR beamformers (or generally filters that offer superdirectivity) for the suppression of diffuse noise and a binaural CASA-based speech processor for the suppression of coherent interference as well as residual diffuse noise.*

general signal model. The general aim is to enhance the target speaker s in a mix of multiple interference, using an array of N microphones. For this purpose, the output $\tilde{x}$ at each microphone ℓ is band-limited and sampled over ι time intervals. To obtain a frequency representation of the signal in short time-frames—in which speech signals can efficiently be filtered (Paliwal and Wojcicki, 2008)—a transformation to the

short-time Fourier transform (STFT) domain is performed with a N_d-point discrete Fourier transform (DFT) over time-frames of short duration:

$$x_\ell(d,n) = \sum_{\iota=0}^{N_\mathrm{d}-1} \chi(\iota)\tilde{x}_\ell(n\Delta T+\iota)e^{-j2\pi\iota\frac{d}{N_\mathrm{d}}}, \tag{2.1.1}$$

where $d = 0, 1, \ldots, N_\mathrm{d}-1$, n, ΔT and χ are the frequency bin, the frame index, the frame shift and a window function, respectively. The microphone signals $x_\ell(d,n)$ can be written as a vector $\mathbf{x}(d,n)$, and considering the mix of signals, $\mathbf{x}(d,n)$ can be expanded into:

$$\mathbf{x}(d,n) = \mathbf{s}(d,n) + \mathbf{v}(d,n) = \mathbf{a}(d)s(d,n) + \mathbf{v}(d,n), \tag{2.1.2}$$

in which $s(d,n)$ denotes the source signal in the STFT domain, and $\mathbf{a}(d)$ is the propagation path vector:

$$\mathbf{a}(d) = (a_0(d), a_1(d), \ldots, a_{N-1}(d))^\mathrm{T}, \tag{2.1.3}$$

between the source s and each microphone ℓ, which is considered stationary throughout this work. In Equation (2.1.3) the subscript T denotes the transposition of a vector and $\mathbf{v}(d,n)$ is the noise vector that comprises all distortions, i.e. room reverberation, interfering sources or the microphone self noise. In the following, the time and frequency index are omitted without a loss of general validity, for notational convenience.

To separate the speaker s from the noisy mixture in the output y, the input vector $\mathbf{x}$ can be multiplied with the multichannel filter coefficients $\mathbf{w}$:

$$y = \mathbf{w}^\mathrm{H}\mathbf{x}, \tag{2.1.4}$$

where H indicates the Hermitian transposed. The filter vector $\mathbf{w}_\mathrm{opt}$ that is obtained from the MMSE solution is:

$$\mathbf{w}_\mathrm{opt} = \underset{\mathbf{w}(d,n)}{\operatorname{argmin}} \mathrm{E}\left[\left(s - \mathbf{w}^\mathrm{H}\mathbf{x}\right)^2\right]. \tag{2.1.5}$$

The solution constitutes a multichannel Wiener filter. As the practical realization of such a filter is an unsolved problem for the broadband speech-in-noise problem considered, combined processing schemes have been developed for an approximation of this solution. Usually these consist of a beamformer and a post-filter solution. Simmer et al. (2001) formalized those practical approximations of the MMSE filter with a factorization into a complex weight vector of a MVDR beamformer and a scalar single channel Wiener post-filter in the following way:

$$\mathbf{w}_\mathrm{opt} = \underbrace{\frac{\phi_\mathrm{ss}}{\phi_\mathrm{ss}+\phi_\mathrm{vv}}}_{w_\mathrm{post}} \underbrace{\frac{\mathbf{\Phi}_\mathrm{vv}^{-1}\mathbf{a}}{\mathbf{a}^\mathrm{H}\mathbf{\Phi}_\mathrm{vv}^{-1}\mathbf{a}}}_{\mathbf{w}_\mathrm{MVDR}}. \tag{2.1.6}$$

In this equation $\mathbf{\Phi}_{\mathrm{vv}}^{-1}$ is the inverse cross power spectral density matrix of the noise across the microphones, and ϕ_{ss} and ϕ_{vv} are the power spectral densities of the target signal and the noise, respectively. The calculation of the power spectral densities is introduced in Chapter 2.3.

The MMSE factorization applied in the present work is outlined in Figure 2.1 **B**. Bilaterally applied MVDR beamformers (or superdirective beamformers) form the front-end to a central post-processor, which subsequently applies a soft-mask, i.e an approximation of the Wiener gain, to both channels.
Compared to pure single-channel approaches that generally estimate the noise in speech pauses, the here presented binaural CASA approach provides a constant noise estimate by using the spatial information contained in the binaural parameters to derive the filter-gains in each channel. Therefore, using the factorized weights of Equation (2.1.6), Equation (2.1.4) can be rewritten for the combined binaural system proposed here:

$$\begin{pmatrix} y_{\mathrm{l}} \\ y_{\mathrm{r}} \end{pmatrix} = w_{\mathrm{post}} \begin{pmatrix} \mathbf{w}_{\mathrm{MVDR}}^{\mathrm{H}} \mathbf{x}_{\mathrm{l}} \\ \mathbf{w}_{\mathrm{MVDR}}^{\mathrm{H}} \mathbf{x}_{\mathrm{r}} \end{pmatrix}, \tag{2.1.7}$$

where the indices l and r denote the left and right ear signal, respectively. Hence, subsequent to the multiplication of each array vector with the beamformer weights $\mathbf{w}_{\mathrm{MVDR}}$, the binaural output is multiplied with the real-valued post-filter gain function w_{post}. This approach implies that binaural phase and level differences exist at the output of the beamforming front-end, which can subsequently be accessed in the source segregation process by the post-filter. Consequently, natural interaural disparities are altered twice: once by the beamformer and then once by the post-filter gain. Nevertheless, the combined binaural system delivers a binaural signal to the ears, which is known to give speech intelligibility improvement, a topic that is discussed later in this work.

The total benefit of an $\mathrm{SNR}_{\mathrm{total}}$ enhancement of the combined processing scheme is calculated as the sum of the respective logarithmic noise suppression gains (Simmer et al., 2001):

$$\mathrm{SNR}_{\mathrm{total}} = \mathrm{SNR}_{\mathrm{MVDR}} + \mathrm{SNR}_{\mathrm{post}}. \tag{2.1.8}$$

The enhancement of the array $\mathrm{SNR}_{\mathrm{MVDR}}$ can be expressed as the inverse of the MVDR array gain, i.e. the ability to suppress a diffuse noise field:

$$\mathrm{SNR}_{\mathrm{MVDR}} = 10 \log_{10}(\mathbf{w}^{\mathrm{H}} \mathbf{\Gamma}_{\mathrm{vv}} \mathbf{w}), \tag{2.1.9}$$

in which $\mathbf{\Gamma}_{\mathrm{vv}}$ denotes the complex coherence matrix that is equal to the normalized noise correlation matrix $\mathbf{\Gamma}_{\mathrm{vv}} = \frac{\mathbf{\Phi}_{\mathrm{vv}} N}{\mathrm{trace}(\mathbf{\Phi}_{\mathrm{vv}})}$. The SNR enhancement of the post-filter is $\mathrm{SNR}_{\mathrm{post}} = 10 \log_{10}(|h_{\mathrm{post}}|^2)$, with h_{post} being the transfer-function of the post-filter.

By these means, the processing units are complementary for approximating the MMSE solution. Their respective contributions in suppressing coherent and incoherent noise signals are, however, subject to practical limitations. That is, there is

generally no realizable MMSE solution to the considered speech-in-noise problem. For instance, because of the serial processing of the beamformer and post-filter, the latter depends on the output characteristics of the former. This dependency has a direct consequence on the overall performance. A high attenuation of diffuse noise through a beamformer might help a post-processor to detect and attenuate coherent noise sources. However, it might also lead to a deterioration of spatial cues that consequently hampers the separation process and thus the post-filtering quality. Therefore, it is a central question of the present study, whether audiologically inspired post-filters are able to utilize the binaural disparities at the output of bilateral beamformers in the source separation task. Practical simulations, later in this work, aim to answer this and other questions that relate to the problem of speech intelligibility enhancement in real-world scenarios. The remainder of this chapter introduces practical realizations of the MVDR beamformer and the binaural CASA post-filters applied in this work.

2.2 Practical superdirective beamforming

Spatial filtering by application of beamforming comes in three guises: the delay-and-sum beamforming, the gradient processing and the MVDR beamforming. Whereas delay-and-sum beamforming maximizes the amplitude of the target signal by correcting the sound-travel time differences between the microphones in a preferential direction, the gradient processing and the MVDR beamforming minimize the energy of the array output by a decorrelation process of the sound field and enhance the target by a steered unit filter gain. In cases the wavelength λ is greater than twice the microphone spacing l, the gradient and MVDR processing can achieve superdirectivity.

The quality of the beamformer solution can physically be assessed with the Directivity Index (DI):

$$\mathrm{DI} = 10\log_{10}\left(\frac{|\mathbf{w}^{\mathrm{H}}\mathbf{a}|^2}{\mathbf{w}^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{vv}}\mathbf{w}}\right), \tag{2.2.10}$$

and the White Noise Gain (WNG):

$$\mathrm{WNG} = 10\log_{10}\left(\frac{|\mathbf{w}^{\mathrm{H}}\mathbf{a}|^2}{\mathbf{w}^{\mathrm{H}}\mathbf{I}\mathbf{w}}\right). \tag{2.2.11}$$

These expressions specify the two opposing aims of a practical beamforming solution, i.e. the suppression of diffuse noise, as expressed in a coherence matrix of the sound field $\boldsymbol{\Gamma}_{\mathrm{vv}}$ in the denominator of the DI, and the quality of the suppression of uncorrelated noise among the sensors, as expressed through the identity matrix $\mathbf{I}$ in the denominator of the WNG.

The MVDR solution, as introduced in Equation (2.1.6), can be calculated by minimizing the signal power $\mathbf{w}^{\mathrm{H}}\boldsymbol{\Gamma}_{\mathrm{vv}}\mathbf{w}$ at the beamformer's output, while obeying $\mathbf{w}^{\mathrm{H}}\mathbf{a} =$

1 (Bitzer and Simmer, 2001). Although this leads to the directivity optimized solution, depending on the characteristics of the array, i.e. the number of microphones, their location and the spacing between them, it is an impractical solution because it amplifies uncorrelated noise among microphones at low frequencies, which results in a negative WNG. The well-known approach for a balance between directivity and self-noise amplification is provided through the WNG-constrained MVDR solution. This constrained solution is obtained through the stabilization of the matrix inversion:

$$\mathbf{w}_{\mathrm{MVDR}}\,|_{\mathrm{stab}} = \frac{(\mathbf{\Gamma}_{\mathrm{vv}} + \kappa\mathbf{I})^{-1}\mathbf{a}}{\mathbf{a}^{\mathrm{H}}(\mathbf{\Gamma}_{\mathrm{vv}} + \kappa\mathbf{I})^{-1}\mathbf{a}}, \tag{2.2.12}$$

in which $\kappa\mathbf{I}$ represents an adjustable amount of uncorrelated noise. Hence, by modifying κ, a compromise between the conflicting characteristics of target gain and noise robustness can be reached. As there is no simple relation among κ and the WNG, usually an optimization routine is used to reach optimal directivity at a given minimum WNG (Merks, 2000). By virtue of this stability constraint, the mutual uncorrelated self-noise of the transducers due to gain and phase mismatches as inevitable consequences of the production process, numerical noise due to errors in the signal processing, and also noise due to wind turbulence, can efficiently be counteracted.

To exemplify the working principle of the constrained MVDR beamformer, a two microphone endfire array (collinear with the target source) is simulated. We assume the microphones to be in the far-field of the source, and the presence of an ideal diffuse noise field that is to be attenuated. The distance l between the microphones is 5 cm, which results in a spatial Nyquist frequency ($c/2l$, with c being the speed of sound) of 3.4 kHz. Figure 2.2 shows two adjustments of the constraint MVDR solution. One solution with a $\kappa = 0$ results in the unconstrained MVDR solution, which is equal to a first order gradient solution with hypercardioid weights (Merks, 2000). The other setting is given by the a constrained MVDR solution with $\kappa = 10$. In this case, the constrained MVDR processing approximates the robustness and the gain of the delay-and-sum solution.
The DI for the unconstrained solution is 6 dB. For enfire-arrays it can be calculated as $20\log_{10}(N)$ (Merks, 2000). The WNG shows a strong amplification of uncorrelated noise at low frequencies. By inspecting the beamforming coefficients in Figure 2.2, a strong amplification towards low frequencies can be seen. As the unconstrained MVDR beamforming imposes a decorrelation of the correlated noise—which is seen as a phase difference of π among the filter coefficients, as long as sensor distance is smaller than half of the wave length, i.e. $l < \lambda/2$, with λ being the wave-length, in the right-hand plot of Figure 2.2—, the correlated target source will be decorrelated as well. To offset this attenuation, high filter gains guarantee an undistorted target signal, i.e. a unity gain. However, the inevitable consequence of this approach is the amplification of uncorrelated noise (Bitzer and Simmer, 2001).
The opposite nature in terms of directivity and stability is found for the constrained MVDR solution with $\kappa = 10$ in Figure 2.2. Almost no directivity is found for

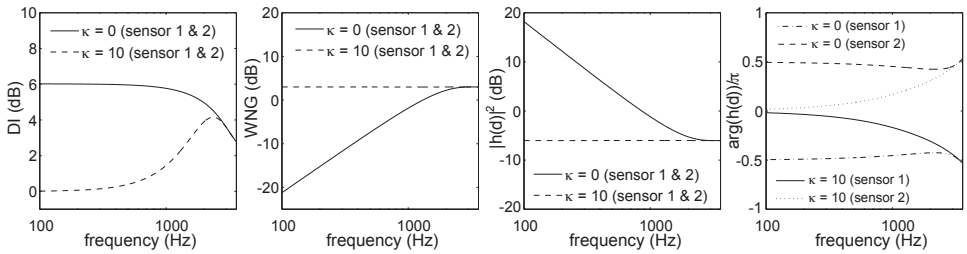

Figure 2.2: *The DI and WNG for an endfire array with two omni-directional microphones exemplify the working principle of the unconstrained MVDR beamforming with $\kappa = 0$ and the constrained MVDR beamforming with $\kappa = 10$. The right-hand plots show the concomitant complex beamforming weights.*

$f < 1$ kHz. Only if the wavelength is small compared to the dimensions of the array (up to the Nyquist limit), the solution generates a moderate gain. In contrast to the unconstrained MVDR solution, uncorrelated noise is highly attenuated in the entire spectrum, which is indicated by a constant positive WNG. As mentioned, the solution boils down to an approximation of the delay-and-sum beamformer where all delays are corrected with f-dependency for the target direction and constant gains of $1/2$.

Returning to the introduction of the three general beamforming solutions, the narrowband low gain and robust delay-and-sum solution, and the broadband high gain and unstable gradient solution, mark the extremes of the beamforming method and can be seen as two sub-solutions of the constrained MVDR solution (Merks, 2000). In practical applications, a compromise between the conflicting objectives of target gain and self-noise amplification has to be found.

2.2.1 Bilateral beamforming and the effect on binaural cues and speech intelligibility

As explained and illustrated, the beamformer approach exploits the spatial diversity of sources to improve the SNR. If, however, this spatial diversity is cancelled in a monaural or diotic signal, the benefit of the spatial filter is often practically compensated by the lack of binaural cues, which are strongly used in the auditory scene analysis. In order to realize an improvement of speech intelligibility, it is important to convey the spatial diversity to the listener by providing a binaural output signal (Desloge et al., 1997; Hamacher et al., 2008). Additionally, the hearing efforts are relaxed through a natural spatial listening experience. Together with a low target distortion as well as noise suppression, this effect adds to the listening ease and, thus, acts against a lessening of mental attention and fatigue.

Several head-related array systems have been developed to comply with the requirement of binaural cue preservation. In a recent study, Rohdenburg (2008) compared different binaural noise reduction schemes, which were applied to the output of bilateral BTEs with each three microphones. In an evaluation he found that head-related arrays that accommodate superdirective beamforming and post-filtering in a joint real-valued transfer-function, are well suited to provide an ideal balance between noise suppression and binaural cue preservation. In a similar fashion, Merks (2000) demonstrated that bilaterally applied endfire arrays using beamforming allow for a subjective localization performance comparable to the localization with natural binaural cues. As a matter of course, binaural cue fidelity is counteracted by the width of the main-lobe and the beampattern, i.e. the directivity. As a consequence, if the main lobe is narrow and the array gain is high, binaural cues tend to be unnaturally modified. Moreover, binaural cues may strongly fluctuate in the vicinity of inevitable zeros in the beampattern. Wearers of hearing aids using beamforming usually adapt to these differences to some extent. Rohdenburg (2008) confirmed this relative subjective robustness in favour of noise suppression. In an evaluation with normal hearing people at SNRs around 0 dB and above, he found binaural cues to be less important than the preclusion of target distortion. Merks (2000) compared a binaural broadside-array that partially conveyed ILDs with a binaural endfire array that maintained interaural time differences (ITD) as well as small ILD values. In an SRT evaluation, he found no advantage due to dichotic presentation of the ILDs as compared to the diotic stimulus. However, he observed an SRT advantage due to the ITDs of 1.6 dB with the binaural endfire array, which is in consideration of the above-mentioned tradeoff between practically achieved directivity gains and the binaural cue-preservation, of an expected magnitude.

2.2.2 A review of the state-of-the-art of beamforming solutions

If we survey current solutions of binaural beamformers on a global level, one may differentiate fixed and adaptive beamformers as well as bilaterally independent applied arrays and head-based arrays. Classical fixed schemes often consist of small endfire arrays in BTEs, made up of two to three microphones over a distance of about 2 cm, feeding a fixed gradient or MVDR beamformer processing scheme. The above-mentioned hearing glasses (Varibel Innovations BV) employ two independent arrays in endfire orientation and an MVDR scheme with four microphones that are non-uniformly distributed over a length of 7.2 cm (Boone, 2006). Other solutions extend the array over bilateral BTEs through a binaural link (Rohdenburg, 2008). These bilateral head-related arrays need to account for the propagation model of the array mounted on a head. Hence, these systems ideally employ individually measured head-related transfer functions (HRTF). However, the resulting HRTFs are found to be susceptible to the positioning of the array system. Together with the suboptimal optimization of fixed MVDR filters for real-world applications, this leads to processing artifacts (Rohdenburg, 2008).

A solution to alleviate this instability was introduced: improving the beamformer robustness. However, a tradeoff with the WNG constraint leads to less noise suppression. Other approaches to circumvent the instability are the application of parametric head-models and the implementation of an adaptive MVDR beamformer in a GSC framework, a method pursued by (Rohdenburg, 2008). Both concepts will be described briefly.
The recording of individual array-based HRTFs is not acceptable for practical application, because the individual audiological aid is constrained by the lack of time and money. Therefore, head-models of different complexity could be employed for future head-related array systems, which capture the most important characteristics of an individual head (Lotter and Vary, 2006; Rohdenburg, 2008). Although the applied models to date only permit a weak compromise with respect to measured HRTFs, the approach may result in more favourable results in the near future as parametric head-models improve. See, e.g. Fels (2008) for recent developments.
Adaptive solutions of head-related array systems using a GSC structure are often based on a delay-and-sum solution in parallel with an adaptive path, or on an MVDR solution in parallel with an adaptive path (Simmer et al., 2001; Rohdenburg, 2008). Therein, the adaptive path captures an estimate of the noise field and adjusts the spatial nulls of the beampattern towards the directions of coherent interferers. The delay-and-sum solution in parallel with the adaptive path structure is expected to be less effective in suppressing incoherent signals (Simmer et al., 2001). In theory, the fixed MVDR solutions in parallel with an adaptive path structure is expected to achieve a considerable suppression of coherent and incoherent noise. Unfortunately, in realistic conditions, the advantage of such an adaptive structure remains below expectations due to stearing errors and the above-mentioned noise sensitivity (Greenberg and Zurek, 2001).

Rohdenburg (2008) compared the fixed and adaptive (GSC) MVDR beamformer with a post-filter based on interaural coherence. His findings are summarized:

- As compared to the fixed processing scheme, a small SNR improvement of 2 dB, using an instrumental measure, was found for the adaptive structure in optimal conditions using HRTFs instead of head-models. However, it showed a much lower stability towards inevitable steering errors and other factors that perturb the propagation model. When employing a parametric head-model the advantage was eliminated, even under perfect steering conditions.

- The performance of an adaptive procedure depends on the sound field conditions. In real-world scenarios, when background babble noise and multi-path direct sound propagation occur, the adaptive approach is no longer well determined, i.e. the number of coherent noise sources impinging on the array is greater than the number of microphones in the array.

- A reliable detection of speech pauses in speech can improve the adaptive processing. Gap detection (VAD), however, proves to be difficult in critical real-world

conditions.

Also bilaterally applied fixed MVDR beamformers suffer from positioning displacement and other deviations from the propagation model that was used during their optimization. This was also reported by Merks (2000), who found that the beampattern of a free-field optimized endfire array is perturbed at high frequencies by the head. For this reason, he compared the binaural endfire array performance based on a free-field propagation model optimization and a head-mounted optimization using an artificial head. By means of an evaluation in a simulated diffuse noise field with a female speech shaped noise colouration, he could not confirm a subjective benefit due to the more accurate propagation model. Besides, the bilaterally applied endfire array showed a robust and high SNR gain in this noise field, which differed from the noise field in which the beamformer was optimized.

Recent advancements incorporate a localizer, or in general terms, a scene classifier. For instance, Rohdenburg (2008) developed a head-mounted array with an adaptive target-tracker based on a fixed MVDR beamformer and a coherence-based post-filter. Using instrumental measures, an improvement of signal quality as well as speech intelligibility over the same system without target-tracking was found, even when the target is moving relative to the head-based array. In a similar fashion, Boone et al. (2010) developed a system based on a combined processing scheme of a bilaterally applied MVDR beamformer and a binaural CASA-based post-filter. Therein the CASA-based localizer of Albani et al. (1996) served as a scene classifier that triggered the aperture of the post-filter with a data-driven Bayesian classification approach. Based on the complexity of the scene, the aperture of the post-filter was adapted to guarantee an optimal signal quality, at the same time with a speech intelligibility enhancement. The underlying concept of using the parametric output of a localizer to estimate varying soft-gains for a high quality output had been previously introduced by Madhu (2009b).

To roundup this review, mainly two approaches of binaural and constrained MVDR beamforming solutions that allow for a combination with a post-filter have been developed. One is based on bilateral endfire arrays that are connected via a binaural link to extend the array dimensions. These systems need to take the diffraction of the head into account by applying a propagation model. The other approach is based on bilateral MVDR beamformers that work independently and do not require a propagation model of the head.
Head-based arrays are a promising approach for future application in small BTE hearing aids. What these require are suitable propagation models, including individualized HRTF approximations. Adaptive implementations using a GSC structure of head-based systems did not demonstrate an advantage over fixed systems in real-world applications. However, a head-based array system with a target tracker demonstrated an improved speech intelligibility. Additionally, the head-based systems of Lotter and Vary (2006) and Rohdenburg (2008) integrate a post-filter that employs the binaural waveform coherence as a classifier of diffuse noise. These

systems can be considered as approximations of the MMSE solution. Using two bilateral endfire arrays that establish separate beams, offers the advantage of a lower susceptibility to deviations from the propagation model used in the optimization of the MVDR filters, and positioning errors (Merks, 2000). Constrained superdirective endfire solutions are characterized by a simple, robust and efficient processing. As a result, these are already successfully applied in today's hearing aids. Moreover, using a non-adaptive beamforming structure, i.e. neither adaptive "nulling" towards interference nor a beam steering towards the target speaker offers the advantage of a stable binaural image, which can be exploited by a binaural post-processor. Moreover, as compared to algorithms in which binaural cues constantly change, a hearing aid with stable binaural cues is considered to contribute substantially to the listening ease.

This work primarily studies the improvement of speech intelligibility that can be obtained with a set of binaural CASA-based post-processors at the output of different bilateral front-ends in various acoustical conditions. Therefore, it is the general approach of this study to create the boundary conditions of the analysis of the post-processors as realistic as possible. For that purpose, exclusively genuine off-the-shelf hearing aids are applied as beamforming front-ends. The following section presents and analyzes the choice of hearing aids applied in this work.

2.2.3 Analysis of three bilaterally applied beamformers

In this thesis, three examples of real-world superdirective front-end solutions are given, which are all based on bilateral endfire arrays featuring separate beamformers. These are a commercially available BTE with and without a directivity mode (GN ReSound type Canta 470-D) and the commercially available hearing glasses (HG) in two settings, one of moderate directivity and one of high directivity (Varibel Innovations BV).
A measurement of the DI with the hearing aids mounted on an artificial head (KEMAR manikin, Knowles Electronics) in anechoic conditions showed a speech intelligibility weighted DI[2] of 1.3 dB, 4.4 dB and 7.2 dB for the BTE (directivity mode) and the HG in the low and high directivity mode, respectively (Boone, 2006). During the development of the HG in the high directivity mode, Merks (2000) conducted several evaluations and found an audiological benefit of 6.2 dB for hearing impaired people and an improvement of 7.5 dB for people with normal hearing in a diffuse noise field. Hence, the hearing glasses realize more than a 5 dB SNR improvement and, consequently, can be considered as being beneficial for the majority of hearing impaired people (Duquesnoy and Plomp, 1983).

In the following, the three directional filters are reassessed with a set of physical measures and compared to the unaided case. To that purpose, the hearing aids were

[2] $\mathrm{DI} = \sum_{\mathrm{i}}^{\mathrm{N}_{1/3}} \delta_{\mathrm{i}} \mathrm{DI}_{\mathrm{i}}$, with δ being the band importance weights per one-third octave band between 500 and 5000 Hz that are taken from ANSI/ASA (2007).

applied without an audiological profile, a frequency-independent amplification gain and no compression (the electrical signal was tapped prior to the receiver of the hearing aids). Furthermore, the aids were mounted on a non-symmetric artificial head system, the ITA mannequin head of Schmitz (1995), which is the head system applied in this work. In the following the ITA mannequin will be referred to as the Aachen head. See Section 3.2 for more details.
Impulse response measurements were performed in an anechoic chamber. Head and torso resided on a rotating turn table, which was remotely controlled to record impulse responses at both ears at steps of one degrees ($\Delta\theta = 1$ deg) in the horizontal plane, i.e. zero degree elevation ($\vartheta = 0$ deg). For details on the measurement procedure see Chapter 3.2.

In a first analysis, the transfer functions of the hearing aids and the artificial head (without an ear channel simulator) for frontal sound incidence, $\theta = 0$ deg, are given in the upper plots of Figure 2.3. The transfer functions show a frequency-dependent behavior. Also the Aachen head offers an improved frequency-transfer at frequencies higher than 2 kHz, due to the shape of the pinna. The observed high-pass roll-off of about -25 dB/decade below 1 kHz at the output of the BTEs, is responsible for attenuating the self-noise of the gradient-solution at low frequencies, as well as low-frequency environmental noise. In comparison with these transfer functions, the HG in the low and the high directivity mode show a rather frequency-independent transfer of sound energy from the frontal direction. In addition, a deviation between the left and the right channel can be observed at the output of each front-end as a result of the unmatched microphones and asymmetries in the measurement setup and the head-shape.

In a second analysis the front random (FR) index for frontal incidence, i.e. $\theta = 0$ deg, of the hearing aids and the artificial head are measured. The logarithmic two-dimensional $\mathrm{FR}_{\theta=0}$ is:

$$\mathrm{FR}_{\theta=0}(d) = 10\log_{10}\frac{|h(d,\theta=0)|^2}{\frac{1}{N_\theta-1}\left(\sum_{u=1}^{N_\theta}|h(d,\theta)|^2\right)}, \tag{2.2.13}$$

in which N_θ is the number of azimuthal measurement positions (Merks, 2000). Hence, the FR determines the ratio of the squared array response in target direction to the averaged squared array response due to all-sided sound incidence. Generally, the FR equals the DI if the array's target response equals the most sensitive array direction. However, in contrast to the DI definition in Equation (2.2.10), which represents a theoretical formulation, the DI for measurements is defined as the ratio of the maximized squared array response with respect to the angles θ and ϑ, to the average squared array response due to omnidirectional sound incidence (Merks, 2000). If head diffraction or processing errors introduce off-axis maxima, the DI calculation method will incorporate these and, hence, deviate from the target direction. Therefore, in the present work, the FR is chosen with $h(d,\theta=0)$ in the numerator of Equation (2.2.13) to quantify the target gain of frontal direction in the

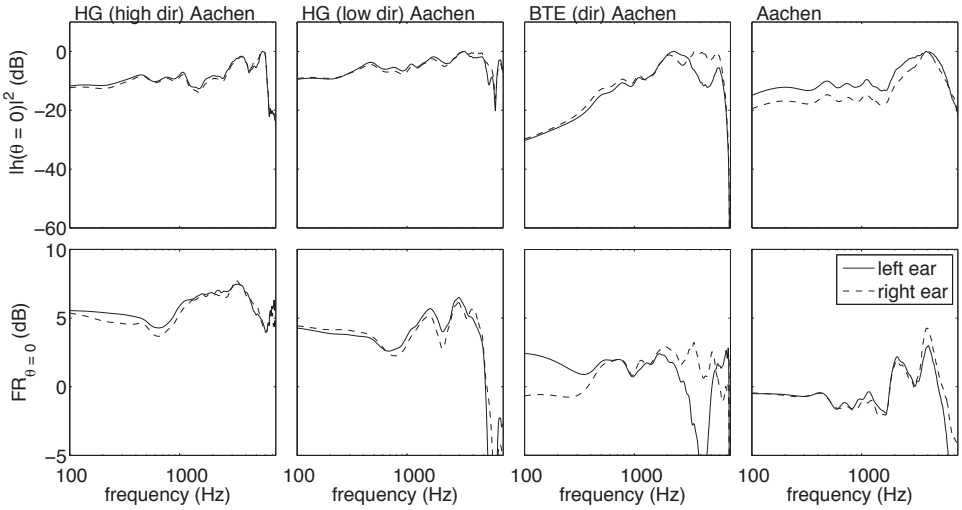

Figure 2.3: *Left and right ear transfer functions (dB) and the FR indexes (dB) at $\theta = 0$ deg for the hearing glasses (HG) in two different directional modes and the applied BTE hearing aid in a directional mode (all mounted on the Aachen head). The results are contrasted with the unaided case, by tapping the binaural output of the Aachen head mannequin.*

left and the right channel. The second row of plots in Figure 2.3 gives the results. As expected, the $FR_{\theta=0}$ is highest for the HG in the high directivity mode. Except for high frequencies, a congruent but attenuated curve is observed for the $FR_{\theta=0}$ of the HG in the low directivity mode. However, at high frequencies the $FR_{\theta=0}$ decays. The $FR_{\theta=0}$ of the BTE shows strong fluctuations. The imbalance might be a result of variations in the directional behavior of the applied microphones and asymmetries in the mounting and alignment. An analysis of the FR of the same BTE hearing aid in the omni-directional mode showed only small differences between the left and the right channel. This comparison supposedly isolates the directional processing as being the reason for the observed imbalance of the left and right FRs in the directional mode of the BTE. Moreover, listening to the output of the hearing aids revealed a considerable internal noise level. Despite these processing errors, the BTE in the directional mode shows a $FR_{\theta=0}$ of about 1 to 3 dB in a range from 0.5 to 2.5 kHz, due to the applied first-order gradient processing.
The $FR_{\theta=0}$ of the Aachen head hovers around 0 dB and below, until approximately 2 kHz, when the directivity is shown to rise due to the shape of the pinna.

To gain a better understanding of the beam-pattern, the two-dimensional FR index was subsequently calculated for all azimuthal directions and is depicted in Figure 2.4. The approach offers a two-dimensional beampattern analysis, which provides an

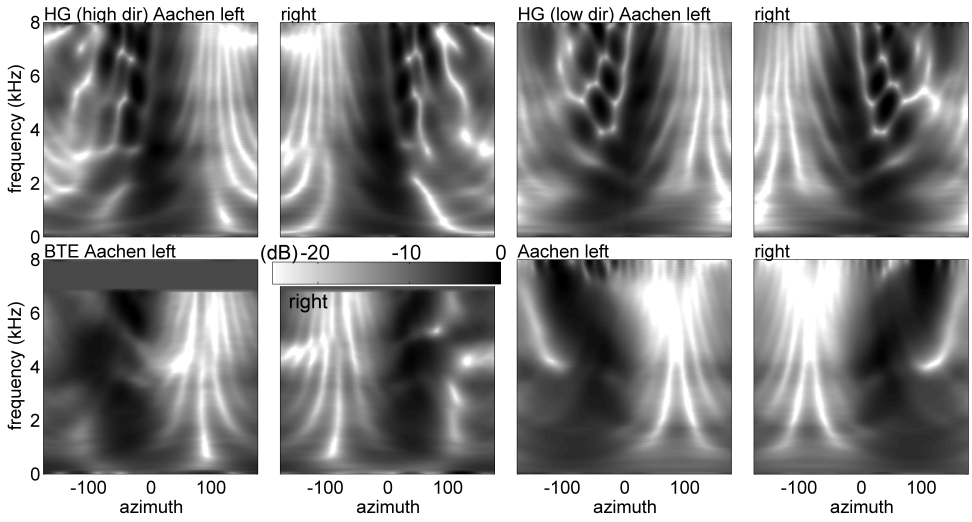

Figure 2.4: *The FR-based directivity patterns (dB) in the horizontal plane as a function of azimuth and frequency of the two-channel front-ends applied in this work are given. All hearing aids are mounted on the Aachen head. For a comparison with the unaided case, the directivity of the Aachen head is given. The FR of each plot is normalized to 0 dB. As regards the BTE front-end, the directional programme mode is shown. Because of the limited frequency transfer of this font-end, the FRs are only calculated until 7 kHz.*

overall qualitative insight into the directivity of the front-ends.
Beginning with the HG in the high directivity mode, a high degree of sensitivity around the median plane ($\theta = 0$ deg) is observed, for the left and the right beam-pattern. However, the patterns are broadened as compared to the initial results of Merks (2000). This finding is likely a result of the practical implementation, i.e. restricted filter and digital word length, but may also be due to the Aachen head. This head mannequin is larger—about 2 cm in diameter—than the KEMAR head, which was used at the time of the HG development. Already Merks (2000) had analyzed a broadening of the main-lobe, the closer the temples are situated with respect to the head. Moreover, Figure 2.4 shows how the frequency-dependent directional response deviates from a high gain in look-direction due to the diffractions around the head from above 3 to 4 kHz. At high frequencies, the directional characteristics of the applied microphones can no longer be assumed omni-directional, which makes the free-field beam-pattern susceptible to slight deviations throughout the propagation path.
Furthermore, the decline of directivity at high frequencies of the HG in the low directivity mode that was illustrated with the $\mathrm{FR}_{\theta=0}$ in Figure 2.3, can be explained by observing the respective beam-pattern. As Figure 2.4 shows, the beam-pattern

extends broadly along the median plane and splits up in spatial notches above approximately 4 kHz. These spatial notches at high frequencies coincide with the $FR_{\theta=0}$ analysis direction. The DI, which maximizes the array response in the numerator of Equation (2.2.13), would consequently be higher. Note that the DI is used as a cost-function in the optimization process of the MVDR filters with the aforementioned constraints of maximum noise sensitivity and a distortionless transfer function for the target-direction.
The filters of the here applied HG have been optimized in a simulated free-standing setup, without being mounted on a head mannequin. As previously mentioned, despite this deviation from real-world application, the approach has been found to be equally efficient and more robust as the optimization including the head-model. Practical measurements showed that head-related MVDR filters depend considerably on the mounting and the peculiarities of the head-model (Merks, 2000).
Non-optimal beam-patterns are observed with the bilaterally applied BTEs in the directional mode. The maximum of the directivity turns out to be off-axis, at approximately -45 and 45 deg for left and right ear, respectively. In the range of 3.5 to 5.5 kHz, the beam-patterns clearly differ. At these frequencies, the left ear BTE shows a lateral deviation of the main-lobe, while the right ear BTE offers the main-lobe at 45 deg. Consequently, the deviation of the FRs at $\theta = 0$ deg in Fig. 2.3 is found to be a consequence of the beam-pattern and not a consequence of a general absence of directivity.
Regarding the Aachen head, and as previously observed, directivity shows to be building up above approximately 2 kHz, due to the focussing direction of the pinnae. Furthermore, at higher frequencies, the main sensitivity is found laterally, supposedly due to waves of small dimensions exciting the diaphragms of the microphones, which reside at the entrance of the ear-channel, most efficiently out of the perpendicular direction. Consequently, the observed directivity should differ from the directivity when using genuine HRTFs.

In summary, three variants of real-world bilaterally applied beamformers have been studied. The directional filters realize a compromise between super-directivity and robustness and can be considered as possible front-ends for a future application of binaural CASA post-filters. Whereas the HG showed an advantage in terms of symmetry and frequency-transfer into the looking direction, the directional processing of the BTE is clearly impeded by the placement of the microphones behind the pinna and by an increased noise-sensitivity of the gradient-method.[3] Moreover, asymmetries due to mounting differences and, as found in the comparison with the omnidirectional programme mode, unmatched microphones, result in substantial differences between the left and right ear beam-pattern. Finally, the self-noise amplification at low frequencies of the gradient-processing requires a high-pass filter, which produces an artificial sound character and, more important, decreases speech

[3]The noise sensitivity of the BTE in the directivity mode has not been measured, but was qualitatively experienced during the measurement of the HRTFs by a low internal SNR of the BTE in the directivity mode, hence, a high self-noise.

intelligibility in the low frequency range.

The presented beamforming solutions sample the range of today's array processing in commercially available hearing aids. A challenge exists in the adaptation of the CASA post-filter to the peculiarities of the interaural transfer functions of the different front-ends. In Chapter 3.3.3 a statistical method will be introduced that allows for adaptation in an efficient way. Up to this point, the beamforming front-ends of the combined processing scheme have been introduced and analyzed. In the second part of this chapter the CASA-based post-filters will be studied.

2.3 Varying filter-gain functions

This section is divided into four parts. The first part introduces the concept of real-valued gain functions that are established in binaural CASA-based post-filters, which are used later in this work. These gain functions are binary or smooth, i.e. well-known as ideal binary masking, or soft-masking approaches, respectively. The well-established concept of the Wiener filter is the linear variant of the soft-mask approach. It requires perfect knowledge of the noise power as well as a perfect analysis and synthesis filtering technique. Needless to say, these requirements are not so stringent in reality. Consequently, the Wiener gain can only be approximated. In a second part of this section, soft-masks and ideal binary masks are compared. In a subsequent subsection, a statistical analysis is applied to the local (or bin-wise) SNR values in the STFT domain. This allows us to gain an insight into the complexity of the speech-in-noise problem across different conditions. The last part of this section deals with the suppression of the musical noise phenomenon, which is an inevitable artifact of time-varying real-valued gain functions. To this purpose, the cepstral smoothing technique will be introduced.

2.3.1 The Wiener filter approach in the STFT domain

The post-filter defined in Equation (2.1.6) has been established as a single channel Wiener filter. It can be calculated with the Fourier Transform of the single channel MMSE solution (see Appendix A.1). Depending on a certain SNR, the Wiener filter realizes the optimal single channel filter. Based on the power spectral density estimates of signal and noise, the single channel Wiener post-filter is a spectral gain function and it introduces signal distortion if $\phi_{\mathrm{vv}} > 0$. Therefore, the multi-channel MMSE solution, and the here performed factorization into a MVDR beamformer and single channel post-filter, does not imply a distortionless reconstruction of the original signal.
The strength of the distortion can be constrained by introducing a multiplier ν as (Madhu, 2009a):

$$w_{\mathrm{post}} \,|_{\mathrm{stab}} = \frac{\phi_{\mathrm{ss}}}{\phi_{\mathrm{ss}} + \nu\phi_{\mathrm{vv}}}. \tag{2.3.14}$$

Any value of ν between zero and one leads to a particular balance between distortion and interference suppression. As a result, a speech distortion weighted Wiener filter can be formulated. Recent advances use varying distortion weights by calculating the short-time probability of speech presence. Hence, these approaches apply a lower value of ν in periods when the target is absent and noise dominates (Madhu, 2009a). An alternative to this distortion constraint method is well-known as flooring by a lower-bounding of the gain-function at $\max[A_{\min}, w_{\text{post}}]$, in which $A_{\min}$ is the lower bound of the final gain.

Originally Wiener's approach was a linear and time-invariant filter that was based on the assumption of random stationary processes (Simmer et al., 2001). Stationarity, however, does not apply to speech. The requirement can be alleviated by the fact that speech offers short-time stationarity. Block-wise analysis-synthesis filterbank approaches are appropriate for short-term magnitude processing of speech signals. Critical filterbanks, i.e. filterbanks with bandwidths that are approximately equally distributed on a logarithmic frequency scale (hence, broadly comparable to the tonotopic organization in humans), allow for a high resolution at low frequencies. Due to their physiological and psycho-acoustical resemblance, critical filterbanks are generally found in algorithmic approaches that aim to approximate certain percepts, as e.g. speech intelligibility. In speech enhancement tasks, the direct DFT-IDFT[4] transformation method is widely preferred for reasons of computational efficiency and a higher spectral resolution at mid to high frequencies.

Generally, when processing speech with a short-time filterbank approach, a trade off has to be found that realizes the highest possible frequency resolution while it must not violate the short-time stationarity of speech. Motivated by the determination of such an optimal frame length for speech processing in the STFT domain, Paliwal and Wojcicki (2008) found that a window-length of 15 to 35 ms leads to optimal speech intelligibility when speech is reconstructed from the short-time magnitude spectrum. If the analysis frames are too long, the short-time stationarity of speech can no longer be used. At the opposite extreme, i.e. if the frames are too short, the spectral estimates of power spectral densities become less consistent, which is caused by the stochastic nature of the speech signal. In addition, the shortest analysis frame-length is dictated by the lowest harmonic, i.e. the fundamental pitch frequency that has to be resolved. To yield a stable estimate of the power spectral densities, usually two to three times the pitch period should be accommodated in a frame (Paliwal and Wojcicki, 2008). For example, a pitch-frequency range from 80 to 400 Hz has a period range from 2 to 12 ms.
Regarding the design of hearing aids, the system delay and the computational additionally load constrain the processing method. For instance, longer frames lead to less computational load but increase the system delay. Altogether, the multiple conflicting objectives that underly the time-frequency processing furthermore hamper an ultimate Wiener filter solution in speech enhancement approaches.

[4]IDFT denotes the inverse discrete Fourier transform. See Appendix A.2 for an definition of the here applied DFT-IDFT filterbank method.

2.3.2 Soft-masks versus ideal binary masks

The term mask originates from the psycho-acoustical phenomenon of masking, that describes the process or the amount by which the threshold of audibility of one sound is raised by the presence or the masking of another sound (Moore, 2003). This auditory concept has been transferred to gain-functions, in which a mask is created to isolate the target signal from a mix (Wang and Brown, 2006).[5] Accordingly, a mask based on the Wiener rule suppresses interference proportional to the ratio of signal to noise. This particular mask is known as a linear soft-mask, i.e. $\mathcal{M}_{\text{soft}} = w$. The introduction of non-linearity, i.e. by compression or an expansion of the gain-values, may increase interference suppression, and results in graduations of non-linear soft-masks. The utmost non-linear gain function is well-known as the ideal binary mask (IBM), which can be formulated as:

$$\mathcal{M}_{\text{IBM}}(d,n) = \begin{cases} 1 & \text{if} \quad 10\log_{10}\left(\frac{\phi_{\text{ss}}(d,n)}{\phi_{\text{vv}}(d,n)}\right) > \varepsilon \\ 0 & \text{otherwise} \end{cases}, \tag{2.3.15}$$

in which ε is an SNR criterion that can be chosen a priori or according to the global SNR of the mix. If ε is equal to a fixed SNR-criterion, the IBM is dependent on the SNR. If, however, the ε is determined by the mixing SNR, one speaks of a local criterion, and the IBM is independent of the SNR.
Figure 2.5 shows a comparison of a linear smooth mask and an IBM for a mix of three vowels in stationary speech shaped noise at 0 dB, with a local criterion ε of 0 dB. As can be seen, the IBM results in a unit gain clustering around the harmonics. The linear soft-mask, on the other hand, scales the target power according to the power ratio at each time-frequency bin. Listening to the two weighted results shows an advantage for the IBM, which offers in this setup a higher quality perception due to the unit gain for most of the harmonics and a rigorous interference suppression. In a different setup, however, when the target speech power is softer and more widely distributed over the time-frequency plane, the IBM might suppress transitional parts of speech that particularly contribute to intelligibility.

A controversy exists in the field of speech enhancement about the optimal mask approach. The IBM concept offers the advantage of a high-level interference suppression. Therefore, the method requires a priori knowledge of the power spectral densities at each time-frequency bin. However, such knowledge is not available outside the laboratory. Due to its considerable separation power, IBM masks have frequently been suggested as a ceiling measure to define the ultimate CASA goal (see e.g. Wang and Brown, 2006). Based on a broadband interference, IBMs even allow for a full recovery of speech intelligibility in an SNR situation of -60 dB

[5] With regard to terminology it suggests itself to replace the widely familiar term 'mask' for varying filter-gain functions in CASA approaches with 'pattern'. This would also be more consistent for the terminology of this thesis, where the term 'pattern' is used in the description of the binaural classification approach. Nevertheless, in order to omit confusion with literature, the term 'mask' is adopted for describing the varying filter-gain functions in this thesis.

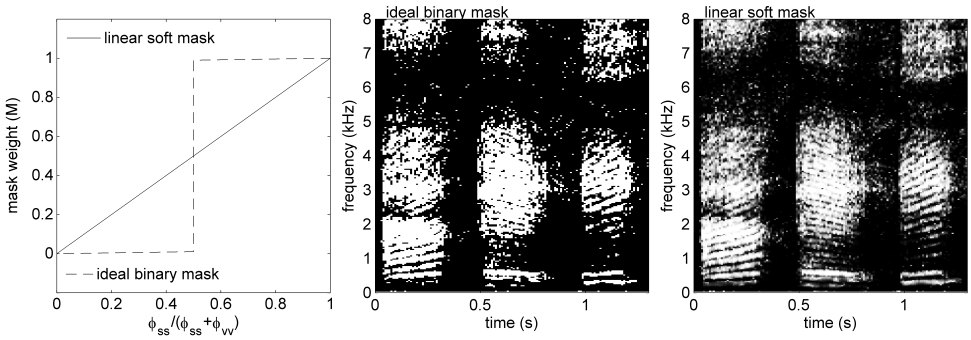

Figure 2.5: *An ideal binary mask (IBM) (middle plot) and a linear soft-mask (right-hand plot) with the associated gain functions (left-hand plot) in the STFT domain for the three vowels /a/, /e/ and /i/ are given. Signal and noise were mixed at an SNR of 0 dB. The IBM has a local criterion ε of 0 dB and the noise is stationary and equal to the long-term spectrum of speech.*

(Kjems et al., 2009). This is the effect of noise gating, also known as the vocoder principle. The IBM method can be further improved. In a recent study, Kim and Loizou (2010) showed that an overestimation of the noise power may result in a considerable improvement of speech intelligibility using IBMs. Nonetheless, the problem of estimating the power spectral density of the noise remains as a real-world challenge. Therefore, any error in this estimation process is amplified by the IBM hard-clustering approach. In addition, Madhu et al. (2010) demonstrated the superiority of the Wiener rule or linear soft-masks to the IBM method, in terms of speech intelligibility. In their work it is shown that IBM using a local SNR criterion generally offers a much higher intelligibility output than the IBM with a fixed SNR criterion. Both concepts, however, by far cannot achieve the intelligibility gains of the Wiener filter, especially in situations when the power of the target signal and the interference are allocated in equal time-frequency regions. Given an accurate estimation of the power spectral densities, the Wiener gain allows for full intelligibility under SNR conditions as low as -35 dB in babble noise and single talker interference conditions (Madhu et al., 2010).

At a later point in the present study, soft-masks will be derived based on spatial classification. Subsequently, these soft-masks will be optimized using an instrumental measure of the binaural speech intelligibility. This approach is different from the general method of optimizing the MMSE criterion, such that it allows for an optimal intelligibility-based handling of distortions, which are inherent to the single-channel filtering process (see e.g., Loizou and Kim, 2011).

2.3.3 Implications of the SNR calculation method

An important measure to assess a particular speech-in-noise problem is the analysis of the local SNR, i.e. the SNR_l calculated per time-frequency bin. Regardless of this fact, throughout most comparable studies, the SNR is calculated from the root-mean-square (RMS) power of the temporal waveforms. Thereby a global SNR calculation method offers a common reference, everybody is familiar with. The global SNR, however, can only approximately define the speech-in-noise problem. In this subsection, the implications of the global SNR mixing method are briefly studied.

In order to analyze the distributions of the signal power in the time-frequency plane at a typical global SNR level, four speech-in-noise conditions were generated and the local SNRs were analyzed. The speech signals used consisted of zero mean, RMS-normalized and phonetically balanced sentences of the TNO-corpus in the Dutch language (TNO, 2000). The sentences were concatenated and silent periods were excluded with the application of a simple VAD algorithm. The signals were convolved with the HRTF of the Aachen dummy head for frontal incidence at the left ear, lowpass filtered at 8 kHz and digitized at a sampling frequency of 16 kHz. Subsequently, the time series were segmented in blocks of 256 samples and Hanning-weighted. Each block was padded with zeros to yield a vector of 512 bins. Consequently, a DFT was applied to generate spectro-temporal signal representations. Throughout all speech-in-noise conditions, the global SNR was set to -5 dB, based on the long-term RMS levels of the waveforms. In order to make the distribution of the signal power visible, a hard clustering of signal power was generated using the IBM method with a fixed criterion ε of 6 dB. To calculate the IBMs as in Equation (2.3.15), the power spectral densities were calculated with a first-order recursive smoothing technique. Hence, the power spectral density of the signal is determined from:

$$\phi_{\mathrm{ss}}(d,n) = \alpha\phi_{\mathrm{ss}}(d,n-1) + (1-\alpha)s(d,n)s^*(d,n). \tag{2.3.16}$$

For the noise we have:

$$\phi_{\mathrm{vv}}(d,n) = \alpha\phi_{\mathrm{vv}}(d,n-1) + (1-\alpha)v(d,n)v^*(d,n). \tag{2.3.17}$$

The variable $\alpha = \exp\left(-\Delta T/\breve{\tau}\right)$ is a smoothing constant that depends on the filterbank frame-shift ΔT and the time constant $\breve{\tau}$. The parameter $\breve{\tau}$ is typically in the range of 8 to 30 ms. In this experiment, $\breve{\tau}$ was set to 8 ms. Speech enhancement in hearing aids benefits from such a recursive filtering, because it introduces no additional system delay.

Four noise conditions were created. The target speaker was a female speaker, and the target signal was mixed with:

A a male speaker,

B two male speakers,

C speech babble of a lively canteen,

D speech babble of a lively canteen with additional reverberation[6] of RT = 0.4 s, applied to the target signal using the mirror image source model software (MISM) of Van Dorp Schuitman (2009).

Fig. 2.6 shows the resulting signal power distributions in the time-frequency plots as well as the probability density functions (PDFs) of the local SNRs. The PDFs were generated using time series of 15 s length. For each PDF, the standard deviation σ and the mean μ were calculated. Additionally, PDF portions of bins with an amplitude ratio of more than 2, hence, a local SNR greater than 6 dB, were calculated.

Figure 2.6 gives the results. Beginning with condition A, one finds a homogeneous distribution around $\mu = -5$ dB. Local and global SNR correspond in this condition. The standard deviation of $\sigma = 18$ dB indicates that the two signals occupy different regions in the observed time-frequency plane. This conclusion is verified by inspecting the IBM of condition A, in which target signal components cluster around dominant signal portions. About one third of all bins hold a local SNR of more than 6 dB.
If in condition B two speakers interfere at the global SNR of -5 dB, σ of the local SNR decreases to 17 dB and μ shifts to -8 dB. Only one-fifth of the time-frequency bins offer a local SNR ≥ 6 dB. Compared with condition A, the speech-in-noise problem is more difficult, although both conditions can be characterized with the same global SNR value.
In condition C, i.e. the babble background of a lively canteen at a global SNR of -5 dB, σ decreases to 14 dB and indicates that the signals are no longer separated in the observed time-frequency representation. Regions of high signal power appear as sparsely scattered outliers in the IBM mask. The local SNR has a μ of -16 dB, which reveals a difference of more than 10 dB in local SNR, as compared to condition A. Only 8 % of the mix offers a local SNR of more than 6 dB.
Finally, condition D, which is equal to condition C apart from the fact that reverberation was applied to the target signal, shows a μ of -12 dB and a σ of 13 dB. The reduction of the standard deviation is a result of the reverberation. As the reverberated target signal is smoothed, its dynamic range decreases and the PDF of the local SNR has a narrower distribution. Although μ increases—as compared to condition C, only about one-tenth of the bins offer a local SNR greater than 6 dB.

Overall, the experiment describes a series of issues that have to be considered when describing a particular speech-in-noise problem. First, single and coherent speech interferers are generally well separated from the target signal in a time-frequency representation. Consequently, these can be separated efficiently. Secondly, as the

[6] RT is the reverberation time. It is defined as the length of time for which the sound pressure level attenuates to a relative value of -60 dB, after the driving sound source is switched off.

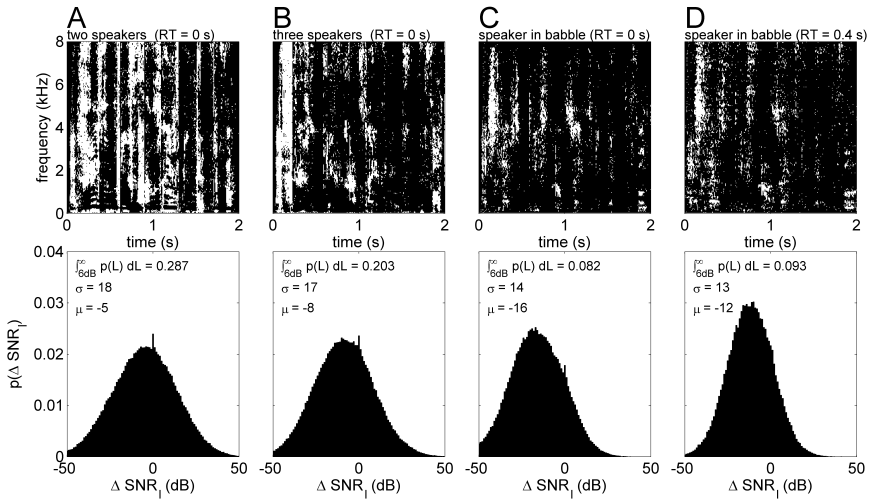

Figure 2.6: *Four speech-in-noise conditions, each mixed at a global SNR of* -5 *dB. The first row shows IBM masks of the mixtures. A fixed mask criterion* $\varepsilon = 6$ *dB is chosen. The bottom row shows the associated PDFs of the local SNRs (*ΔSNR_{l}*). In each plot the cumulative probability for a local SNR greater than* 6 *dB, the PDF standard deviation,* σ*, and the PDF mean,* μ*, are given.*

background becomes more continuous, e.g. a white spectrum, the separability decreases and the local SNR substantially falls off relative to the global SNR. Hence, the global SNR is a limited measure to describe the speech-in-noise problem. Finally, if reverberation is applied to the target signal, its dynamic range tends to decrease. Consequently, the signals become evenly mixed, i.e. most time-frequency bins share approximately equal signal portions. As a result, signal classification and enhancement is complicated.

The problem of the long-term waveform-based SNR is well-known. The crest factor, defined as $o(\iota) = \max(|x(\iota)|)/x_{\mathrm{RMS}}(\iota)$, represents one physical solution to the problem by calculating the ratio of the maximum signal value to its RMS (Zwicker and Zollner, 1987). Thereby a small o indicates a smooth signal and the transmission of a constant power-rate, whereas a high o indicates a spiky signal and a varying power transmission. As has been shown, a local SNR measure constitutes a better criterion for assessing the speech-in-noise problem. At a later point in this work, a segmental SNR measure (coarser than the local SNR), calculated across critical bands and time-frames of 32 ms length will be used together with a measure of speech intelligibility, to assess the algorithmic signal enhancement.

With respect to the composition of different sound scenes in the present work, the global SNR method will be adopted, despite its limited explanatory power of the speech-in-noise problem. However, as mentioned, the global SNR calculation method facilitates a comparison with other studies, for instance with the work of Nix and Hohmann (2006) in the following chapter. Where possible throughout the assessment in Chapter 5, the global SNR will be referred to an objective measure of speech intelligibility. It was shown by Miller (1947) that similar deviations as described above between the local and the global SNR, exist between speech intelligibility and the global SNR measure.

A term that has frequently been used in the previous experiment is separability of signals in the time-frequency domain. If the signal are disjoint,[7] the speech intelligibility of the model hearing system is considerably supported by the possibility of glimpsing in pauses, harmonic grouping as well as clustering across time and frequency. The more complex the background gets, in the experiment above, going from condition A to condition D, the stronger the smearing of different signals over the time-frequency atoms becomes. Based on an IBM mask, dominant signal power was shown to disperse into a sparse cloud, under complex conditions, and no clustering across neighboring bins was observed. Hence, in the STFT domain, the disjointness assumption is generally only justified for two speakers. For the most part, this condition is attributed to the high dynamic range and the compressive representation that the dynamic range of each of the two speaker signals requires (Barker et al., 2000). Mathematically, disjointness can be expressed for speaker $s_\jmath(d,n)$ and speaker $s_{\jmath'}(d,n)$ as (Madhu, 2009b):

$$s_\jmath(d,n)s_{\jmath'}(d,n) \approx 0 \qquad \forall \jmath' \neq \jmath. \tag{2.3.18}$$

In summary, this subsection presents a statistical experiment that illustrated two important realities if speech intelligibility is to be enhanced. The general temporal waveform-based calculation method of the SNR offers little information on the actual difficulty of the speech-in-noise problem. A better measure to describe the complexity of a scene is the mean SNR at the local time-frequency bin scale. However, as will be shown later in this work, also the local SNR is not strictly correlated with speech intelligibility.
The second important fact inferred from the experiment is the absence of signal disjointness in the STFT domain in most real-world noise conditions. This substantially complicates the estimation of the noise power at low SNRs in single channel applications (Barker et al., 2000). It is one of the research questions of this work, whether further signal decompositions, as e.g. an analysis of interaural disparities of the binaural signal, allow for a robust signal classification, even in complex and low SNR conditions.

As it was shown in the last two subsections, mask-based single-channel filtering in-

[7] Note the concept of disjointness is applicable to other domains. Later in this work, a statistical analysis will be applied to study this property in the binaural time-frequency (centre frequencies) and the binaural temporal centre and modulation frequency domain.

troduces distortions, irrespective of whether the Wiener filter or the IBM method is applied. Furthermore, these distortions increase if the noise power is estimated from the mixture in real-world situations. Faulty gains in the masks result in the well-known musical noise phenomenon. The next subsection deals with the attenuation of this kind of distortion.

2.3.4 Cepstral smoothing of masks[8]

As stated before, the success of varying gain-filter functions, or mask approaches, is subject to the quality of the estimation procedure of the signal and noise powers. When applying these varying gain-filter functions for noise suppression, musical noise, or narrow-band bursts, of short duration are an inevitable consequence of the filtering process (Breithaupt and Martin, 2008). Also, when masks are determined in multidimensional signal spaces, e.g. by using binaural cues, they remain (often coarse) estimations of the underlying signals. Consequently, the introduction of faulty gains is inherent in the procedure and has to be counteracted.
In particular, fluctuating noise and passages of low SNR lead to non-stationary artifacts that show a duration and a spectral width of mostly just one bin for musical noise, and may sustain over a couple of bins, as frequently observed with babble noise interference (Breithaupt and Martin, 2008).
A method for the attenuation of these artifacts without affecting the quality of the filter in terms of signal distortion and noise suppression is a difficult problem. There exist a handful of remedies, among which the noise flooring and an overestimation of the noise power are the classical solutions (both methods are introduced in Chapter 2.3.1). Another general method to circumvent the problem is to smooth the mask along time and frequency. However, all these techniques have their shortcomings. Noise-flooring and the overestimation of noise power lead to lower interference suppression and signal distortion. Smoothing a mask in time and frequency deteriorates speech intelligibility, as the fine-structure of onsets and in consonants becomes blurred by such means.

As a consequence, the suppression of musical noise artifacts is closely related with the tradeoff between signal suppression and target distortion. A method to lessen this interdependence was introduced by Breithaupt and Martin (2008) and Madhu et al. (2008), who applied the cepstral shaping technique to the problem.

The basis for this solution is the observation that musical noise is fluctuating randomly, with high frequencies in the spectrum, in contrast with the spectral speech features of the vocal tract filter and the vocal cords. Thereby, the latter shows a rather high spectral frequency too, i.e. the fundamental frequency, but no randomness. Because of this difference, these signal components become separable in the cepstral domain (Oppenheim and Schafer, 1975). With the definition of dif-

[8]The techniques related to the cepstrum are distinguished from spectrum methods by a set of established anagrams, e.g. frequency becomes quefrency.

ferent quefrency-regions in the cepstrum, which correspond to different features of speech and noise, a time-based smoothing can be applied that preserves the speech-components while suppressing musical noise. The method is explained below.

A typical mask that realizes a compromise between interference suppression and target distortion can be expressed as:

$$\mathcal{M}(d,n) = \begin{cases} 1 & \text{if} \quad \phi_{\mathrm{ss}}(d,n) > \phi_{\mathrm{vv}}(d,n) \\ A_{\min} & \text{otherwise} \end{cases}, \tag{2.3.19}$$

in which $A_{\min}$ determines the maximum suppression that is permitted. To modify independently the component signals of the mask $\mathcal{M}(d,n)$ in the cepstral domain, the multiplicative mixture in the spectral domain[9] is first linearized with the application of the logarithm function and subsequently transformed to the cepstrum by an inverse DFT:

$$\mathcal{M}_{\mathrm{c}}(g,n) = \frac{1}{N_{\mathrm{d}}} \sum_{d=0}^{N_{\mathrm{d}}-1} \left\{ \log_{\mathrm{n}} \mathcal{M}(d,n) \right\} e^{j2\pi d \frac{g}{N_{\mathrm{d}}}}, \tag{2.3.20}$$

where $g = 0, 1, \ldots, N_{\mathrm{d}} - 1$ denotes the quefrency-bin. At each quefrency coefficient of the mask representation in the cepstrum, a first order recursive averaging of the time index is applied with:

$$\mathcal{M}_{\mathrm{c}}(g,n) = \alpha_{\mathrm{x}} \mathcal{M}_{\mathrm{c}}(g,n-1) + (1-\alpha_{\mathrm{x}}) \mathcal{M}_{\mathrm{c}}(g,n). \tag{2.3.21}$$

In here, four ranges are defined that cover the elemental parts of speech and musical noise in the cepstral domain. Each cepstral range features a particular time constant α_{x} that accounts for the different component signals:

$$\alpha_{\mathrm{x}} = \begin{cases} \alpha_{\mathrm{loE}} & \text{if} \quad g \in \{0, \ldots, g_{\mathrm{loE}}\} \\ \alpha_{\mathrm{hiE}} & \text{if} \quad g \in \{g_{\mathrm{loE}}+1, \ldots, g_{\mathrm{hiE}}\} \\ \alpha_{\mathrm{p}} & \text{if} \quad g \in \{g_{\mathrm{p}}\} \\ \alpha_{\mathrm{n}} & \text{if} \quad g \in \{g_{\mathrm{hiE}}+1, \ldots, N_{\mathrm{d}}/2\} \setminus \{g_{\mathrm{p}}\} \end{cases}. \tag{2.3.22}$$

The lowest range contains the slowly varying spectral broadband envelope of speech, i.e. the formants with maxima at resonances of the vocal filter. Any smoothing decreases the SNR (the target speech is damped) and slurs the phonemes. Therefore the preservation of these speech characteristics is crucial to speech intelligibility and a smoothing constant α_{loE} close to zero should be applied. The second range comprises the fluctuating envelope components in the speech spectrum, i.e. the voiced fine-structure of the speech spectra, that is largely dominated by dynamic articulators in speech sounds. Moreover, this quefrency range is a stage of increasingly

[9] For a multiplicative mixture of the target speech and the distortion in the frequency domain, the signal model of this work in Figure 2.1 needs to be extended by convolutional distortion, which is given through the room impulse response by which the target speech is convolved. This reality has, however, been omitted in this work for a better readability.

occurring musical noise artifacts. Consequently, the smoothing constant α_{hiE} has to constitute a compromise between preservation of detailed spectral speech features and a suppression of musical noise. The third broad range carries, with high probability, the random unwanted peaks that cause the musical noise phenomenon. A smoothing constant α_{n} close to one is applied in this cepstral region, to reduce the variance of the gain function in low SNR sections. The pitch of speech also resides in this range of higher quefrencies. Therefore it has to be excluded from the high smoothing. Fortunately, the cepstrum offers a robust way of estimating the pitch, by taking the maximum value in the cepstral range of the first harmonic, i.e. $g_{\mathrm{p}} \in \{70\ \mathrm{Hz} \ldots 500\ \mathrm{Hz}\}$. With the relation $g_{\mathrm{p}} = f_{\mathrm{s}}/\mathrm{F0}$, the pitch quefrency is calculated with:

$$g_{\mathrm{p}} = \underset{g}{\operatorname{argmax}} \left\{ \mathcal{M}_{\mathrm{c}}(g,n) | g_{\mathrm{p-low}} \leq g \leq g_{\mathrm{p-high}} \right\}. \tag{2.3.23}$$

Since the pitch is mainly present in voiced speech and cannot always be perfectly determined with Equation (2.3.23), a small time constant α_{p} is usually applied that realizes a compromise between pitch-preservation and musical noise suppression.
After the cepstral smoothing, the signal is transformed back to the spectral domain by calculating the DFT and by element-wise exponentiation:

$$\mathcal{M}(d,n) = \exp\left\{ \sum_{g=0}^{N_{\mathrm{d}}-1} \mathcal{M}_{\mathrm{c}}(g,n) e^{-j2\pi g \frac{d}{N_{\mathrm{d}}}} \right\}, \tag{2.3.24}$$

with $d = 0, 1, \ldots, N_{\mathrm{d}} - 1$. Figure 2.7 illustrates a female utterance, the vowels of /a/, /e/ and /i/, in the spectral and cepstral domain (for algorithmic details of the DFT approach see below). An inspection at equal time-instances of the spectrum and

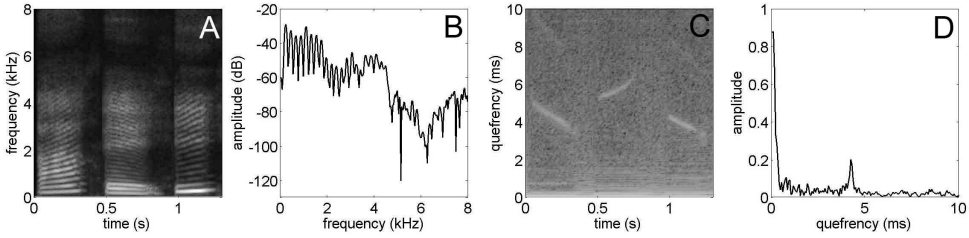

Figure 2.7: *A female utterance of the vowels /a/, /e/ and /i/ in a spectrogram representation (plot* A*). Plot* B *shows an amplitude spectrum thereof at the time instant of* 0.1 *s. Plot* C *gives the referring cepstrogram and in plot* D *the amplitude of the cepstrum at the time instant of* 0.1 *s.*

the cepstrum shows clearly how the component signals are resolved in the cepstral

domain after taking the logarithm of the spectrum and transformation to the cepstrum. While the broadband envelope resides at the low quefrency end, the higher quefrency bins feature the highly fluctuating spectral components of speech as well as the pitch, which additionally shows to be well resolved. Furthermore, one can observe that the main features of these vowels are sustained for a relative long time span in the cepstrum, whereas high quefrency bins mainly show a noise-like pattern, due to turbulence in the vocal tract (with the exception of the quasi-periodic excitation of the vocal folds in the event of voiced speech). Consequently, smoothing in that region should not have a vital impact on speech intelligibility.

In a second example, the influence of the four different smoothing constants on the spectrum is analyzed. Therefore, again, the threefold female vowel utterance is mixed at an SNR of 0 dB with the long-term spectrum of average male speech of Dutch sentences taken from the TNO corpus (TNO, 2000). The signals were lowpass filtered at 8 kHz and sampled at 16 kHz. 256 samples were segmented and Hanning-windowed prior to a 512-point DFT. The window-overlap was 50 %. In order to create a highly non-linear mask, an IBM was established with a fixed $\varepsilon = 0$ dB (see Equation (2.3.15)). The local SNR was calculated from the true power spectral densities according to Equation (2.3.16) and (2.3.17), with an α of 8 ms.
The cepstrum was partitioned in the following sections: the low envelope cepstrum ranging from the 0th to the 5th quefrency-bin (g_{loE}) and the high envelope cepstrum from ranging the 6th to 15th quefrency bin (g_{hiE}). The remainder comprised the fine-structure.

The upper left-hand plot in Figure 2.8 shows the IBM. The dominating formant regions stand out with a distinct pitch pattern that stretches throughout the entire frequency range. Random peaks that produce musical noise are scattered sparsely across the plot. If the high quefrencies are strongly smoothed ($\alpha_{\mathrm{n}} = 0.99$, upper middle plot in Figure 2.8), the random peaks disappear. However, also the fine-structure of the signal is affected, indicating that the filter-source model of speech production is not entirely separable in the cepstral domain. If only the pitch is smoothed ($\alpha_{\mathrm{p}} = 0.99$, upper right-hand plot in Fig. 2.8), a contrast enhancement of the pitch can be observed throughout the frequency range, especially at changes of the pitch. When smoothing the low quefrency bins, the low frequency envelope is affected and strongly smoothed ($\alpha_{\mathrm{loE}} = 0.99$, lower left-hand plot in Figure 2.8). Regions that show a fine-structure, like onsets and spectral details, are less modified. As can be seen in the lower middle plot of Figure 2.8, these regions are influenced when higher quefrencies are smoothed ($\alpha_{\mathrm{hiE}} = 0.99$). Finally an example is given with a practically possible combination of filter coefficients (lower right-hand plot in Figure 2.8) in the cepstral domain, which results in a signal with considerably attenuated musical noise artifacts and a perceptually decreased target distortion.

As the cepstrum is calculated from the real-valued mask, a symmetrical spectral function, the cepstrum itself is also real-valued and symmetrical. This simplifies the processing and allows for exploitation of the symmetry property of the cep-

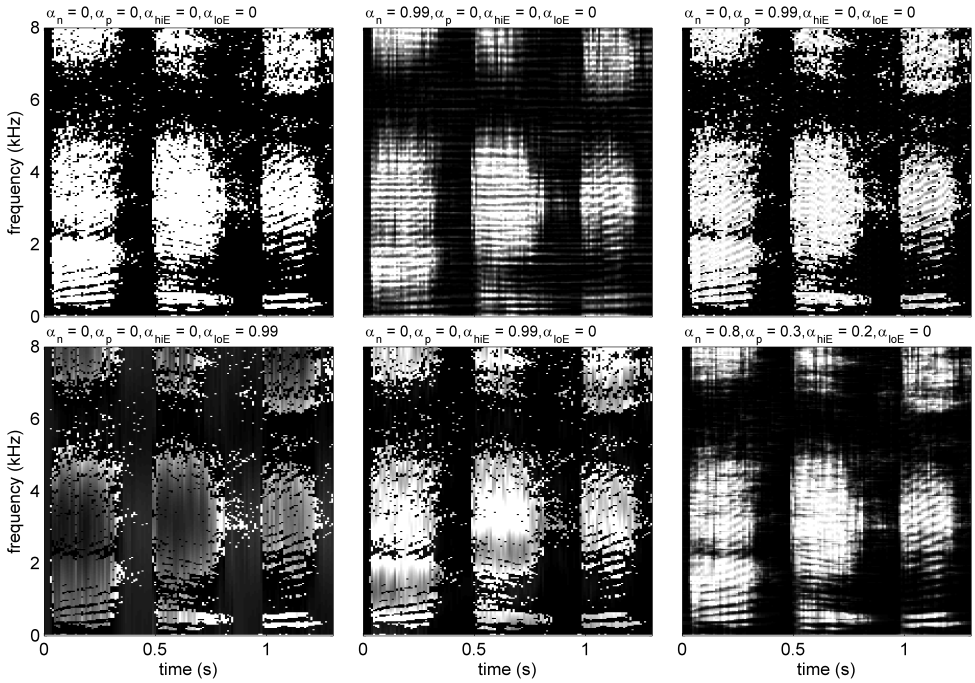

Figure 2.8: *An IBM (using a fixed mask criterion of $\varepsilon = 0$ dB) of a female utterance of the vowels /a/, /e/ and /i/ in a spectrogram, shown in the upper left-hand plot. The speech was mixed at an SNR of 0 dB with time-invariant long-term speech shaped noise of a male voice. The remaining plots show the influence of cepstral smoothing at different quefrency regions in the STFT domain (see titles).*

stral coefficients among the negative and the positive quefrencies, i.e. if $g > N_\mathrm{d}/2$: $\mathcal{M}_\mathrm{c}(g, n) = \mathcal{M}_\mathrm{c}(N_\mathrm{d} - g, n)$. Care has to be taken with respect to the time constants used in the estimation of the power spectral densities, as e.g. done in the Equations (2.3.16) and (2.3.17). If the time constants are too high, the cepstral smoothing operator becomes ineffective (Breithaupt and Martin, 2008). Although the cepstral smoothing technique was applied with IBMs in the examples of this section, it is equally applicable to linear and nonlinear soft-masks.
In Chapter 5.3, the cepstral smoothing constants will be optimized in terms of speech intelligibility for a set of direct DFT-IDFT implementations. The aim is to apply the cepstral smoothing technique at the output of the combined speech enhancement scheme, thereby increasing the output quality without a reduction of speech intelligibility.

2.4 Binaural CASA speech processors

This section introduces three speech enhancement processors that computationally replicate different functional models of the binaural auditory scene analysis (i.e. CASA). This section describes the basic algorithmic procedures and serves as a preparation for the following chapters, in which the CASA processors are applied downstream of several beamforming front-ends, analyzed in a statistical study, optimized and eventually assessed throughout a suite of noise conditions.

To date, many of the functions, as opposed to the underlying functionalities, of the auditory system are understood (Adamy et al., 2003). When examining the auditory system as a whole, a first functional classification may separate the lower neural stages in the brainstem and the midbrain, from the higher stages of the auditory cortex in the outer layer of the cerebrum. The principal function of the lower neural stages is the establishment of an inner representation of the auditory scene via an analysis of auditory cues. Four neural nuclei perform this task with neurons that each respond to specific characteristics of the acoustic stimulus. Thereby the nuclei contribute to a continuously updated multidimensional feature space.

With respect to the capability of the auditory system to tune into a single source amid many sources, it is widely accepted that conscious choice combines bottom-up grouping cues and top-down hypotheses by means of a temporal correlation of response patterns along the auditory path (Brown and Wang, 2006). This mode of operation is referred to as the binding process, an active grouping process, which is accomplished via downward nerves from the auditory cortex. Hence, an alternating manner of listening—signal-based and hypothesis-driven—forms the auditory system. It is the fastest sense and it draws its excellence from its evolutionary importance as an alarm system to most of the vertebrates.

Although much of the auditory functions can be modeled based in offline simulations, for hearing aids only the CASA-based lower neural stages, i.e. the signal-based processes of the auditory scene analysis, just start to become manageable in real-time applications with today's technological possibilities. Despite the computational complexity to model speech patterns as top-down processes, as it is for instance performed in automatic speech recognition, those phoneme or sub-phoneme patterns are difficult (if not impossible) to establish in real-time applications. The reason for this shortcoming is for the most part imposed by the coarticulation phenomenon, which describes that an instantaneous representation of speech on the sub-phoneme level is influenced by the preceding and the following phoneme (Moore, 2003). That is, pattern recognition needs to be retrieved from longer time spans and can possibly not be accomplished in low-delay applications. Hence, in addition to the fact that target patterns are deteriorated and altered by the superposition with interfering sounds, a speech recognition top-down solution turns out to be insolvable for real-time based sound separation approaches.
In consideration of these restrictions and challenges, it is unlikely that CASA-based

speech processors might soon achieve human performance. Moreover, the objective of CASA-based speech enhancement processors is the reconstruction of the waveform of the target speaker. The approach of the model hearing process, however, is to transform the acoustic input into neural response patterns and to bind these into streams. As Ellis (2006) has pointed out, this represents a fundamental difference and obviously offers a much better noise-handling. Despite the difficulty of the speech enhancement approach, a set of binaural CASA processors has been developed that proves to solve the speech-in-noise problem (Wang and Brown, 2006).

As already described in the introduction, this cannot be regarded as a matter of course. Compared to single-channel approaches, which generally fail to generate a speech intelligibility benefit (Hu and Loizou, 2007), also binaural CASA-based speech processors turned out to enhance speech intelligibility only under specific conditions.

When reviewing variants of binaural CASA speech processors, three basic designs can be found (Peissig, 1992; Wittkop et al., 1997). One frequently applied system originates from the binaural algorithm of Gaik and Lindemann (1986), in the following referred to as the carrier-level-phase (CLP) algorithm. This speech processor accomplishes a bilateral frequency decomposition and subsequently calculates the interaural phase and level differences (IPD and ILD, respectively) of the acoustic waveform, to employ these parameters as a directional classifier in an amplitude weighted separation process. The CLP algorithm can be considered an implementation of the coincidence model of Jeffress (1948), which first explained the binaural processing observed in subjective tests.
A second group of binaural CASA algorithms adopts the concept of the multi-channel spatial coherence algorithm of Allen et al. (1977), hereafter referred to as the carrier-coherence (CC) algorithm. Based on primitive grouping, this algorithm exploits the binaural waveform coherence at zero lag, to suppress diffuse sound.
A third well-known binaural CASA algorithm filters the signal in a joint centre and modulation frequency domain and was developed by Kollmeier and Koch (1994); in the following referred to as the envelope-level-time (ELT) algorithm. Herein the separation process is based on the level and time differences of the binaural envelope signal in the range of the fundamental frequency of speech. As the envelope of the signal was considered to be more robust towards noise than the acoustic waveform, this algorithm triggered much hope for an efficient speech enhancement in highly adverse conditions, at the time of its development.
All these algorithms offer a binaural output, which is known to add to the audiological benefit based on a cue-supported hearing.

Although based on the principles of the before mentioned basic binaural processors, there are many variants of these algorithms. Kollmeier et al. (1993) combined the coherence and the coincidence method in one processor. Their approach was revised and extended by Wittkop and Hohmann (2003) to improve speech quality. Albani et al. (1996) combined the principles of algorithm CLP with a lookup table of bin-

aural cues for incidence directions across the upper hemisphere. This lookup table is based on the principles of neural response patterns of binaural cues, originally found in the barn owl (Brainard et al., 1992). The algorithm incorporates the natural cone-of-confusion artifact,[10] a way to overcome it by an across-frequency interaction, and a mechanism for the facilitation to dominant percepts.
What most of these algorithms did not account for was the altered nature of binaural cues in the presence of noise. Rather, many algorithms use a reference lookup table of binaural cues, recorded for a set of directions in anechoic conditions. An exception is the algorithm of Gaik and Lindemann (1986), which uses a lookup table of binaural cues that has been recorded in the presence of noise. A further advancement was presented with the cocktail-party processor[11] of Bodden (1993), using the same core algorithm, i.e. algorithm CLP. The algorithm incorporates the adaptation to HRTFs, contralateral inhibition as well as the precedence effect model, to estimate a Wiener filter.
Roman et al. (2003) developed a binaural cocktail-party processor that uses a maximum a posterior (MAP) classifier for estimating a binary mask. Harding et al. (2005) further developed this method by employing a posteriori statistics in the classification process for estimating a soft-mask. Their approach, also based on algorithm CLP, was deemed a front-end in automatic speech recognition. Nix (2005) and Nix and Hohmann (2006) applied a MAP classification approach to CASA-based localization, using the framework of algorithm CLP. In addition, Nix (2005) proposed a statistical non-Gaussian multidimensional source separation method which draws on the localization of sources and the dynamical spectro-temporal evolution of speech. Compared to previous approaches that simulate primitive grouping, these algorithms introduced the principles of schema-driven source segregation to CASA, by employing the knowledge of patterns.[12] Therefore, the incorporation of patterns of binaural cues can be understood as an attempt to mimic the binaural top-down processing. As recently proposed by Blauert (2011) and Kolossa (2011), if the ultimate goal of CASA is to achieve human performance, future CASA algorithms need to include more sophisticated models of top-down processes.
Multi-layered bottom-up CASA processors that combine binaural and monaural cues were proposed by Woods et al. (1996), Woodruff et al. (2010) and Weiss et al. (2011). Recently, as an alternative to the common Jeffress-model, Li et al. (2011) proposed a speech enhancement model based on the equalization-cancellation model of Durlach (1960). There are also algorithms that explicitly, as opposed to the implicit approach of algorithm CC, simulate the precedence effect,[13] e.g as proposed by Martin (1997).

[10] The cone of common IPDs and ILDs around the cranical axis is known as the cone-of-confusion artifact. The ambiguity resolves for broadband sounds.

[11] Often binaural CASA speech processors are referred to as 'cocktail-party processors'. A term that was initially established by Cherry (1953) to describe the human ability of listening to a speaker in noisy surroundings.

[12] The ASA processes of 'primitive grouping' and 'schema-driven segregation' are studied in (Bregman, 1990, p. 395 ff.)

[13] The precedence effect describes the psycho-acoustic phenomenon in which sounds gain a dominant perceptual localization cue if these are associated with the first wavefront (Blauert, 1997).

These approaches showed to be mainly capable of the localization of transients (Opdam, 2010), rather than generating a speech intelligibility enhancement.

In this work we present a conceptual study of the three principal binaural speech processors, i.e. he algorithms CC, CLP and ELT, with and without a beamforming front-end. A similar approach can be found in Wittkop et al. (1997), who compared the algorithm of Kollmeier et al. (1993), which is a combination of algorithm CC and CLP, with the algorithm of Kollmeier and Koch (1994), hence algorithm ELT in the present work, in a listening test. Despite the important result that confirmed an audiological success for normal hearing and hearing impaired people under specific conditions, Wittkop's review and analysis of binaural speech processors did not include a beamforming front-end and presented mainly a hands-on approach, with few algorithmic details, for a small set of speech-in-noise conditions. Moreover, the combined application of algorithm CC and CLP in Wittkop's study was not supported by a thorough analysis of the benefit given by each of the underlying processing schemes. In addition, new algorithmic insights, as well as new statistical approaches that had arisen in recent years to classify binaural cues in noise, suggest that binaural CASA algorithms need to be revised.
In the following, the algorithmic frameworks of the three binaural speech processors are introduced. Figure 2.9 juxtaposes the schemes. During the study of this work, the initial conceptual designs of Gaik and Lindemann (1986); Allen et al. (1977) and Kollmeier and Koch (1994) were looked at. With respect to the implementation, the sampling frequency of each algorithm was 16 kHz. Algorithm ELT, using a sampling of 16 kHz in the time domain, deviates in terms of frequency transfer because it offers an internal bandpass sampling frequency of only 8 kHz, which implies that the temporal Nyquist frequency was limited to 4 kHz. However, considering speech intelligibility, a frequency limitation of this degree does not result in a distinct disadvantage with respect to the other algorithms, which is a requirement for a comparative study. To account for the non-stationarity and non-whiteness of speech signals, each algorithm uses a direct DFT-IDFT analysis-synthesis approach with Hanning-weighted frames of 16 ms length and an overlap of 50 %.[14] Throughout this work, the target source location is limited to the frontal direction.
This chapter exclusively presents the basic working principles of each algorithm. It does not introduce the pattern-based classification, nor the optimization of the algorithmic parameters, which will be presented in Chapters 3.3.3 and 5.2, respectively. Besides, it should be emphasized that algorithm CC is based on the standard primitive grouping scheme, using the non-directional magnitude squared coherence at zero lag, as a noise classifier. Hence, only algorithm ELT and CLP will be extended by a pattern-based classifier.

[14]See Appendix A.2 for a definition of the DFT-IDFT filterbank approach.

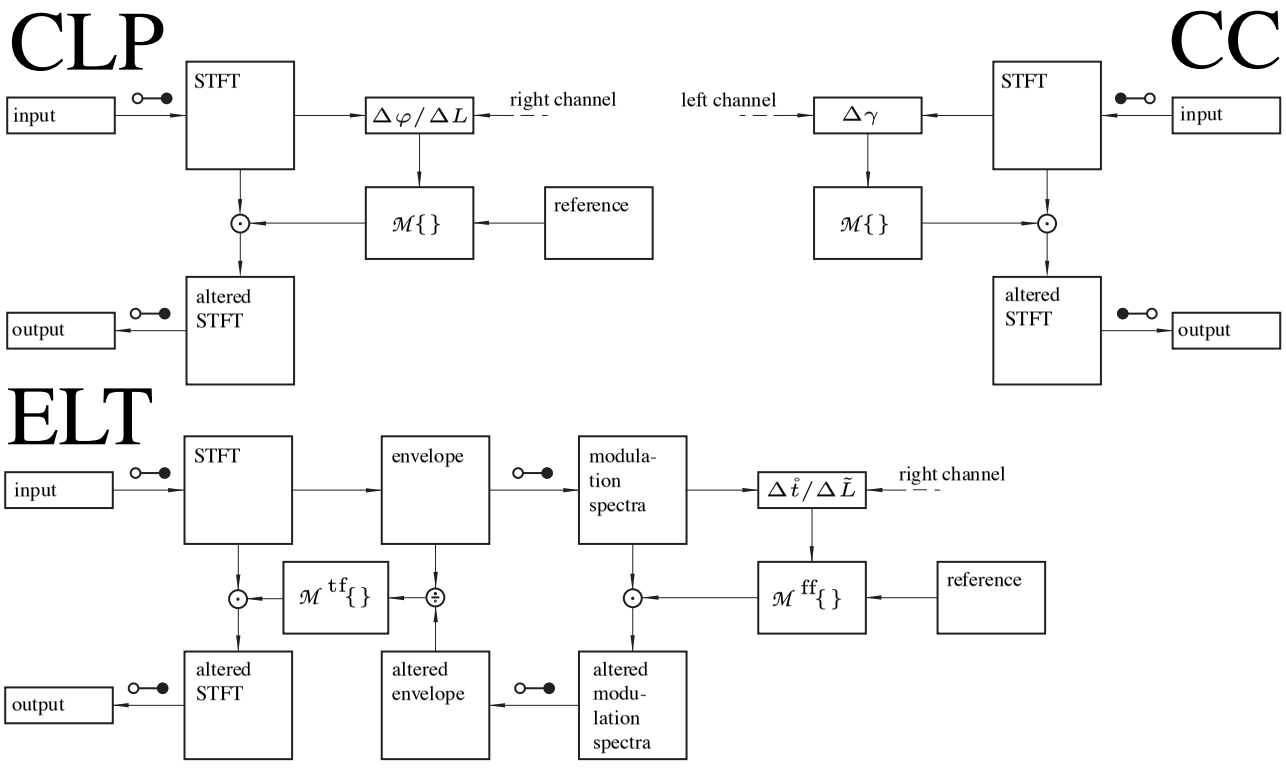

Figure 2.9: *Block diagrams of the model-based binaural speech processors presented in this work. The CLP algorithm refers to the algorithm of Gaik and Lindemann (1986), while algorithm CC is a binaural version of the spatial coherence-based algorithm of Allen* et al. *(1977). Algorithm ELT is based on the algorithm of Kollmeier and Koch (1994). As the binaural algorithms are symmetric around their binaural stages, only one side is drawn. The barbells indicate a frequency transformation, or its inverse.*

2.4.1 Algorithm CC

It is one of the capabilities of the auditory system to suppress reverberation and diffuse noise, presumably through a physiological decorrelation process and the head shadow effect. It was found that reverberation has a severe impact on monaural detection thresholds and a less severe impact on binaural detection thresholds, which results in a significant binaural advantage when listening with two ears (Brown and Palomaki, 2006).
The perception of reverberation on stimuli observed in experiments can be modeled by the normalized spatial correlation function, i.e. the coherence, which extends the interaural correlation process to the phenomenon of the precedence effect (Faller and Merimaa, 2004).
The maximum of the normalized correlation function signifies the coherence between the signals and results in a value between zero and one. Consequently, if the signals are highly coherent, the value of the coherence function is tending towards one. Reverberation and diffuse sound are generally indicated by a low correlation between the left and right ear signal and therefore can be identified with a low binaural waveform coherence value. Thereby, the coherence function shows a main lobe with a maximum that depends on the spatial displacement of the sound source. The width of its main lobe is inversely proportional to the distance between the microphones (or ears) and the diffuseness of the sound field. In order to perform speech enhancement, Allen et al. (1977) selected the latter characteristic of the coherence and applied it as a gain function that distinguishes between coherent and incoherent signal portions, with the aim to improve speech quality. Their approach represents the conceptual basis of the speech processor CC in this work.
Unfortunately, the multi-channel coherence function has a dependency on a set of factors that impede their applicability. These are mainly the distance between the microphones, i.e. a dependency on the wavelength of the sound, and the distance towards the target speaker in relation to the reverberation radius.[15]

For two omni-directional receivers in an ideal diffuse sound field, the absolute magnitude spatial coherence can be expressed as:

$$\Delta\gamma(f,l) = \left|\frac{\sin(kl)}{kl}\right|, \tag{2.4.25}$$

with $k = 2\pi f/c$ being the wave number, c is the speed of sound and l is the distance between the receivers. The function has a first zero at $f = c/2l$. As a result, the coherence function can only be used as a meaningful agent for the differentiation between coherent and incoherent signal portions if the distance between the receivers

[15]The reverberation radius describes the distance around a source, at which direct sound and reverberated sound have equal energy. With the assumption of a perfect diffuse reverberation field, it can be calculated with: $r_{\mathrm{RT}} = 0.1\sqrt{\frac{V}{\pi \mathrm{RT}}}$, where V is the volume of the room and RT the reverberation time.

and the frequency are sufficiently high. In addition and as above-mentioned, coherent signals should be tending towards one in order to be detected as the target signal. Therefore, a target speaker should ideally be located within the reverberation radius. A certain direct-to-reverberation ratio, however, always constitutes a certain amount of decorrelation and, moreover, reverberation acts on the direct speech too.

In order to improve the direct-to-reverberation ratio, Martin (2001) suggests the combination of the coherence-based post-filter with directional microphones. Based on measurements with an artificial head, Jeub et al. (2009) showed that binaural impulse responses expand the coherence function with respect to a free-field measurement due to the head shadow effect. Consequently, the coherence function is applicable at lower frequencies. A performance improvement of algorithm CC by about 1 dB when using binaural impulse responses instead of free-field impulse responses was demonstrated in the attenuation of reverberation.

As the algorithms, which are considered here, are designed to enhance a speaker in the frontal direction, the estimated spectral power densities in both channels as well as the cross power spectral density are directly used to calculate the coherence function at zero lag. If $x_{\mathrm{l}}(d,n)$ denotes the input of the left channel and $x_{\mathrm{r}}(d,n)$ denotes the right channel in the STFT domain, the spectral power density estimates can be calculated with the previously introduced first-order recursive smoothing of the STFT signals:

$$\phi_{\mathrm{ll}}(d,n) = \alpha_\gamma \phi_{\mathrm{ll}}(d,n-1) + (1-\alpha_\gamma)|x_{\mathrm{l}}(d,n)|^2, \tag{2.4.26}$$

$$\phi_{\mathrm{rr}}(d,n) = \alpha_\gamma \phi_{\mathrm{rr}}(d,n-1) + (1-\alpha_\gamma)|x_{\mathrm{r}}(d,n)|^2, \tag{2.4.27}$$

$$\phi_{\mathrm{lr}}(d,n) = \alpha_\gamma \phi_{\mathrm{lr}}(d,n-1) + (1-\alpha_\gamma)x_{\mathrm{l}}(d,n)x_{\mathrm{r}}^*(d,n), \tag{2.4.28}$$

in which $*$ and α_γ are the complex conjugate operator and the integration constant for estimating the density functions, respectively. From these spectral power density estimations, the normalized absolute magnitude coherence at zero lag is calculated as:

$$\Delta\gamma(d,n) = \frac{|\phi_{\mathrm{lr}}(d,n)|}{\sqrt{\phi_{\mathrm{ll}}(d,n)\phi_{\mathrm{rr}}(d,n)}}. \tag{2.4.29}$$

The function $\Delta\gamma(d,n)$ can directly be multiplied with the modulus of the left and right STFT signal, to enhance coherent target speech (preferably) from the front in a diffuse sound field. In algorithm CC the square of the coherence function in Equation (2.4.29) is calculated, $\Delta\gamma^2(d,n)$, which is the well-known magnitude squared coherence (MSC) function.

As the benefit of the algorithm across many different conditions can be improved by an empirical adaption processes, a set of algorithmic parameters is introduced. The parameters of algorithm CC are given in Table 2.1.
Therein the parameters $A_{\min}$, e and d_{x} denote the maximum suppression of the magnitude of the STFT bins, the compression or expansion of the weighting function, i.e. of the MSC function, and a lower cutoff frequency bin for applying the

Table 2.1: *Parameters and parameter ranges of algorithm CC.*

fixed parameters

f_s	N_χ (analysis window size)	N_d (DFT size)	ΔT (overlap)
16 kHz	256 bin (16 ms)	512 bin	128 bin (8 ms)

algorithmic parameters

A_min	e	α_γ	d_x
0.01 - 0.5	0.5 - 3	0.01 - 0.9	0.1 - 7 kHz

coherence function, respectively. Parameter d_x is introduced to limit the influence of a coherence-based weighting at low frequencies, where the coherence function is broadened and, as before mentioned, a weak agent for interference suppression. In Chapter 5.2, these algorithmic parameters are tuned in specific speech-in-noise conditions to attain optimal speech intelligibility.

Using a particular parameter set, the magnitude weighting function of algorithm CC is calculated with:

$$\mathcal{M}_\mathrm{cc}(d,n) = \begin{cases} 1 & \text{if} \quad d < d_\mathrm{x} \\ \max\left[\Delta\gamma^2(d,n), A_\mathrm{min}\right]^e & \text{if} \quad d \in \{d_\mathrm{x}, \ldots, N_\mathrm{d}/2\} \end{cases}, \tag{2.4.30}$$

and multiplied with the magnitude of the STFT representation of the left and right channel signal, as sketched in Fig. 2.9. The filtered signal is transformed back to the time domain with the IDFT and an overlap-add technique. The original STFT phase is left unchanged throughout the filtering process.

The MSC, calculated with the fixed parameters in Table 2.1, is analyzed in Figure 2.10. The upper row of plots shows the influence of the smoothing constant $\alpha_\gamma = \exp\left(-\Delta T/\check{\tau}\right)$ with different values for $\check{\tau}$ for a canteen and a workshop background over a duration of 20 s. See Chapter 3.2 for a description of the sound material. The results are compared to the theoretical MSC in an ideal diffuse noise field, i.e. the square of Equation (2.4.25) with $l = 20$ cm.
The importance for averaging the magnitude squared normalized cross-power density for inferring the coherence can be seen. Due to the practical existence of stochastic signals in the left and the right channel per STFT atom, an MSC based on a short sample mean does not represent the amount of linear relationship between the channels. It rather tends to one because of the approximate instantaneous energy correspondence in the numerator and denominator of Equation (2.4.29). For this reason an MSC close to 1/2 is observed if $\check{\tau}$ is as short as 8 ms, although the sound field is highly diffuse. Only if an averaging over several bins is performed, the coherence converges towards the theoretical quantities of a diffuse noise field. For a practical application of algorithm CC, the averaging of the coherence, however, must not be too long, as this will slur transients in speech when applied in the STFT

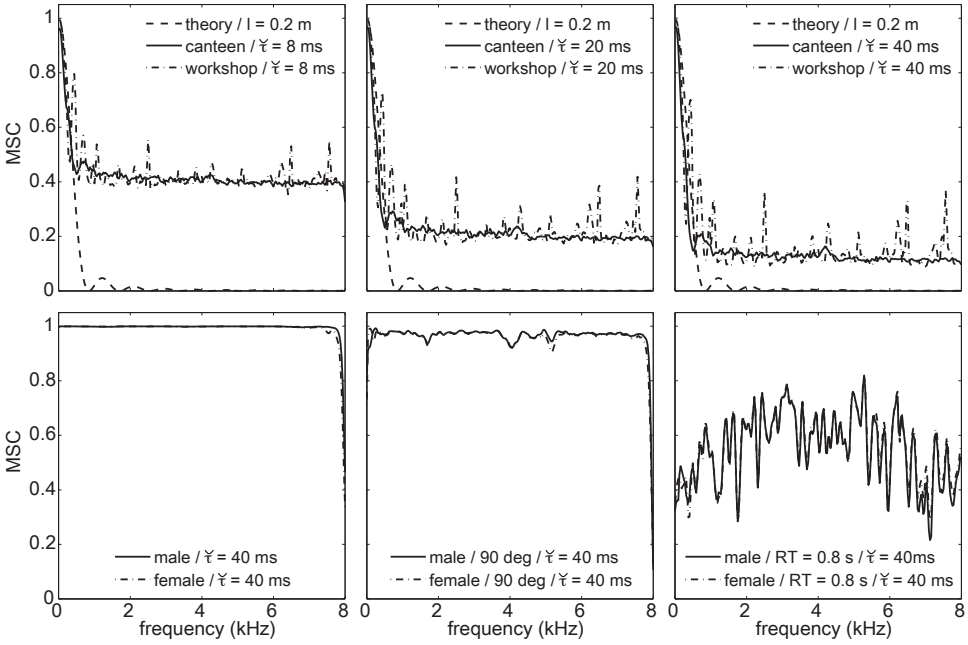

Figure 2.10: *Application examples of algorithm CC are presented using the binaural impulse responses of the Aachen head. The first row of plots gives the MSC over frequency of different spatial scenes at different time constants ($\alpha_\gamma = \exp(-\Delta T/\check{\tau})$) for averaging the magnitude squared normalized cross-power density. The results are compared to the theoretical MSC in an ideal diffuse noise field, i.e. the square of Equation (2.4.25) with $l = 20$ cm. The second row of plots gives the MSC of speakers in frontal direction at an SNR of 60 dB in a diffuse sound field. The middle plot at the bottom shows the MSC when the same speakers are located at 90 deg. The right-hand plot at the bottom shows the MSC of artificially reverberated speech at frontal position with an RT of 0.8 s.*

weighting process.

The left-hand plot at the bottom of Figure 2.10 shows the MSC over frequency for coherent speech of a male and a female speaker for a duration of 20 s. A high value of the coherence is attained throughout the entire spectrum. The middle plot at the bottom of Figure 2.10 shows the inability of algorithm CC to distinguish frontal from lateral sources with an MSC calculated at zero lag. Hence, also for lateral sources at 90 deg the MSC remains close to one. The right-hand plot at the bottom of Figure 2.10 shows the influence of artificial reverberation (RT = 0.8 s, $r_{\mathrm{RT}} = 0.6$ m, distance of the speaker 1 m) acting on the direct speech from frontal direction. The coherence is severely decreased and the contrast between target and diffuse noise is

weakened.

Today the approach of Allen et al. (1977) forms the basis of numerous speech processors. For instance, coherence-based filters are applied in speech enhancement for voice intercom systems (Martin, 2001) and binaural noise suppression systems (Dörbecker and Ernst, 1996; Lotter and Vary, 2006; Rohdenburg, 2008). As an alternative to a joint gain function for the left and right channel, Wittkop and Hohmann (2003) introduced separate coherence-based magnitude gain factors, which are calculated from a binaural input. However, due to the complexity of their algorithm, it offers three combined schemes of speech enhancement as well as a scene analyzer, it is not directly clear whether this approach offers advantages over the here presented method.

2.4.2 Algorithm CLP

Localization and intelligibility are closely related (Stern et al., 2006). This observation was made with concurrently active speakers, generally, such that correct localization implies correct understanding. The auditory system is remarkably sensitive to changes of interaural parameters. For pure tones, humans are able to discriminate ITDs with just noticeable differences (JNDs) of 10 μs, and JNDs of ILDs are at 1 dB. The angular resolution for speech-like sounds is 5 deg in the median plane and reduces to 20 deg in lateral directions. Clicks on the contrary can be localized with a difference of 1 deg in the frontal direction and about 5 deg in lateral directions (Stern et al., 2006).

As previously mentioned, the coincidence model of Jeffress (1948) accounts for many of the phenomena that are associated with binaural processing. The model spans a two-dimensional space, with one axis formed by the centre frequencies, the tonotopic axis, and the neural coincidence detectors of spatial displacement, forming the other axis.
In the following, the model is implemented with a cross correlation function per centre frequency bin. Therefore, subsequent to the estimation of the cross power spectral density representation of the signals, as in Equation (2.4.28),[16] interaural disparity based on IPDs can be calculated in the standard way:

$$\Delta\varphi(d,n) = \arg(\phi_{\mathrm{lr}}(d,n)). \tag{2.4.31}$$

The ILDs are computed from the left and right auto power spectral density representations of Equation (2.4.26) and (2.4.27):[17]

$$\Delta L(d,n) = 10\log_{10}\left(\frac{\phi_{\mathrm{ll}}(d,n)}{\phi_{\mathrm{rr}}(d,n)}\right). \tag{2.4.32}$$

[16] Note Equation (2.4.28) changes for algorithm CLP, such that α_γ is replaced with α_{PSD}.

[17] Note Equations (2.4.26) and (2.4.27) change for algorithm CLP, such that α_γ is replaced with α_{PSD}.

The calculation of the final weighting function $\mathcal{M}_{\mathrm{clp}}(d,n)$ follows a series of steps. Again, several algorithmic parameters are introduced to adapt the processing of the binaural algorithm to different scenes. First, a raw weighting function $\mathcal{M}'_{\mathrm{clp}}(d,n)$ is computed for the suppression of time-frequency bins that show IPDs and ILDs which deviate from the frontal direction. In addition, if one of the binaural parameters fails as a classifier in a certain centre frequencies range, the classification with that cue is restricted by an adjustable weighting function to its efficient working range. Therefore, $\mathcal{M}'_{\mathrm{clp}}(d,n)$ is based on the following definition:

$$\mathcal{M}'_{\mathrm{clp}}(d,n) = \begin{cases} \mathcal{M}_{\Delta\varphi}(d,n) & \text{if} \quad d \in \{0,\ldots,\min[d_{\mathrm{x}\varphi}, d_{\mathrm{xL}}]-1\} \\ (1-\xi)\mathcal{M}_{\Delta L}(d,n)\ldots & \\ +\ \xi\mathcal{M}_{\Delta\varphi}(d,n) & \text{if} \quad d_{\mathrm{xL}} \leq d_{\mathrm{x}\varphi} \cap d \in \{d_{\mathrm{xL}},\ldots,d_{\mathrm{x}\varphi}\} \\ 1 & \text{if} \quad d_{\mathrm{xL}} > d_{\mathrm{x}\varphi} \cap d \in \{d_{\mathrm{x}\varphi},\ldots,d_{\mathrm{xL}}\} \\ \mathcal{M}_{\Delta\mathrm{L}}(d,n) & \text{if} \quad d \in \{\max[d_{\mathrm{x}\varphi}, d_{\mathrm{xL}}]+1,\ldots,N_{\mathrm{d}}/2\}. \end{cases} \tag{2.4.33}$$

Therein $\mathcal{M}_{\Delta L}$ and $\mathcal{M}_{\Delta\varphi}$ represent soft-masks of the ILD cue and the IPD cue, respectively. These soft-masks are based on the a posteriori probability p that the signal at a certain time-frequency atom in the STFT domain is generated by the target. For instance, the soft-mask $\mathcal{M}_{\Delta L}$ is computed with:

$$\mathcal{M}_{\Delta L}(d,n) = p(\theta_{\mathrm{t}}|\Delta L(d,n)), \tag{2.4.34}$$

in which θ_{t} is the direction of the target signal. The soft-mask $\mathcal{M}_{\Delta\varphi}$ is calculated analogously. The computation of the a posteriori lookup tables is presented in Chapter 3.3.3.

As it has been formalized in Equation (2.4.33), prior to the composition of the raw weighting function, the soft-masks $\mathcal{M}_{\Delta L}$ and $\mathcal{M}_{\Delta\varphi}$ are restricted to specific frequency ranges. The crossover frequencies d_{xL} and $d_{\mathrm{x}\varphi}$ are tuning parameters of the algorithm and allow for full overlap, a partition of the IPD at low frequencies and ILD at high frequencies, and no overlap.[18] Furthermore, in case when $d_{\mathrm{xL}} \leq d \leq d_{\mathrm{x}\varphi}$, the soft-masks are balanced with the parameter ξ. Using this variable assemblage of the soft-masks, the CLP algorithm can be optimized with respect to the efficient domain of each binaural cue and their weighted combination.

The weighting function $\mathcal{M}_{\mathrm{clp}}(d,n)$ in Figure 2.9 is then calculated as:

$$\mathcal{M}_{\mathrm{clp}}(d,n) = \max\left[(\mathcal{M}'_{\mathrm{clp}}(d,n))^e, A_{\min}\right], \tag{2.4.35}$$

whereby the raw weighting function $\mathcal{M}'_{\mathrm{clp}}(d,n)$ is compressed or expanded with the exponent e and lower bounded with $A_{\min}$. Subsequently, $\mathcal{M}_{\mathrm{clp}}(d,n)$ is multiplied with the modulus of the original STFT signal in both channels. The original STFT phase is left unchanged throughout the filtering process. In a last step, the filtered

[18] At around 1.5 kHz the ILD and IPD of the fine-structure of the waveform are unstable, often equivocal, cues. As a consequence a unity gain weighting in this frequency range might be advantageous for attaining a speech intelligibility gain.

Table 2.2: *Parameters and parameter ranges of algorithm CLP. ϵ_{hist} determines the main-lobe of the algorithm. A description of this parameter is given in Chapter 3.3.3.*

fixed parameters

f_{s}	N_{χ}	N_{d}	ΔT	α_{PSD}
16 kHz	256 bin (16 ms)	512 bins	128 bins (8 ms)	0.36 ($\breve{\tau}$ = 8 ms)

algorithmic parameters

d_{xL}	$d_{\text{x}\varphi}$	$[n_{\text{L}}\ n_{\text{L}}]$	$[n_{\varphi}\ n_{\varphi}]$	ϵ_{hist}	ξ	A_{min}	e
0 - $N_{\text{d}}/2$	0 - $N_{\text{d}}/2$	1 - 10	1 - 10	1 - 5	0 - 1	0.01 - 0.5	0.5 - 2

signals are transformed back to the time domain with the IDFT and a following overlap-add technique.

The parameters of the algorithm CLP are outlined in Table 2.2. In order to improve the performance further, three algorithmic parameters are introduced to modify the raw weighting function. These are the bin sizes of sub matrixes (n_{L} by n_{L} for the ILD values and n_{φ} by n_{φ} for the IPD) that are used to cluster adjacent bins in the binaural domains with a smoothing process. This approach has been suggested in Peissig (1992). A third parameter is introduced, ϵ_{hist}, to control the width of the main lobe, in analogy to beamformers, of the binaural spatial filter. A description of this parameter, which controls the a posteriori lookup table as well as the subsequent optimization of parameters, is given in Chapter 3.3.3.

Algorithm CLP can be regarded as the classical binaural processor of time-frequency masking. The speech processor has shown to work efficiently if the interferers are disjoint in the time-frequency representation (Peissig, 1992). The performance, however, declines in diffuse noise fields, in the presence of reverberation and towards negative SNRs. For that reason, algorithm CLP does generally not have the capability to unravel a target speaker from interference in diffuse sound fields. Nevertheless, the algorithm can be considered a suitable candidate for speech enhancement in many situations, in particular as it allows for a pattern-driven classification and, hence, a high degree of plasticity towards many different conditions.

2.4.3 Algorithm ELT

In explaining the auditory deficiency of a mere binaural temporal disparity-based discrimination of continuous sounds, Stern et al. (2006) refer to psycho-acoustical tests, which show that identification according to interaural time difference (or equivalently phase differences) can be easily achieved by modulating the stimuli with modulation frequencies typically found in speech, i.e. from the articulation rate to pitch frequencies. For this reason, it is assumed that concurrent speakers are grouped with respect to the combination of modulation frequency and interaural temporal

disparity.
For one part, this finding corresponds to the psycho-acoustical phenomenon of the co-modulation masking release, which describes that signal detection can be improved in masking experiments, if the envelope of the masker is modulated and this modulation is coherent or correlated across different frequency bands (Moore, 2003). The observation of Stern et al. (2006) is also supported by physiological findings, which provide evidence of neurons that are tuned to modulation frequencies, and which are supposedly organized perpendicularly to the mapping (hence, independent) of centre frequencies, i.e. the tonotopic neural coding (Kollmeier and Koch, 1994).

Several monaural speech processors have been developed that mimic the auditory modulation-based grouping of speech mixtures (Hu and Wang, 2004; Schimmel et al., 2007). Based on a different concept, but still, predominantly based on modulation perception, Mesgarani et al. (2004) developed a noise suppression algorithm which simulates spectro-temporal response fields (STRFs) of the auditory cortex. The algorithm performs a filtering at spectro-temporal modulations of less than 32 Hz, which allows a flexible auditory pattern analysis in speech enhancement. In subjective evaluation the algorithm demonstrated an improvement of signal quality over a noise estimation-driven Wiener filter approach.

Based on physiological and psycho-acoustical models of the lower neural stages of the auditory apparatus, Kollmeier and Koch (1994) combined the two-dimensional representation of centre and modulation frequencies, those in the range of pitch frequencies, i.e. up to 400 Hz, with a model of binaural interaction. The algorithm performs interference suppression based on binaural cues of the envelope at different modulation frequencies; it was introduced as algorithm ELT in this work and is schematically depicted in Figure 2.9. As can be seen, the main difference of this algorithm to the algorithms CC and CLP is a location-based separation using the binaural envelope signal in the centre and modulation frequency domain. Two weighting functions, $\mathcal{M}_{\text{elt}}^{\text{ff}}$ and $\mathcal{M}_{\text{elt}}^{\text{ft}}$, are subsequently calculated, in which the latter is to be used as a magnitude-based weighting in the two channel STFT signal.

As such an integral expression converges slowly (Hartmann, 1997), The algorithm starts with a decomposition of the signals into complex bandpass representations by a DFT (length N_{d} in Table 2.3) over Hanning-weighted analysis windows (length N_{χ} in Table 2.3). The analysis windows overlap by 0.125 ms, which results in a sampling frequency of the complex bandpass signal of 8 kHz. To avoid aliasing in the newly time-sampled bandpass representation, the signal is only processed up to the Nyquist-frequency of 4 kHz. Additionally, a low-pass filter with a fourth order FIR filter is applied to the complex band-pass time series with a cutoff frequency of 3.6 kHz, in order to limit the frequency of the envelope estimates. No delay is introduced with this filter by using the MATLAB (The MathWorks TM) function `filtfilt.m`.

Subsequently, the envelope of the bandpass signal, e.g. for the left side $\mathcal{E}\{x_{\text{l}}(d,n)\}$,

is extracted using the Fourier transform. In the present implementation of algorithm ELT, a window-wise 512-point DFT/IDFT analysis/synthesis approach across n is calculated using a Hanning window of 128 bins, by which the bandpass signal is weighted, and an appended array of 384 bins of zero amplitude, see Appendix A.2. Having the signal transformed with the DFT, the negative frequency components are set to zero. Subsequently, the IDFT is taken and the absolute value is computed. The result is multiplied by a factor of two in order to arrive at an approximation of $\mathcal{E}\{x_\mathrm{l}(d,n)\}$. The method is similar to the envelope extraction with the Hilbert transform from real-valued signals (Hartmann, 1997).

As a means to transform the signal in the centre and modulation frequency domain, the envelope in the STFT domain is first frame-wise weighted (in temporal direction) with a vector $\mathring{\chi}(n)$ that consists of a Hanning window of length $N_{\mathring{\chi}}$ and an appended array of $N_\mathrm{dd} - N_{\mathring{\chi}}$ zeros. Subsequently, the modulation spectrum is computed with a DFT of length N_dd:

$$\mathring{x}_\mathrm{l}(d,m,o) = \sum_{n=0}^{N_\mathrm{dd}-1} \mathring{\chi}(n)\mathcal{E}\{x_\mathrm{l}(d,o\Delta\mathring{T}+n)\}e^{-j2\pi n\frac{m}{N_\mathrm{dd}}}, \tag{2.4.36}$$

where $m = 0, 1, \ldots, N_\mathrm{dd}-1$, o and $\Delta\mathring{T}$ are the modulation frequency coefficient, the frame index and the frame shift, respectively. See Table 2.3 for the details of the present implementation of algorithm ELT. Note that $\Delta\mathring{T}$, which also specifies the order of the system delay of algorithm ELT, is as short as 8 ms. In accordance with the implementation of Kollmeier and Koch (1994), the resulting complex modulation spectrum beyond 400 Hz (corresponding to $m > 27$) is discarded in the following cue-based filtering process. In the inverse transform, the modulation spectrum $m > 27$ is retained unaltered.

The estimated modulation spectral auto and cross power spectral densities (MPSD) are computed using the standard first-order recursive averaging method:

$$\mathring{\phi}_\mathrm{ll}(d,m,o) = \mathring{\alpha}\mathring{\phi}_\mathrm{ll}(d,m,o-1) + (1-\mathring{\alpha})|\mathring{x}_\mathrm{l}(d,m,o)|^2, \tag{2.4.37}$$

$$\mathring{\phi}_\mathrm{rr}(d,m,o) = \mathring{\alpha}\mathring{\phi}_\mathrm{rr}(d,m,o-1) + (1-\mathring{\alpha})|\mathring{x}_\mathrm{r}(d,m,o)|^2, \tag{2.4.38}$$

$$\mathring{\phi}_\mathrm{lr}(d,m,o) = \mathring{\alpha}\mathring{\phi}_\mathrm{lr}(d,m,o-1) + (1-\mathring{\alpha})\mathring{x}_\mathrm{l}(d,m,o)\mathring{x}^*_\mathrm{r}(d,m,o), \tag{2.4.39}$$

and $\mathring{\alpha} = \exp(-\Delta\mathring{T}/\breve{\tau})$, where the time constant has been set to 8 ms.

Based on the MPSD representation in the left and right channel, the interaural phase and level differences can be calculated as:

$$\Delta\mathring{\varphi}(d,m,o) = \arg(\mathring{\phi}_\mathrm{lr}(d,m,o)) \tag{2.4.40}$$

and:

$$\Delta\mathring{L}(d,m,o) = 10\log_{10}\left(\frac{\mathring{\phi}_\mathrm{ll}(d,m,o)}{\mathring{\phi}_\mathrm{rr}(d,m,o)}\right), \tag{2.4.41}$$

respectively. As opposed to the implementation of Kollmeier and Koch (1994) in which the IPD was employed for the noise suppression task, in this work the related ITD:

$$\Delta\mathring{t}(d,m,o) = \frac{\Delta\mathring{\varphi}(d,m,o)}{2\pi m} \tag{2.4.42}$$

is applied.[19] As it is shown in Chapter 3.3.2, ITDs of low modulation frequencies are a better indicator in the source separation process than the corresponding IPDs. Subsequently, the binaural representations of $\Delta\mathring{L}(d,m,o)$ (i.e., $\Delta\tilde{L}(d,m,o)$, see footnote) and $\Delta\mathring{t}(d,m,o)$ are averaged with an adjustable sub matrix of $[n_\mathrm{L}\ n_\mathrm{L}]$ and $[n_\mathrm{t}\ n_\mathrm{t}]$ bins, respectively. Hence, similar to the implementation of algorithm CLP, a simple method of clustering, or in terms of ASA a primitive grouping, due to the proximity along the tonotopic and periodotopic, i.e. modulation frequency, dimensions is allowed.

Hereafter, the DFT coefficients and the frame index of the masks are omitted for notational convenience.

As it was proposed by Kollmeier and Koch (1994), the robustness of the algorithm can be improved by identifying unreliable ITD and ILD values by calculating the standard-deviations σ_t and σ_L, respectively, across a sliding sub matrix of centre and modulation frequencies of five by five bins (i.e., parameters $n_{\sigma\mathrm{t}}$ and $n_{\sigma\mathrm{L}}$ in Table 2.3). The approach is adopted, by which two masks, one for the standard deviation of the ILD values, denoted $\mathcal{M}_{\sigma\mathrm{L}}$ and one of the ITD values, denoted $\mathcal{M}_{\sigma\mathrm{t}}$, are calculated as:

$$\mathcal{M}_{\sigma\mathrm{L}} = \begin{cases} 1 & \text{if} \quad \sigma_\mathrm{L} < 0.6 \text{ dB} \\ (3.5 - \sigma_\mathrm{L})/2.9 & \text{if} \quad 0.6 \text{ dB} \leq \sigma_\mathrm{L} \leq 3.5 \text{ dB} \\ 0 & \text{if} \quad \sigma_\mathrm{L} > 3.5 \text{ dB} \end{cases} \tag{2.4.43}$$

and

$$\mathcal{M}_{\sigma\mathrm{t}} = \begin{cases} 1 & \text{if} \quad \sigma_\mathrm{t} < 0.6 \text{ ms} \\ (3.5 - \sigma_\mathrm{t})/2.9 & \text{if} \quad 0.6 \text{ ms} \leq \sigma_\mathrm{t} \leq 3.5 \text{ ms} \\ 0 & \text{if} \quad \sigma_\mathrm{t} > 3.5 \text{ ms}. \end{cases} \tag{2.4.44}$$

The probability-based soft-masks of the directional classifiers, i.e. $\mathcal{M}_{\Delta\mathrm{L}}$ and $\mathcal{M}_{\Delta\mathrm{t}}$, are calculated in the same way as shown for algorithm CLP in Equation (2.4.34), however, using the respective lookup tables of the ELT algorithm. The computation of the a posteriori lookup tables is presented in Chapter 3.3.3. These masks are then averaged with the concomitant sub matrixes $[n_\mathrm{L}\ n_\mathrm{L}]$ and $[n_\mathrm{t}\ n_\mathrm{t}]$, followed by a frequency-dependent cue combination and multiplication rule with the masks of the statistical penalty measure:

$$\mathcal{M}_\mathrm{elt}^\mathrm{ff} = \begin{cases} \mathcal{M}_{\Delta\mathrm{L}}\mathcal{M}_{\sigma\mathrm{L}}^{e_{\sigma\mathrm{L}}} & \text{if} \quad m \leq m_\mathrm{xo} \\ (1-\xi)\mathcal{M}_{\Delta\mathrm{L}}\mathcal{M}_{\sigma\mathrm{L}}^{e_{\sigma\mathrm{L}}} + \xi\mathcal{M}_{\Delta\mathrm{t}}\mathcal{M}_{\sigma\mathrm{t}}^{e_{\sigma\mathrm{t}}} & \text{if} \quad m > m_\mathrm{xo} \end{cases}, \tag{2.4.45}$$

[19] In addition, Equation (2.4.41) will be replaced with a magnified ILD calculation method, $\Delta\tilde{L}$, which is introduced in Equation (3.3.7) in Chapter 3.3.2. For the purpose of speech enhancement Equation (3.3.7) is executed at the DFT resolution.

where ξ, $e_{\sigma \mathrm{L}}$, $e_{\sigma \mathrm{t}}$ and m_{xo} are the balancing factor of the binaural cues, the expansion/compression exponents of the statistical penalty masks and a lower cutoff modulation frequency parameter for employing the ITD of the envelope, respectively. Based on the statistical analysis on binaural cues in Chapter 3.3.2, the lower cutoff frequency parameter m_{xo} has been adjusted to a modulation centre frequency of 78 Hz.

By means of the subsequent magnitude-based multiplication of the soft-mask $\mathcal{M}_{\mathrm{elt}}^{\mathrm{ff}}$ with the modulus of the original modulation spectra in both channels, lateral interference as well as centre and modulation frequency bins showing unreliable binaural cues are attenuated. The high frequency modulation spectrum above 400 Hz is appended to the filtered modulation spectrum and left unchanged. In addition, the original phase of the complex modulation spectrum is not altered in the mask-based filtering process. The approach was found beneficial during the adjustment of the algorithm as well as in view of the inverse transform of the altered modulation spectrum to the altered envelope signal. Regarding this algorithmic choice, Paliwal et al. (2011) showed that the phase of a complex modulation spectrum does hardly contribute to speech intelligibility in frames of short duration.

In the following, an example is given to illustrate the modulation-based weighting process. Therefore a mix of two speakers is created. The target speaker is in frontal direction and an interfering speaker at 270 deg (clockwise). The upper row of Figure 2.11 images different processing steps of algorithm ELT. First, the upper left-hand plot presents the combined centre and modulation frequency spectrum of two speakers in the left channel at a certain point in time. The upper middle plot shows the weighting function based on the ILD cue. The mask $\mathcal{M}_{\mathrm{elt}}^{\mathrm{ff}}$ is designed such that the energy components that belong to the target in the frontal line are preserved with a weighting value close to one (white). On the other hand energy components that belong to the speaker at 270 deg are attenuated with a weighting value close to zero (black). Finally, the upper right-hand plot shows the centre and modulation frequency spectrum subsequent to the multiplication by the weighting function. Regions with centre and modulation frequency bins that belong to the interfering speaker at 270 deg are attenuated, as identified by darker shades of grey. The centre and modulation frequency spectrum in this example shows mainly a laminar energy distribution. Nevertheless, a distinct modulation pattern at low frequencies, i.e. up to 500 Hz and around 1.8 kHz, can be observed. These regions exemplify how sources that share the same centre frequencies are separated by means of different modulation frequencies. Note that it would be beneficial for the modulation-based filtering process if the periodotopic axis is finer resolved. However, constrained by the well-known uncertainty principle, this implies longer analysis windows in which the short-time stationarity of speech is easily exceeded.

Subsequent to a first filtering in the centre and modulation frequency domain, the filtered envelope $\mathcal{E}'$ of the left and right channel is retrieved by applying the inverse of Equation (2.4.36), i.e. the IDFT of the altered centre and modulation frequency

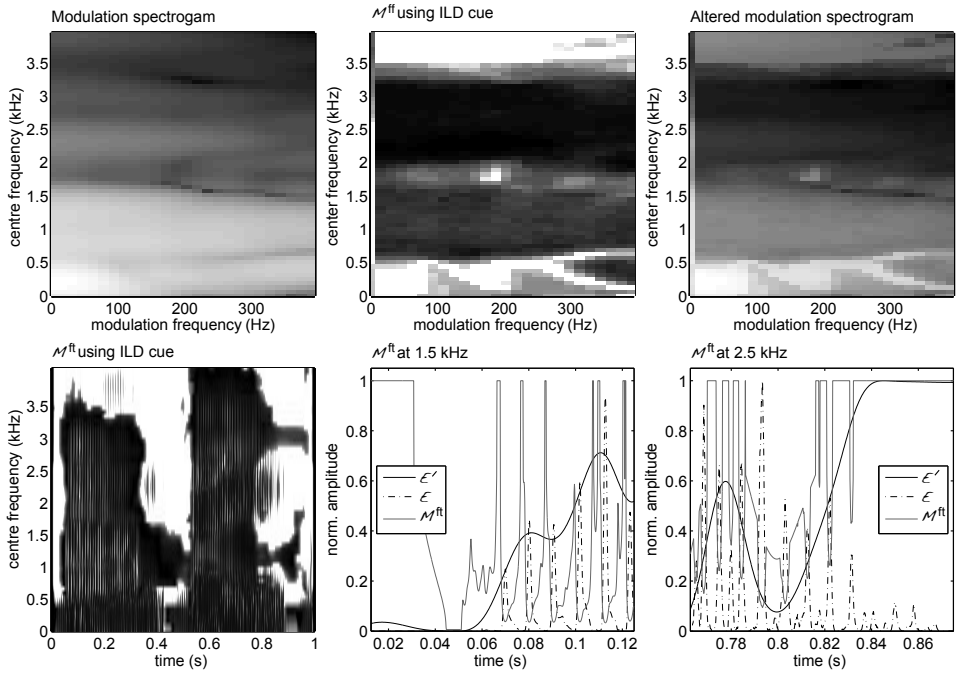

Figure 2.11: *The upper plots show a series of the algorithm ELT-based speech enhancement in the carrier and modulation frequency domain from left to right: the upper left-hand plot shows a mixture of two speakers. The upper middle plot shows a masking function $\mathcal{M}_{\mathrm{elt}}^{\mathrm{ff}}$ using only the ILD cue, and the upper right-hand plot shows the resulting centre and modulation frequency spectrum subsequent to the multiplication of the original mixture with the mask. The lower plots illustrate a typical weighting function $\mathcal{M}_{\mathrm{elt}}^{\mathrm{ft}}$: the lower left-hand plot is a top view on the mask in the time-frequency domain, the lower middle plot and the lower right-hand plot show time slices through the mask and the envelope functions at different time ranges and frequencies.*

spectrum.

Kollmeier and Koch (1994) proposed a correction factor, i.e. the root of the quotient of the altered envelope and the original envelope, as a means to compensate for a phase loss of the altered envelope signal.[20] In our research, this approach was adopted because of our empirical observation of an improved noise suppression, despite the fact that the present implementation of algorithm ELT retains the original phase throughout the signal flow. However, not the square root of the quotient is

[20]It remains unclear to the author whether Kollmeier and Koch (1994) implied with that statement a loss of the group delay of the envelope or the absence of the sinusoidal phase information.

taken (cf., Kollmeier and Koch, 1994), but an adjustable exponent e, of the quotient is introduced, which allows for a compression, or expansion, of the weighting function:

$$\acute{\mathcal{M}}_{\mathrm{elt}}^{\mathrm{ft}}(d,n) = \left(\frac{\overline{\mathcal{E}'\{x_{\mathrm{l}}(d,n)\}}}{\max[\mathcal{E}\{x_{\mathrm{l}}(d,n)\}, \epsilon]} \right)^{e}. \tag{2.4.46}$$

The parameter ϵ in Equation (2.4.46) is introduced to avoid a division by zero, and the overscore signifies a lowpass operation, in which the filtered envelope is smoothed with a finite impulse response in the form:

$$\overline{\mathcal{E}'\{x_{\mathrm{l}}(d,n)\}} = \sum_{q=0}^{N_{\mathrm{lp}}-1} \Xi(q)\mathcal{E}'\{x_{\mathrm{l}}(d,n-q)\}, \tag{2.4.47}$$

where $\Xi(q)$ is an impulse response with $q = 0, 1, \ldots, N_{\mathrm{lp}} - 1$ tabs.[21] The lowpass operation introduces an overall interference suppression, which enables a broad masking, i.e. time-frequency bins are bound to bigger clusters at higher frequencies as well as in regions of speaker dominance. A fine-scale masking, i.e. a modulation correction, is observed everywhere else. The lower left-hand plot in Figure 2.11 depicts a typical soft-mask $\acute{\mathcal{M}}_{\mathrm{elt}}^{\mathrm{ft}}$ in the centre frequency domain. Thereby the filter order N_{lp} and the cutoff frequency d_{xo} of the lowpass filter determine the balancing between low frequency and high frequency masking. These parameters were found to be crucial for the overall algorithmic success, and are, hence, part of the optimization in Chapter 5.2.4.

Subsequently, the raw soft-mask is lower and upper bounded with:

$$\mathcal{M}_{\mathrm{elt}}^{\mathrm{ft}}(d,n) = \min[\max[\acute{\mathcal{M}}_{\mathrm{elt}}^{\mathrm{ft}}(d,n), A_{\min}], A_{\max}]. \tag{2.4.48}$$

The lower middle and lower right-hand plot in Fig. 2.11 display a range of possible proportions between the original envelope $\mathcal{E}$, the altered envelope $\mathcal{E}'$ as well as $\mathcal{M}_{\mathrm{elt}}^{\mathrm{ft}}$ at two centre frequencies (see titles). From the relative amplitudes, there are three cases that can be differentiated:

(1) $\mathcal{E} >> \mathcal{E}'$ the weighting factor is close to zero. A tendency of low-gain clustering with neighbouring bins is observed.

(2) $\mathcal{E} \approx \mathcal{E}'$ dominant peaks are suppressed to about the proportion of $(\overline{\mathcal{E}'}/\mathcal{E})^{e}$, which results in a fine-scale masking.

(3) $\mathcal{E} << \mathcal{E}'$ the weighting factor is close to one. A tendency of high-gain clustering with neighbouring bins is observed.

[21] The lowpass filter was designed using the MATLAB (The MathWorks TM) function `fir1.m` with the filter order N_{lp} and the cutoff frequency d_{xo}. The filtering process was performed with the MATLAB function `filtfilt.m`, which introduces no phase delay by the filtering process.

Finally, the magnitude of the original left and right channel STFT representation are multiplied by the mask $\mathcal{M}_{\text{elt}}^{\text{ft}}$ and transformed back to the time-domain stimulus with an IDFT, followed by an overlap-add technique. As opposed to the algorithms CLP and CC, algorithm ELT uses a Hanning synthesis window of length N_ι to interleave the reconstructed short-time frames, which can have an offset due to high amplitudes in $\mathcal{M}_{\text{elt}}^{\text{ft}}$. In the filtering process, the original carrier phase information is retained, an approach that has shown to be beneficial for speech intelligibility enhancement using the modulation-based filter approach (Paliwal et al., 2011).

The fixed and algorithmic parameters of processor ELT are given in Table 2.3. In order to provide a comparison with the original implementation of Kollmeier and Koch (1994), the published parameters of their implementation are given in the second row. An important difference with the original implementation lies in the analysis frame length of the complex bandpass signal ($N_{\mathring{\chi}}$ in Table 2.3). The reason for this modification was inspired by our statistical observation that the requirement of short-time stationarity of speech is violated with the application of an analysis frame-length of 40 ms, as proposed by Kollmeier and Koch (1994). As previously mentioned, although modulation frequencies resolve better on longer time analysis windows, a mix of sources tends to merge in the binaural modulation domain and this leads to incorrect envelope ITD (or envelope-based IPD) and envelope ILD values. Therefore, the analysis of the bandpass envelope signal was performed with the shorter frame-length of 16 ms and an overlap of 50 %.

In listening tests Kollmeier and Koch (1994) and Wittkop et al. (1997) showed that algorithm ELT yields a small but robust improvement of speech intelligibility at very low mixing SNR conditions, but generally no improvement in coherent noise setups. It is the question of the present work whether this improvement can be verified or even improved with the present reviewed and optimized implementation of the algorithm.

Summary

This chapter introduced the signal model of the combined processing scheme, which comprises a bilaterally applied beamforming front-end and a CASA-based post-processor. Practical beamforming solutions were presented that realize a compromise between directivity and stability. In the second part of this chapter, the concept of time-frequency masking was introduced. A statistical analysis of the energy distribution of several sources in the time-frequency domain brought to light to what extent speech-in-noise problems may differ despite a common global RMS-based SNR. Subsequently, the cepstral smoothing technique was introduced to possibly improve the quality of the post-filter output. In the last section of this chapter three binaural speech processors were introduced, that showed the potential for speech and quality improvement.

The present work focusses on the improvement given by a set of binaural speech

Table 2.3: *Parameters and parameter ranges of algorithm ELT. ϵ_{hist} determines the main-lobe of the algorithm. A description of this parameter is given in Chapter 3.3.3. The second row in this table gives the parameters of the implementation of Kollmeier and Koch (1994). Note that not every parameter can be directly compared, since the present algorithm deviates in several details form the original implementation. The differences are discussed in this chapter.*

fixed parameters

f_{s}	N_{χ}	N_{d}	ΔT	$\mathring{\alpha}$
16 kHz	64 bin (4 ms)	256 bin	2 bin (0.125 ms)	$\tau = 8$ ms
25 kHz	64 bin (2,6 ms)	128 bin	4 bin (0.16 ms)	unknown

f_{s} (STFT domain)	$N_{\mathring{\chi}}$	N_{dd}	$\Delta\mathring{T}$
8 kHz	128 bin (16 ms)	512 bin	64 bin (8 ms)
6.25 kHz	256 bin (40 ms)	512 bin	64 bin (10 ms)

$[n_{\sigma\text{L}}\ n_{\sigma\text{L}}]$	$[n_{\sigma\text{t}}\ n_{\sigma\text{t}}]$
5×5 bin	5×5 bin
5×5 bin	5×5 bin

algorithmic parameters

$[n_{\text{L}}\ n_{\text{L}}]$	$[n_{\text{t}}\ n_{\text{t}}]$	ξ	$e_{\sigma\text{L}}$	$e_{\sigma\text{t}}$
n×n bin; n ∈ 1…10	n×n bin; n ∈ 1…10	0 … 1	0.1 … 3	0.1 … 3
5×5 bin	5×5 bin	0.7	2	2

N_{lp}	d_{xo}	ϵ	e
50 … 350	20 … 800 Hz	1×10^{-6} … 0.01	0.1 … 2
unknown	unknown	unknown	0.5

A_{max}	A_{min}	ϵ_{hist}
2 … 10	0.01 … 0.5	1 … 5
10	0.1	–

processors that are serially connected to bilaterally applied beamforming front-ends. To this end, the following chapter presents a statistical analysis of binaural cues at the output of different bilaterally applied front-ends in varying noise conditions. Subsequently, a novel pattern-based weighting method that accounts for the statistics of binaural cues in dissimilar acoustics is introduced. Finally, in the last chapter of this work, the binaural CASA algorithms are optimized and assessed through a range of noise conditions.

3

Binaural parameter statistics and optimal pattern-based noise suppression

As a means for efficient improvement of intelligibility of noise-corrupted speech, the previous chapter introduced a series of binaural CASA speech processors, in combination with several bilaterally applied beamforming front-ends. These CASA processors share the application of interaural cues[1] in the separation process. While humans use a combination of multiple interaural and monaural cues in the localization and source separation process, each of the here presented binaural algorithms has only access to one or two parameters of interaural disparity.[2] Although of high spatial acuity in anechoic conditions, interaural parameters show a sensitive nature and become weak indicators of direction in diffuse noise conditions. It is, therefore, required to understand their behavior in different noise environments and to apply them in the best possible way in the source separation process.

To achieve this, the current chapter studies interaural parameters of the fine-structure and the envelope of the waveform in a statistical analysis. Following that, a Bayesian classification method is introduced for the establishment of pattern-based weighting functions in algorithms CLP and ELT. This probabilistic approach increases the robustness of the binaural algorithms at low SNRs and under diffuse conditions. Prior to the statistical analysis, it appears worthwhile to summarize briefly a set of features of the twofold auditory scene analysis (ASA) process and the role of interaural parameters therein, on the basis of psycho-acoustical and physiological findings. By examining their nature, much of the localization and eventual separation possibilities with interaural cues in noise can readily be understood.

[1] Herein 'interaural cues' are interchangeably used with their computational analog, i.e. 'interaural parameters'.

[2] Briefly it has been mentioned in Chapter 2 that the algorithms CLP and ELT additionally make use of the proximity and concurrency of signals throughout their feature spaces in terms of an averaging process. However, this constitutes a very rudimentary algorithmic grouping scheme as opposed to the conjectured schema-based binding approach in the auditory processing.

3.1 Psycho-acoustical and physiological background

The two-fold ASA approach, i.e. the decomposition of an ambient scene into its constituent components of perception and the binding of these cues at different neural layers to build coherent streams of the sources involved, is a complex reciprocal system of bottom-up and schema-driven top-down processes (Blauert, 1997). These two processes cannot work independently in a robust speech perception process. What appears obvious for the schema-driven processing, essentially the formulation of hypotheses about the percept based on available cues, is rather surprising for the primitive grouping process.
Barker (2006) summarized a series of psycho-acoustical studies that describe this phenomenon. It was found that central cues in the primitive grouping process can be neglected, such as e.g. the fundamental frequency cue in whispered speech, while the integration into streams remains well maintained. Moreover, primitive grouping fails to deliver a robust basis for the binding process at abrupt phoneme transitions and can generally not explain the integration of words into streams. Obviously, flexible and sturdy track guiding is provided by the schema-driven top-down process. Much of the schemata seem to be learned through life. A striking example are click sounds in some African languages, which do not integrate in coherent speech streams for people not belonging to this speech group.

Therefore, it is assumed that schema-driven processes play a dominant role in speech perception. Barker (2006, p. 308) writes: "Primitive processes perhaps limit the space over which schema-driven processes need to search in order to arrive at a correct hypothesis." Besides the differentiation in bottom-up and top-down processes, there seems to exist a hierarchy of grouping cues. It was found that articulatory features of the vocal tract overrule spatial cues and that features that are invariant on a high level, as accent and speaker identity, play a strong role in the grouping process.

Considering the complex ASA process, a system far from universally understood, the binaural speech processors in this work are simplified equivalents of the bottom-up process.[3] As described in the introduction, algorithms CLP and CC approximate the peripheral analysis with an STFT and subsequently perform an interaural comparison of the acoustic waveform. Algorithm ELT alternatively decomposes the STFT bandpass signals of the left and right ear channel into modulation frequencies. Subsequently, the interaural differences of a certain sound scene are inferred from the temporal structure of the signal envelope across carrier and modulation frequencies.

Physiological and psycho-acoustical studies demonstrate that these CASA approaches functionally mimic complementary parts of the actual binaural ASA processing,

[3]Later in this chapter a pattern recognition strategy for binaural cues is introduced, which improves the source separation process. This approach can be understood as an attempt of bridging the gap between bottom-up and top-down processes of the auditory system, in terms of binaural processing. For practical application, it would require a classification algorithm to switch between different scenes in order to apply the pattern-based separation algorithm efficiently.

where none of them is adopting all necessary components of the underlying binaural model ASA process. In fact, the different binaural IPD calculation methods in algorithms CLP and ELT (fine-structure and envelope, respectively), correspond to the two channel processing of IPDs, as found in physiological and psycho-acoustical studies of mammals (see e.g., Dietz et al., 2009). These findings corroborate the belief of an independent processing of the fine-structure or carrier IPDs in the medial superior olive (MSO) and a combined processing of the envelope IPDs and ILDs in the lateral superior olive (LSO). Based on these insights, the ILD computation from the binaural fine-structure waveform in algorithm CLP can be considered an approximation of the ILD cue assessment from the binaural envelope signal of the model ASA process.

With respect to algorithm CC, i.e. the coherence based speech processors, we find that apart from a strong relation between the binaural interaction process and the binaural waveform coherence, there appears to be no substantiation for auditory interaural coherence processing. Among other opposing reason, Van de Par et al. (2001) for example, outline that the accuracy needed to normalize the cross-correlation cannot be provided by the auditory system. In the light of these results, the binaural coherence processing of algorithm CC is merely an implicit model to explain the underlying auditory process of assigning a high perceptual weight to a sound component, if its binaural waveform is coherent (Rakerd and Hartmann, 2010).

To date the perceptual ranking of the interdependent directional cues, i.e. the onset and steady state fine-structure ITDs[4], the envelope ITDs and the ILDs, have not been clarified to the full extent (Stern et al., 2006). In addition to strong personal differences, many parameters, like loudness, onset slope and envelope shape, play an important role in the perception of each of these cues. Furthermore, the weighting of each of these cues in the localization process is strongly influenced by the SNR and the composition of the sound field, as e.g. the degree of diffuseness of the interfering noise.

With regard to the frequency dependency of each of these directional cues, it is well-known that the envelope IPD and ILD provide lateralization over the entire frequency range. The carrier IPD, on the other hand, is only assessed in the lower frequency spectrum, up to about 1.3 kHz in humans. This frequency limit is given through the ambiguity in the spatial sampling at higher frequencies, i.e. above the spatial Nyquist frequency, which starts from about 900 Hz, and which is additionally imposed by the loss of phase locking of the neural firing rate in that frequency range as the stimulus travels up the auditory nerve.

[4]If not explicitly noted, the IPD and ITD are interchangeably used to describe 'interaural temporal disparity', as both cues equally describe the lateralization phenomena mentioned in this introduction. There are, however, also opposing findings with narrow-band lateralization experiments that support the sole application of the IPD in the auditory processing (Zhang and Hartmann, 2006).

Dietz et al. (2009) analyzed the trading[5] of interaural cues, i.e. the envelope IPD and the carrier IPD, using sinusoidal amplitude-modulated tones. In psycho-acoustical experiments the authors found a linear relationship in the trading between ILD and fine-structure IPD as well as between ILD and envelope IPD. Furthermore, the authors obtained a nonlinear association among the fine-structure IPD and the envelope IPD; specifically, such that the lateralization of a fine-structure IPD at 45 deg and around 1 kHz needs a maximum envelope IPD to be counteracted. By developing a two-channel model for the fine-structure and envelope IPDs, Dietz et al. (2009) were able to model lateralization independent of psycho-acoustical evaluation.

Rakerd and Hartmann (2010) studied interaural temporal cues in noise. The authors confirmed their hypothesis, in which the degree of coherence is the only determining parameter for the perceptual relevance of the carrier and envelope ITD, although the responsiveness varied strongly with frequency. At low frequencies, in their study an octave band with a centre frequency of 225 Hz, where the coherence is physically high (the authors suggest the term 'physically compressed'), the coherence had to be large as compared to mid-bands, in their study a band centred at 715 Hz, to achieve the same degree of localization distinctness. The same accounts for high frequency bands, in their study a band centred at 2850 Hz, where only the envelope ITD allows temporal lateralization. With the restrictions that the authors used noise bands to derive their results (the envelope shape affects the results), they found that envelope ITDs are of no additional use under adverse conditions. Overall, the authors concluded that humans gradually favour other directional cues than the ITDs as the coherence decreases.

To conclude this brief survey, it is expected that the interaural temporal fine-structure differences between the ears are most important if these are available (Stern et al., 2006). In silence, the ILD, together with the monaural cue of the frequency transfer due to the direction from which the wave is impinging onto the pinna,[6] is important in the reduction of front-back confusion and ambiguities caused by the cone of confusion effect. Temporal interaural differences in the binaural envelope signal are used in humans when the coherence of the waveform is high, which is rarely the case in real-world surroundings. In noise, the perceptual ranking of these cues consequently changes, and psycho-acoustical experiments suggests a higher importance of the ILD cue (Rakerd and Hartmann, 2010).

The following statistical study intends to gain an understanding of the applicability of spatial cues in the noise suppression task. The strategy of this study has adopted the methods of the statistical analysis of interaural fine-structure cues of Nix and Hohmann (2006). In their work a DFT is used in the peripheral frequency decompo-

[5]The trading of an IPD cue with an ILD cue has comprehensively been studied in psycho-acoustics to infer the relative importance of binaural cues in the localization process. Usually, the cues are presented from the left and right hemisphere, and are subsequently balanced to retrieve a localization percept from the midline (Dietz et al., 2009).

[6]The monaural phenomenon of localization in the median plane is known by Blauert's Directivity Bands (Blauert, 1997)

sition, prior to an averaging over broader carrier frequency bands. We conceptually repeat this analysis with the binaural processing stage of our implementation of algorithm CLP, on the output of a mannequin and several hearing aids.
As opposed to the statistical analysis of Nix and Hohmann (2006) up to the fourth central moment (i.e. mean, variance, skewness and kurtosis), we limit the analysis to the mean and standard deviation (square root of the variance), as these measures are the most meaningful statistical measures and sufficient within the scope of this work. In the second part of this chapter, we apply the statistical analysis to interaural envelope cues. Consequently, the binaural stage of algorithm ELT is applied to analyze binaural envelope cues up to maximum first fundamental (F0) pitch frequencies, on the output of a mannequin and several hearing aids. Preparatory to the statistical study, the following subsection deals with the data collection.

3.2 Data collection

It is a central theme of this study on binaural statistics and subsequent speech intelligibility enhancement, to analyze contrasting acoustical scenes with different hearing aids on an artificial head. Additionally to the possibility of creating these combinations, the target signal and the interfering signals, such as a coherent jammer or a diffuse background, need to be interchangeable and their spatial distributions readily determined. To simplify the effort of the data collection, directional information is gathered by recording the HRTFs of the left and right ear for each hearing aid, with the artificial head in different positions in a horizontal plane at the ear-level. These HRTFs are used as a convolution kernel for the spatialization[7] of monaural speech recordings at arbitrary directions. By superimposing spatialized speech recordings with the real-world background scenes, which were recorded with the identical recording setups, i.e. with each hearing aid setup, a great variety of acoustical scenes can be created.

In the course of this data collection, the HRTF recordings were measured in the anechoic room of the TU Delft. The setup for the recording consisted of an arc that allows the placement of a loudspeaker at adjustable elevation angles, while keeping a constant radius of 1 m to the centre of the arc, where the artificial head is positioned. As artificial head, the system of the Institute of Technical Acoustics at RWTH, Germany (in this work referred to as the Aachen head) was used without the optional A/D conversion and equalization processor stage (Schmitz, 1995). In the present work, the RME Fireface 400 was used for the amplification and A/D conversion. The Aachen head features the ear moulds of an individual, whose HRTFs were selected on the basis of sound location accuracy and a small front-back error in

[7]The term spatialization is used in this work for describing the spatial arrangement of a sound source with respect to the receiver. It is closely related to auralization (Vorländer, 2008, p. 103), yet emphasizes the spatial rendering, i.e. in terms of space perception the phenomenon of localization.

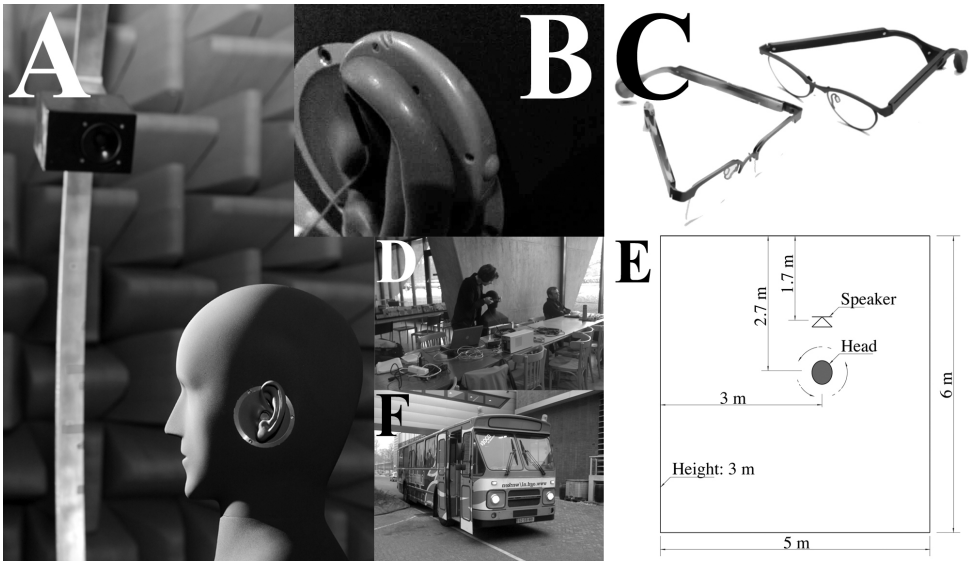

Figure 3.1: *Illustration of the data collection:* A *shows the Aachen head in the anechoic chamber of the TU Delft with the playback setup used in the HRTF recording.* B *displays the BTE hearing aid of ReSound type Canta 4 470-D and* C *depicts two versions of the hearing glasses of Varibel Innovations BV. Photograph* D *shows the preparation of the canteen recording.* F *shows the autobus, which was used for the car noise recording, and* E *depicts the sketch of the room that was simulated with a mirror image source model (MISM) for different degrees of reverberation. In the MISM setup the HRTFs of the Aachen head are applied. The height of the speaker and the head was adjusted to 2 m above the floor.*

a free-field listening test. Torso, shoulders and neck were chosen to be close to the respective dimensions of the individual with the best ears and within the ITU-T P.58 recommendation (Schmitz, 1995). The Aachen head system does not model the ear channel and the ear-drum. Although artificial heads perform poorer than real heads in subjective evaluation, the Aachen head is a particular good head system, as found in a survey (Minnaar et al., 2001) and ideal for the extensive and reference-based testing with different hearing aids that are mounted on it.
The free-field equalization of the HRTFs as well as the equalization of the loudspeaker was performed with the recording of the transfer function of the loudspeaker at the location of the head with a B&K type 4950 1/2-inch free-field microphone. During the HRTF measurements, the Aachen head was mounted on a B&K turntable type 3921 and the B&K control unit type 3922 was used for remote access.

A logarithmic sweep technique was used for all HRTF recordings (Hulsebos, 2004). This recording technique offers a high and constant SNR over the frequency range

with the applied measurement setup, and allows for the isolation of harmonic distortions in the measurement process. Using this property, the resulting HRTFs are free from the harmonic distortions of the recording chain, including the hearing aids, which are particularly sensitive to these distortions at greater sound pressure levels. As the dynamic range was strongly determined by the hearing aid type, the stimulus was three seconds long throughout all measurements and selectively repeated to ensure a high SNR.

The applied speech material is taken from the Multilingual TNO Human Factors database (TNO, 2000) and in the Dutch language. The sentences of this database were developed by Plomp and Mimpen (1979). Prior to the application of this speech material, pauses in the sentences were excluded by a simple VAD method[8] and normalized to a common RMS level. For the calculation of statistical measures of one speech source in a diffuse background noise, the monaural speech recordings of three speakers (one female and two male speakers) were superimposed. In this manner, a continuous stream of speech sounds is gathered, while speech modulation is still present.

In turn, the hearing aids were placed on the Aachen head and different directivity programmes were chosen. The measurement routine was set up to measure in steps of 1 deg in the horizontal plane at ear-level with a sampling frequency of 44.1 kHz and a word length of 16 bit. Both the BTE hearing aid (GN ReSound type Canta 4 470-D) and the hearing glasses (Varibel Innovations BV) were adjusted to have neither frequency dependent amplification, nor compression. Furthermore, the signals were picked up before their respective receivers, i.e. the hearing aid speakers. The HRTFs of the Aachen head were measured with the identical setup, in which the output was taken from the built-in in-ear microphones. Table 3.1 shows the possible combinations of the hearing aids of different directional modes, as well as the Aachen head, with the scenes that have been recorded.
Spatially distributed noise sources were recorded in three different real world environments. The surroundings were selected to sample opposing acoustical characteristics. A lively canteen was chosen for its diffuse ambiance with varying speech sources in time and space. A rather invariant sound scene was recorded in a bus with a dominant diesel engine at constant highway speed. The third real-world recording was made in a workshop of architectural model making. It features a set of high frequency sounds of saws, grinders and model printers. Therein, some sources were rather static in time and space, while others were transient. As already mentioned, with respect to the HRTF measurements, the identical setups of the different binaural receivers were applied. By such means, it is assumed that the HRTFs and the transfer functions of the instruments are matched throughout all measurements.
Graphs of the long-term spectra of the sound scene recordings with different hearing aids and programmes, as well as the Aachen head, are shown in Figure 3.2.

[8]Short-time speech frames of 32 ms length with 50 % overlap were excluded if their RMS level was below 40 dB with respect to the overall RMS level of the connected discourse.

Table 3.1: *Receiver and background scene combinations are given. The BTE hearing aid is a GN ReSound type Canta 4 470-D and the hearing glasses (HG) are the Varibel spectacles. The $*$ denotes a recording error. As a result only one channel was recorded. Based on a temporal decorrelation a binaural signal was later generated. At the time this hearing aid background recording is used in this work, the reader is reminded of this fact.*

Recorder	canteen	autobus	workshop
HG (low directivity)	x	x	x*
HG (high directivity)	x	x	x*
BTE (omni)	x	x	x
BTE (directivity)	x	x	x
Aachen head	x	x	x

The Aachen head (Row E in Figure 3.2) gives a fair reference as the best approx-

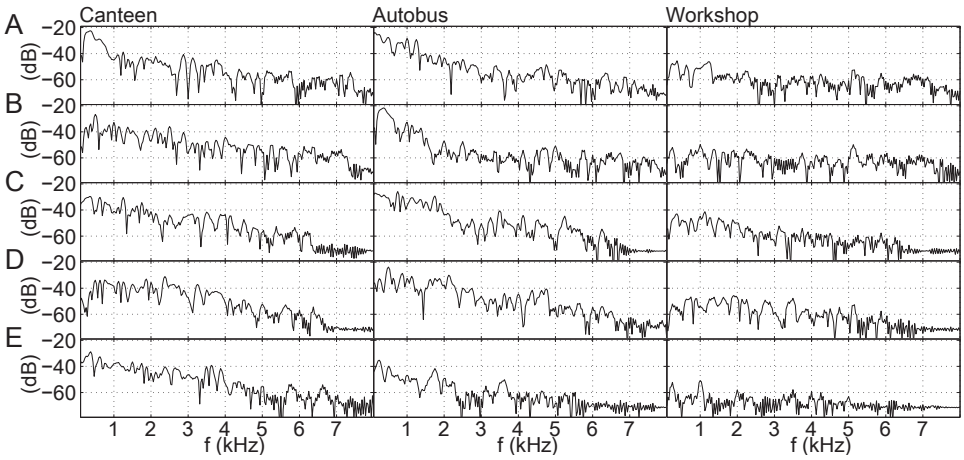

Figure 3.2: *The long-term spectra of different receivers in different acoustic scenes. Prior to the calculation of the spectra, the left and the right channel of a 30 s segment were combined into a monaural signal through averaging. The hearing aids and modes are:* A *hearing glasses low directivity,* B *hearing glasses high directivity,* C *BTE in omni-directional mode,* D *BTE in directivity mode and* E *the Aachen head.*

imation of the actual physical sound field spectra, since the hearing aids generally exhibit highly fluctuating transfer functions in lateral positions. Looking at these spectra, the canteen environment shows a typical long-term speech spectrum with significant energy up to 4 kHz. The autobus spectrum shows prominent signal en-

ergy at low frequencies and the workshop with many high frequency tools features a white noise background character. Compared to the spectra of the hearing aids, the hearing glasses (Row A and B in Fig. 3.2) seem to alter the spectra less than the BTE hearing aid (Row C and D). Moreover, the BTE hearing aid transfer functions roll off above 6.5 kHz. Looking at the workshop spectrum, similar to white noise, the BTE imposes a certain degree of frequency shaping, whereas the hearing glasses offer a flat transfer-function up to 8 kHz in both programme modes.

As mentioned, Table 3.1 lists the combinations of scenes and receivers that can subsequently be digitally superimposed. For the simulation of a particular scene, the global (i.e. waveform-based) SNR between a target speaker and a background scene are calculated at the ear-level. Specifically, the global SNRs are calculated with the long-term and binaurally averaged intensity levels, i.e. the RMSs, of the connected discourse and the noise. This approach is adopted from Nix and Hohmann (2006) and constitutes a crucial point, as it excludes the directional SNR characteristics of the Aachen head and the hearing aids. It can be considered a suitable approach for the statistical analysis of the interaural parameters at a constant SNR, in particular with respect to the comparison among the different hearing aids and the Aachen head. Moreover, as will be shown later in this work, the SNR enhancement of a binaural processor is mainly determined by the ear-level SNR, which supports the hypothesis of a strong dependency between the ear-level SNR and the statistics of the binaural cues.

Figure 3.1 depicts the hearing aids and a set of recording environments. It also gives a sketch (Figure 3.1, drawing E) of the MISM simulation setup, which was used for an additional binaural cue analysis in simulated reverberation (Van Dorp Schuitman, 2009). For that purpose, two different reverberation characteristics with reverberation times of 0.2 and 0.8 s were applied to the HRTFs of the Aachen head in the horizontal plane at 1 deg steps. Using the relation $r_{\mathrm{RT}} = 0.1\sqrt{\frac{V}{\pi \mathrm{RT}}}$, reverberation radii of 1.2 and 0.6 m were simulated. During the simulation first and second order reflections were modeled with the MISM technique, whereas higher order reflection were appended to the IR using statistical modeling. The virtual sources were omnidirectional. This complements the artificial sound scene mixing of clean speech material and a diffuse background, described above, where the target signal lacks reverberation. Based on this data collection across different sound scenes, the following subsection presents the statistical analysis of binaural parameters.

3.3 Statistical analysis of interaural parameters

The analysis of interaural parameters is calculated with the frequency resolution of auditory filters. This approach differs from the binaural speech processors of this work, which operate at the DFT resolution in order to reach a high degree of disjointness of concurrent sources. Yet, by calculating binaural parameters across

auditory filters, we gain a more general insight into the statistics and facilitate the comparison with physiological and psycho-acoustical data. Therefore, throughout the following analysis, the spectral organization of carrier frequencies is based on classical critical bands of Zwicker and Terhardt (1980), and the decomposition of modulation frequencies up to high F0 pitch frequencies is based on constant relative bandwidth filters of 1/3 octave, which is an approximation of the critical modulation filterbank observed in psycho-acoustical tests (Kollmeier and Koch, 1994; Dau et al., 1997).

Instead of applying refined binaural auditory algorithms that are intended to mimic physiological and psycho-acoustical data as close as possible (see e.g., Dietz et al. (2009)), we utilize the algorithms CLP and ELT up to the binaural stage and apply an averaging over auditory-based frequency bands. The application of the analysis stage of binaural speech processors for the calculation of binaural cues obviously leads to differences with more accurate physiological and psycho-acoustical models. Nevertheless, the approach reflects the statistics of the binaural cues as they are available in the speech enhancement process of the applied speech processors. Despite this deviation, the results should still reveal general facts of the model ASA process.
The algorithmic details of the speech processors did not differ from the parameters given in the Tables 2.2 and 2.3. The only difference resides in the execution of the critical band averaging (of the auto and the cross power densities), prior to the calculation of the binaural cues.

As this work offers only a limited scope for a statistical analysis of binaural cues, a selection of scenes and binaural front-ends had to be made. Therefore, out of the measured acoustic background scenes (see Table 3.1), the canteen recording was selected for subsequent statistical analysis. This condition is challenging for speech intelligibility and as such troublesome in daily experience for many people. For the comparison of the binaural cues at the output of a binaural beamformer and natural ears, we opted for the hearing glasses in low directivity mode, in the following referred to as the HG (low directivity), and the Aachen head, respectively. As will be shown, the HG (low directivity) offer well defined binaural parameters in free-field conditions and good listening ease (Merks, 2000). The HG (low directivity) are therefore considered a suitable candidate for a binaural speech processor front-end. Binaural cues are furthermore analyzed in the presence of coherent interference. Lastly, reverberation is applied to concurrent speakers.

In the following subsection, the interaural parameters are first computed from the fine-structure to infer their statistics. Thereafter, the same is done with the binaural parameters of the envelope. With the exception of an analysis of binaural cues of the fine-structure in silence at the output of all binaural receivers employed in this work, both subsections investigate the interaural parameters in identical acoustical scenes, thereby allowing estimation of the separation power of the binaural parameters used in the binaural speech processors.

3.3.1 Inference from the fine-structure of the binaural signal

In this subsection a statistical analysis of interaural parameters is derived from the fine-structure of the waveform. Using algorithm CLP (and the implication of a Nyquist frequency of 8 kHz), an allocation of the DFT bins into 21 critical bands (using the index d_{cb}) of Zwicker and Terhardt (1980) as defined in ANSI/ASA (2007) was performed by adding the squared magnitude spectra and the complex valued cross spectra of the left and right channel:

$$\overline{x_{\mathrm{ll}}(d_{\mathrm{cb}}, n)} = \sum_{d=d_{\mathrm{l}}(d_{\mathrm{cb}})}^{d_{\mathrm{u}}(d_{\mathrm{cb}})} |x_{\mathrm{l}}(d, n)|^2, \tag{3.3.1}$$

$$\overline{x_{\mathrm{rr}}(d_{\mathrm{cb}}, n)} = \sum_{d=d_{\mathrm{l}}(d_{\mathrm{cb}})}^{d_{\mathrm{u}}(d_{\mathrm{cb}})} |x_{\mathrm{r}}(d, n)|^2, \tag{3.3.2}$$

$$\overline{x_{\mathrm{rl}}(d_{\mathrm{cb}}, n)} = \sum_{d=d_{\mathrm{l}}(d_{\mathrm{cb}})}^{d_{\mathrm{u}}(d_{\mathrm{cb}})} x_{\mathrm{l}}(d, n) x_{\mathrm{r}}^*(d, n), \tag{3.3.3}$$

where the overscore denotes the averaged quantities. In the summation process d_{l} and d_{u} denote the lower and the upper cutoff DFT frequency coefficients, respectively, of each critical band d_{cb}. Subsequently, the spectra were subjected to a temporal smoothing according to Equation (2.3.16), i.e. a first order recursive low-pass filter with a time constant $\breve{\tau}$ of 8 ms. The IPDs and ILDs are calculated analog to Equation (2.4.31) and (2.4.32), respectively, per critical band d_{cb}. Note that the smoothing in Equation (3.3.3) results in an intensity-weighted IPD per critical band d_{cb}. See the comment of Hohmann in Goupell and Hartmann (2007) for alternative calculation methods, as well as the above-introduced smoothing method and implications thereof.

The manner of calculating interaural parameters differs from the method used in Nix and Hohmann (2006) in terms of the sampling frequency, DFT frame-length and spectral resolution. For the last one, Nix and Hohmann (2006) used a finer scaling, more specifically, a bandwidth of 0.57 times the equivalent rectangular bandwidth (ERB), leading to 43 adjacent frequency channels in their study. The lower resolution chosen in the present work is considered to be adequate for a general analysis. Since this study is regarded as a preparation for speech enhancement, the algorithmic parameters of Table 2.2 were held constant, as previously mentioned. In doing so, the following analysis is limited to an upper frequency of 8 kHz, although it is generally known that humans are able to analyze ILDs at frequencies higher than this limit.

Based on this short-time analysis of binaural parameters, histograms for the IPD and ILD cues were generated. The histograms have each 200 bins in the range of $-\pi$ to π for the IPD cue and in the range of -40 dB to 40 dB for the ILD cue. After

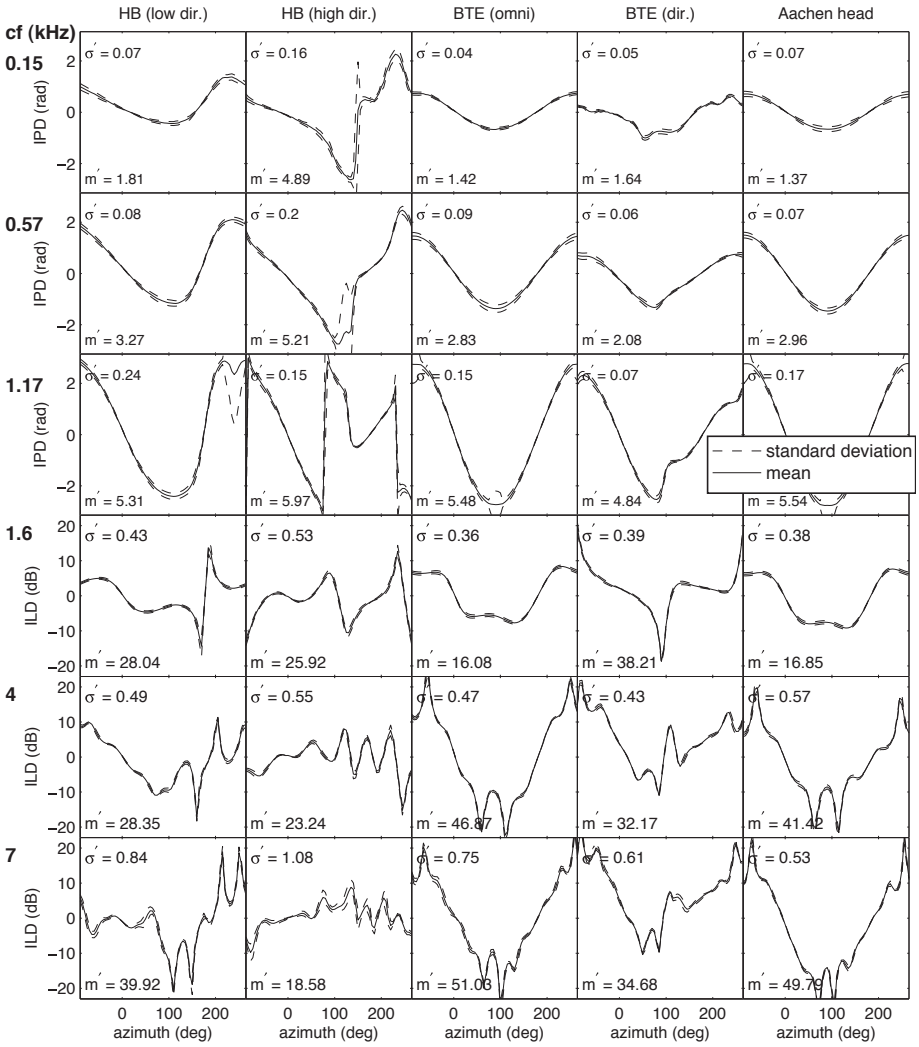

Figure 3.3: *The interaural cues IPD and ILD of the fine-structure waveform are analyzed in terms of parameterized PDFs as a function of azimuth and critical bands, centered at the frequencies (*cf*) of 0.15, 0.57 and 1.17 kHz for the IPD and at 1.6, 4 and 7 kHz for the ILD. The columns juxtapose four hearing aid modes and the Aachen head. In the plots, the solid lines give the mean and the dashed lines give the standard deviation around the mean, both as a function of source azimuth. The value σ' shows the mean standard deviation over all azimuths analyzed, and m' indicates the maximum azimuthal range of the mean.*

the binning processes, the histograms were normalized to yield probability density functions (PDFs).

Interaural parameters inferred from the fine-structure in free-field conditions

Figure 3.3 gives a first analysis of the binaural parameters of the four hearing aids and the artificial head at an SNR of 60 dB. In each subplot, an IPD or ILD of a critical band is depicted for a revolving sound source from −90 to 270 deg in steps of 5 deg, using the above-described speech material. The solid line in each plot represents the mean value of the respective binaural parameter PDFs as a function of source azimuth. At the left bottom of each plot, m' indicates the maximum range (i.e. variation of the mean) of the binaural cue mean as a function of azimuth.[9] The dashed lines in each plot give the standard deviation σ of the respective binaural parameter PDFs as a function of source azimuth around the mean. The single value σ' (top left corner in each plot) accounts for the mean standard deviation over all angles.

In Figure 3.3 the fine-structure IPD parameter is analyzed in the critical bands with the centre frequencies of 0.15, 0.57 and 1.17 kHz and the fine-structure ILD parameter is given in the critical bands with the centre frequencies of 1.6, 4 and 7 kHz. This first analysis highlights several characteristics of the short-time interaural parameters at the output of beamformers.

(1) A juxtaposition of the artificial head with the hearing aids shows a comparable curvature of the mean value for hearing aids with low directivity, as observed with HG (low directivity) and the BTE in the directional mode. This effect is stronger in the absence of directivity, as seen with the BTE in the omnidirectional mode.

(2) The interaural parameters at the output of the hearing glasses in the high directivity mode differ strongly from natural interaural cues. With respect to the IPD, this difference is predominantly observed in the rear horizontal plane. The ILD shows a fluctuating behavior and has no resemblance with the natural ILDs. The reason why the localization with the HG in the high directivity mode shows good results (Merks, 2000), might be attributed to the IPDs in the frontal plane, which are comparable to natural IPDs. The IPD cue of the waveform's fine-structure has shown to overrule the ILD cue easily if both cues are conflicting (Rakerd and Hartmann, 2010). In addition, learning effects are well-known to improve the localization with artificial cues too.

(3) The interaural parameters of the Aachen head and the BTE in the omnidirectional mode show a symmetry around the frontal plane (coronal plane). This symmetry is a consequence of homogeneous IPDs and ILDs on concentric

[9] The equations of the range m' and the standard deviation σ' are given in Appendix A.3.

circles around the intercranical axis, well-known and previously introduced as the cone of confusion.

(4) No symmetry is exhibited by the IPDs and ILDs at the output of the directional hearing aids. Their IPDs expose a difference in slope between the frontal and the rear hemisphere and the carrier ILDs tend to fluctuate in the rear hemisphere. Thus, while the IPD at the output of the directional front-end shows twisted cone of confusion artifacts, the ILD introduces ambiguities. Considering a source in the median plane that is to be enhanced by a binaural speech processor based on a binaural ILD weighting, potential sources at many locations in the rear of the head and in a particular frequency band might not be suppressed. However, the ILD fluctuation is different for each DFT frequency band and a summation over all frequency bands, as performed in the synthesis of the speech signals, should result in the intended attenuation of a coherent interferer in the rear hemisphere. If the interferer is incoherent and the PDFs are broad, obviously less discrimination power will result from interaural parameters.

(5) As expected, the standard deviation is small throughout the analysis in free-field conditions. Nonetheless, some fluctuation of the binaural parameters is present. This fluctuation is higher for sources at lateral positions and around singularities of the arctan function in the calculation of the IPD cue. Nix and Hohmann (2006) indicate the source of this random fluctuation in the short-time processing. Since the windowed short-time analysis has a time aligned hop size in the left and right channel, a directional deviation of sources from the median plane might be analyzed in one channel (window) earlier than in the other. This leads to incorrect interaural parameters, an effect that is cumulative along lateral positions. Therefore, the short-time processing introduces a certain amount of randomness.

Overall, the statistical analysis of interaural fine-structure parameters in free-field conditions shows that these parameters give a fair to excellent indication of direction at the output of all binaural front-ends that have been studied.

Interaural parameters inferred from the fine-structure in noise

When two signals are added, their waveform distributions undergo a convolution (Hartmann, 1997). A similar behavior is observed for interaural parameter distributions if the signals lack disjointness in the time-frequency domain, see e.g. Roman et al. (2003). Hence, an undesirable implication of CASA-based source separation is the manner in which the binaural parameters of a source are subjected to the nature and strength of the interference. Clearly, the binaural parameters in noise differ from the same quantities in free-field conditions. Consequently, a binaural weighting function in a binaural speech processor that applies the binaural reference

parameters of free-field conditions, will likely fail to unravel a complex mixture of sources.

As mentioned before, a seminal study of interaural parameters of the fine-structure in noise was given by Nix and Hohmann (2006). Their results are briefly recapitulated.

- Binaural parameters strongly fluctuate in frequency regions in which the noise level is close to the signal level or higher.
- In free-field conditions, the ILD parameter is a good indicator of direction with a mean corresponding to the direction and a narrow distribution. If, however, the SNR is low, the standard deviation builds up and the mean shifts to the median plane. The ILD becomes a weak parameter of direction.
- It was found that the mean of the IPD shifts to the median plane too, but much less than the ILD. Nevertheless, the standard deviation of its distribution is highly increased at low SNRs.
- A dependence of the statistics of interaural parameters on direction was demonstrated. For the ILD Nix and Hohmann (2006) observe a high standard deviation at lateral positions. This finding is in contrast with the observations of the IPD parameter, for which higher directivity was found.
- The noise type, they concluded, is of lesser importance to the statistics than the SNR in the analyzed frequency band.

In preparation for the formulation of an efficient and flexible approach of a binaural parameter-based weighting process in the second part of this chapter, we will in the following discuss the most salient characteristics, for an adequate placement of the statistical analysis in the context of this work. Therefore, we restrict the analysis, as mentioned before, to the comparison between the HG (low directivity) and the Aachen head. This enables us to compare the interaural parameters at the output of a beamforming front-end with their natural counterparts. The same pair of binaural receivers will be used in the following section, which presents the statistics of binaural envelope parameters.

The succeeding analysis examines the fine-structure ILD parameter in the critical bands of the centre frequencies at 0.15, 2.15 and 7 kHz, and at SNRs of 0 and 10 dB. The speech material is a sample of three superimposed speakers with as background a lively canteen, as described in Chapter 3.2 on the measurement setup.

Figure 3.4 shows the results of the Aachen head recording in row A and the results of the HG (low directivity) recording in row B. A series of observations can be made:

(1) As compared to the free-field conditions, the range m' of the ILD mean is strongly compressed. While a range of about 50 dB is found in the free-field conditions with the Aachen head at a centre frequency of 7 kHz, an m' of only

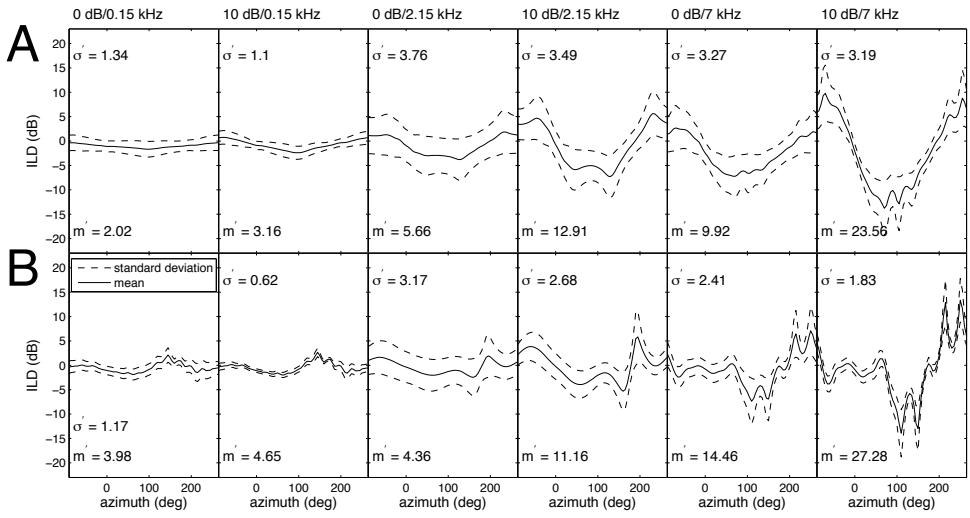

Figure 3.4: *The PDFs with mean (solid line) and standard deviation (dashed line) of the fine-structure ILD parameter as a function of source direction, analyzed in different critical bands and at the SNR of 0 and 10 dB. The centre frequencies are given in the titles. Row* A *refers to the binaural output of the Aachen head and Row* B *presents the results at the output of the HG (low directivity).*

about 10 dB remains in the 0 dB mixing condition. Due to the dependency on frequency of the head shadow effect, m' decreases towards lower centre frequencies. Consequently, the fine-structure ILD parameter is a poor criterion of direction for sounds with a wavelength greater than the dimensions of the head, in particular in noise.

(2) At wavelengths considerably smaller than the dimensions of the head, the fine-structure ILD is shown to moderately mediate an indication of direction, even at low SNRs.

(3) The HG (low directivity) front-end is shown to compress the ILD parameter range m' with respect to the Aachen head. Overall, the ILD parameter shows a much lower slope across the frontal hemisphere and more fluctuation in the rear hemisphere, leading to an increased ambiguity.

(4) The HG (low directivity) exposes a lower averaged standard deviation σ' than the Aachen head. In particular, the growth of σ at lateral positions is not observed with the HG (low directivity). The reason for this behavior might be a combination of the absence of the cone of confusion, as a consequence of the

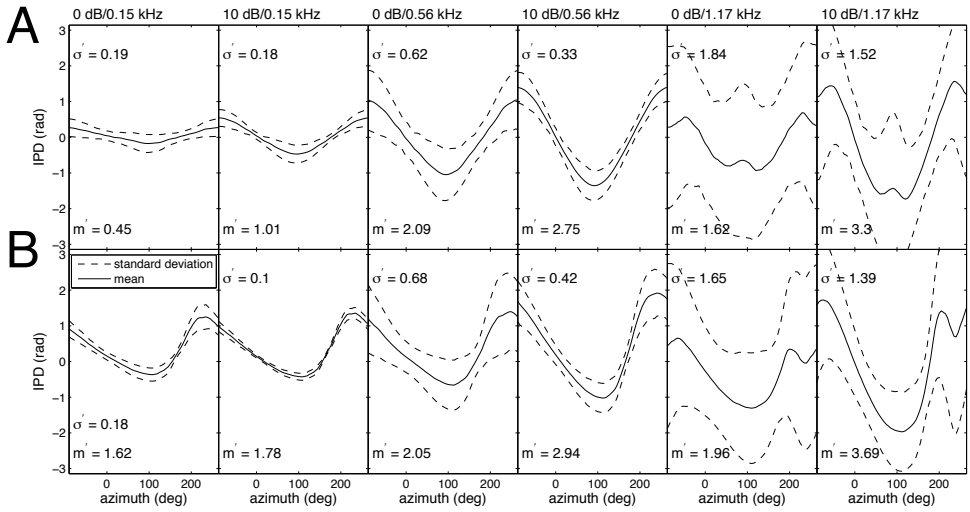

Figure 3.5: *The PDFs with mean (solid line) and standard deviation (dashed line) of the fine-structure IPD parameter as a function of source direction, analyzed in different critical bands and at the SNR of 0 and 10 dB. The centre frequencies are given in the titles. Row* A *refers to the binaural output of the Aachen head and Row* B *presents the results at the output of the HG (low directivity).*

directivity pattern, and the implicitly raised source SNR at lateral positions, due to the suppression of lateral sources by the beamforming of the front-end.

On the whole, the fine-structure ILD parameter as calculated from a short-time analysis, appears to be a moderate source separation criterion in low SNR conditions, for high frequencies. If the wavelength is in the order of the head size and above, the distributions of the ILD parameter for all directions show overlap. In this frequency range, consequently, the ILD is a poor indicator of direction. A disadvantage is observed for the HG (low directivity), which exposes a considerable ILD cue compression in the frontal plane and a high fluctuation in the rear plane.

Next, Figure 3.5 presents the statistics of the fine-structure IPD parameter in noise. Again we list the most important results in row A for the Aachen head and in row B for the HG (low directivity):

(1) As already demonstrated by Nix and Hohmann (2006), we find the mean of the PDFs to be not as much shifted to the median plane as observed with the ILD PDFs at low SNR conditions. Relative to the PDF ranges, the overall standard deviation σ' is higher than found for the ILD, especially at high frequencies.

Moreover, the standard deviation is proportional to the range m' and increases due to the greater phase difference at higher frequencies.

(2) The standard deviation is increased at lateral positions, which is contradicting to the findings of Nix and Hohmann (2006), who found an increased IPD vector strength at lateral positions.

(3) Similar to the observations under free-field conditions (Figure 3.3), at lateral positions and frequencies where the wavelength exceeds the spatial Nyquist limit, a considerable increase of the standard deviation of the IPD can be observed. In Figure 3.5, this is apparent in the critical band with the centre frequency of 1.17 kHz. If we consider IPD lookup tables for binaural speech processors, it is questionable whether the fine-structure IPD, with such a high standard deviation, is a robust indicator of location for the DFT frequency bins beyond the spatial Nyquist limit. For this reason, the possibility of using the IPD for frequencies greater than the spatial Nyquist limit appears to be considerably impeded.

(4) The comparison with the HG (low directivity) reveals a greater range of the mean m' at low frequencies, which is probably due to the increased width of the HG. Together with a lower standard deviation and no symmetry around the frontal plane, the HG (low directivity) front-end might gain an advantage in the binaural separation process. Regardless of this, the spatial Nyquist limit appears to be slightly higher as deduced from the smooth curvature of the mean around 90 deg in the critical band that resides at a centre frequency of 1.17 kHz.

Overall, the fine-structure IPD cue as calculated in a short-time analysis has shown to be a decisive criterion of direction even in low SNR (> 0 dB) conditions up to the spatial Nyquist limit. An advantage in terms of parameter range and standard deviation is observed for the HG (low directivity).

Coherent interference in free-field and reverberant conditions

We finalize this fine-structure analysis of interaural parameters with a brief look at the mix of two speech sources in free-field and reverberant conditions. For that reason, the running speech of two speakers, one female and one male speaker, is mixed at an SNR of 0 dB. Pauses in the speech material are excluded using the above-mentioned VAD procedure. One speech source is kept at 0 deg (female speaker) and the other speech source is changing its location relative to the receiver from -90 to 270 deg (male speaker). In this experiment we skip the moment analysis and directly show the histograms.[10] Figure 3.6 gives the results for the HG (low directivity) in

[10] Although an analysis of the PDFs in terms of mean and standard deviation would simplify the comparison with previous studies in this chapter, this cannot easily visualized for two concurrent

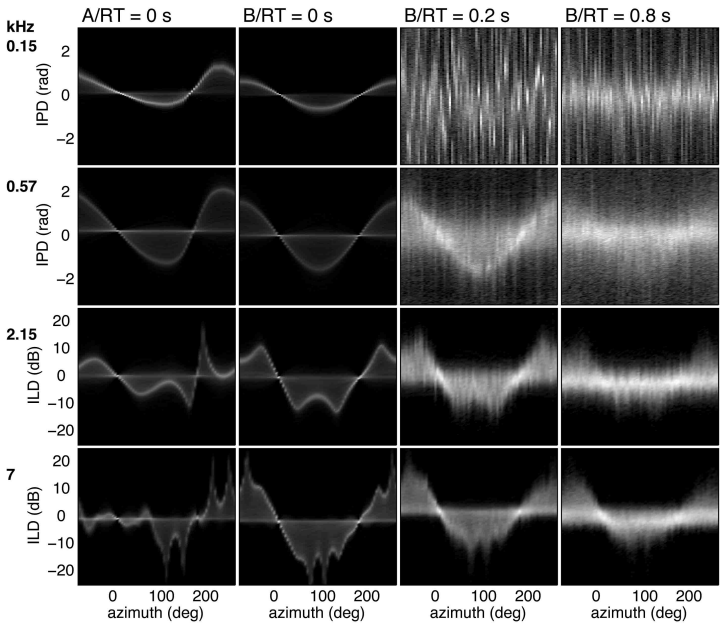

Figure 3.6: *The PDFs of the fine-structure IPD and ILD parameter for two speakers (one at fixed position in frontal direction and one rotating horizontally around the head) as a function of azimuth and critical band. The SNR has been adjusted to 0 dB under free-field and reverberant conditions. Column* A *refers to the HG (low directivity) and the Columns* B *refer to the output of the Aachen head. The simulated reverberation conditions are distinguished through the reverberation time* RT *in the titles.*

anechoic conditions (Column A) and the Aachen head in anechoic conditions as well as two reverberant conditions with reverberation times of 0.2 and 0.8 s (Columns with the title B). We summarize the observations as follows:

(1) In accordance with the findings on the local/global SNR experiment in Chapter 2.3.2, it is expected that about 50 % of the time-frequency bins have a local SNR of 0 dB and higher in favour of one of the two sources. This distribution is clearly seen in the free-field conditions for both the ILD and IPD parameter. The binaural parameter traces reside at quantities (as a function of azimuth) as if no interference was present. This clearly speaks for the disjointness of

sources in a two-dimensional plot. Moreover, the non-symmetrical inclined cumulation due to the interaction of two concurrent sources in the binaural domain can only be appropriately assessed with higher statistics. Therefore, the PDFs are directly plotted here.

two sources in the STFT domain and illustrates the fundamental reason why binaural speech processors can achieve a remarkable suppression of coherent noise sources. Nevertheless, there is a percentage of overlap between the original traces, which should ideally be accounted for in a binaural speech processor weighting function.

(2) For both fine-structure parameters, the main traces show a high compactness, even for the HG (low directivity) and the critical band with a centre frequency of 7 kHz. Due to the fluctuating nature of binaural cues around the mid-line, the HG (low directivity) introduces more ambiguity, especially in the 7 kHz band shown here.

(3) In critical bands at lower centre frequencies (here 150 Hz), there appears to be a higher IPD vector strength of lateral sources than of concurrently active sources in the median plane. Experiments with two male speakers (not presented here), however, demonstrate this to be a result of the spectral difference between the female and male voices.

(4) Artificial reverberation strongly affects the binaural phase at low frequencies. A consistent IPD parameter cannot be measured in the presence of reverberation, in particular in the critical band with a centre frequency of 150 Hz, neither under the condition of a reverberation time of 0.8 s, nor 0.2 s. This observation especially accounts for the 0.2 s situation, which might indicate the MISM implementation to be the cause of this deficiency. Moreover, the resolvability of low frequencies is also impeded by the block-wise DFT-based processing. Psycho-acoustically more plausible is a peripheral Gammatone filterbank analysis method that offers a continuity signal processing. Generally, the loss of binaural phase information complies with psycho-acoustical studies. As the binaural waveform coherence is physically compressed in this frequency range, any diffusion or noise contamination severely degrades binaural phase cues (Rakerd and Hartmann, 2010).

(5) In bands with medium centre frequencies the binaural signals expose greater phase differences (0.57 kHz in Figure 3.6) and, thus, a certain degree of degradation on the binaural fine-structure by reverberation, might still allow for a distinct phase contrast. This is observed in the experiment. At RT = 0.2 s we find a blurred but a clear phase pattern, whereas at RT = 0.8 s the pattern is smooth, but starts to blur beyond recognition due to the increased amount of reverberation. Interestingly, the mean phase difference as a function of azimuth, i.e. the range of the IPD cue, is hardly modified by mild reverberation.

(6) The ILD parameter is shown to be a moderately accurate indicator of sound source localization in reverberation. Although the traces are blurred, the overall mean of the distributions is not as much shifted to the mean as found in the diffuse background of the canteen situation above. Standing waves, as

they occur in small rooms, might impede the ILD localization in reverberating rooms. This characteristic of rectangular rooms is not modeled with the applied MISM method.

(7) A small binaural offset is observed for all histograms and median values. Whether the reason is of acoustical origin or due to a measurement error could not be clarified.

Conclusions

This subsection analyzed interaural parameters of the waveform fine-structure at the output of different binaural receivers in free-field conditions, as well as in the presence of coherent and incoherent interference.

A salient insight of the simulations is that directional hearing aids alter the front-back ambiguity of natural binaural cues. This ambiguity is, in three dimensions, well-known as the cone of confusion artifact for narrow-band sounds. Hence, besides the attenuation of sounds from the side and the rear due to the directional processing of the beamformer, the cone of confusion is warped.[11]

Whereas the fine-structure IPD is only moderately changed by a directional front-end, the ILD parameter offers a narrower range around the mean in the frontal hemisphere and a much increased oscillation around the mean, hence ambiguity, behind the head. Whether this implies a disadvantage in the source separation process will be addressed in Chapter 5.

As a difficult scene of real-life situations, we opted to present the study on the interaural parameters of a speech source in the diffuse and time-invariant canteen noise situation. Apart from the lowest critical bands at the output of the Aachen head, we found the fine-structure IPD to be a useable cue of direction, even at an SNR of 0 dB and up to the spatial Nyquist limit. The same holds to a lesser extent for the ILD above the spatial Nyquist limit. These findings generally correspond to physiological and psycho-acoustical data on binaural fine-structure cues, which commonly state that localization and intelligibility are generally possible above an SNR of 0 dB (Stern et al., 2006).

With respect to the influence of diffuse noise on binaural parameters, the present study corresponds to the initial findings given by Nix and Hohmann (2006). The mean of the fine-structure ILD is strongly shifted to the median plane and the standard deviation is increased. The increase of the standard deviation of the fine-structure IPD is higher relative to the maximum parameter range of the mean. Its mean, however, is less shifted to the median plane, as compared to the fine-structure ILD.

[11] In a side study it was found that this warping of artificial binaural parameters leads to a unique binaural 'labeling' for each source direction. Consequently, the front-back confusion can be widely eliminated with CASA-based localization algorithms (Opdam, 2010; Ketwaru, 2010).

A quantitative comparison between the results of Nix and Hohmann (2006) and our results cannot easily be carried out, due to different bandwidths used in the calculation of the binaural parameters as well as different setups and sound material. However, we have the impression that our results are qualitatively in agreement with their results. By way of example, Nix and Hohmann (2006, Table II) explicitly give the central moments for a source at 60 deg in a lively cafeteria. At an SNR of 0 dB and at the ERB band with a centre frequency of 2.88 kHz, these authors find an ILD of 2.52 dB and a standard deviation of 4.36 dB. The simulation presented here in Figure 3.4 shows a mean of about −3 dB (the sign is a matter of the orientation in the calculation of the ILD) and a standard deviation of about 5 to 6 dB at a critical band of 2.15 kHz (SNR of 0 dB/canteen background).
In the same SNR and location setup, Nix and Hohmann (2006, Table III) report a mean IPD of 2.85 rad and an IPD standard deviation of 1.48 rad at the ERB band with a centre frequency of 540 Hz. Using the Aachen head and the same setting (but different equipment and sound material), we obtain different values but comparable proportions between mean and standard deviation. The mean IPD is of about 1.5 rad and the IPD standard deviation is of about 1 rad in the critical band with a centre frequency of 560 Hz.

In a second study, we analyzed interaural parameters in the presence of a coherent interferer at an SNR of 0 dB in anechoic and reverberant conditions. In anechoic conditions, the quantities of the ILD and IPD parameters clustered around their free-field values. In reverberation we find the ILD to be a more competent parameter in the determination of direction than the IPD parameter.

The disjointness of two sources in anechoic conditions is the key to the success of binaural speech processors in these tasks. Generally, the challenge lies in the correct activation of interaural parameters in more difficult situations. An evolutionary optimization of algorithm CLP in Chapter 5, through a set of difficult conditions will further complement our understanding as to what extent each of the two fine-structure cues can be applied in the computational separation process.

3.3.2 Inference from the envelope of the binaural signal

The sensitivity of the auditory system to amplitude modulation in the absence of spectral cues can be assessed with the temporal modulation transfer function (TMTF). A first graph of this transfer-function dates back to the early 20th century (Riesz, 1928). About fifty years later, psycho-acoustical and physiological studies started to propose an auditory separation of envelopes through modulation filters at higher stages of the auditory nerve. The popular Dau model mimics these findings (Dau et al., 1997). Due to the success of this algorithmic approach in explaining modulation-related phenomena, it is an essential part of many recent auditory models, as e.g. models of speech intelligibility (Christiansen et al., 2010).

The introduced CASA speech processor of Kollmeier and Koch (1994) in Chapter

Table 3.2: *Centre frequencies (Hz) of the 1/3 octave modulation filterbank applied in the analysis of binaural parameters of the envelope. The frequencies have been rounded to the nearest integer.*

40	50	63	79	100	126	159	200	252	317

2.4, algorithm ELT, is conceptually similar in that it establishes a three-dimensional decomposition in the left and the right channel, i.e. a time-varying decomposition of centre and modulation frequencies. The difference with the Dau model lies in the DFT-bin processing, as opposed to an auditory carrier and modulation filter analogy, and in the absence of the nonlinear adaption stage. The ELT processor, on the other hand, is extended by a binaural interaction stage, of which the interaural parameters are analyzed in the following.

To generalize the following analysis, the DFT-bins of the carrier and modulation frequencies were subdivided and averaged into broader frequency bands. The centre frequency bins were again averaged according to the critical band definition given in ANSI (S3.5-1997, Table I), and the modulation frequency bins were averaged with a filter bank of constant relative bandwidth of 1/3 octave. Such a frequency spacing was found to approximate psycho-acoustical modulation tuning curves (see the introduction in (Kollmeier and Koch, 1994)). The centre frequencies belonging to the 1/3 octave filter bank used in this work are given in Table 3.2.

Consequently, the DFT modulation spectra across centre frequencies are grouped and added with:

$$\overline{\mathring{x}_{\mathrm{ll}}(d_{\mathrm{cb}}, m_{\mathrm{cb}}, o)} = \sum_{d=d_{\mathrm{l}}(d_{\mathrm{cb}})}^{d_{\mathrm{u}}(d_{\mathrm{cb}})} \sum_{m=m_{\mathrm{l}}(m_{\mathrm{cb}})}^{m_{\mathrm{u}}(m_{\mathrm{cb}})} |\mathring{x}_{\mathrm{l}}(d, m, o)|^2, \tag{3.3.4}$$

$$\overline{\mathring{x}_{\mathrm{rr}}(d_{\mathrm{cb}}, m_{\mathrm{cb}}, o)} = \sum_{d=d_{\mathrm{l}}(d_{\mathrm{cb}})}^{d_{\mathrm{u}}(d_{\mathrm{cb}})} \sum_{m=m_{\mathrm{l}}(m_{\mathrm{cb}})}^{m_{\mathrm{u}}(m_{\mathrm{cb}})} |\mathring{x}_{\mathrm{r}}(d, m, o)|^2, \tag{3.3.5}$$

$$\overline{\mathring{x}_{\mathrm{rl}}(d_{\mathrm{cb}}, m_{\mathrm{cb}}, o)} = \sum_{d=d_{\mathrm{l}}(d_{\mathrm{cb}})}^{d_{\mathrm{u}}(d_{\mathrm{cb}})} \sum_{m=m_{\mathrm{l}}(m_{\mathrm{cb}})}^{m_{\mathrm{u}}(m_{\mathrm{cb}})} \mathring{x}_{\mathrm{l}}(d, m, o)\mathring{x}_{\mathrm{r}}^{*}(d, m, o), \tag{3.3.6}$$

where the overscore denotes the averaged quantities. In these equations, m_{l} and m_{u} are the lower and the upper cutoff DFT frequency coefficients, respectively, that belong to a particular modulation band with index m_{cb}. The first summation denotes the above-introduced critical band averaging. Subsequently, the auto power and cross power signals in critical bands are time-averaged analog to the Equations (2.4.37) to (2.4.39) with a time constant $\breve{\tau}$ of 8 ms. Then, the interaural parameters are calculated analog to the Equations (2.4.41) and (2.4.42). Due to the time sampling of the complex band pass signals in algorithm ELT with 8 kHz (cf. Table 2.3), the analysis is restricted to an upper frequency of 4 kHz to avoid aliasing.

In the following we restrict the analysis of binaural parameters to focus on the most important questions of this research. Therefore, only the free-field condition and an SNR condition of 0 dB with the previously introduced canteen recording are presented. Furthermore, only a subset of the three-dimensional binaural feature space can be given in the scope of this work. Nevertheless, the study is concentrating on the salient characteristics and a comparison with the binaural parameters of the fine-structure is attempted. The following analysis is linked to Appendix B, where due to the novelty of the present examination, further investigations on the binaural envelope parameters are given.

Interaural envelope parameters at the output of the Aachen head

Based on the binning process of binaural cues, Figure 3.7 gives the interaural differences of the envelope at two centre and modulation (-centre) frequency combinations, $d_{\mathrm{cb}}/m_{\mathrm{cb}}$, as measured on the output of the Aachen head. Since the fundamental fre-

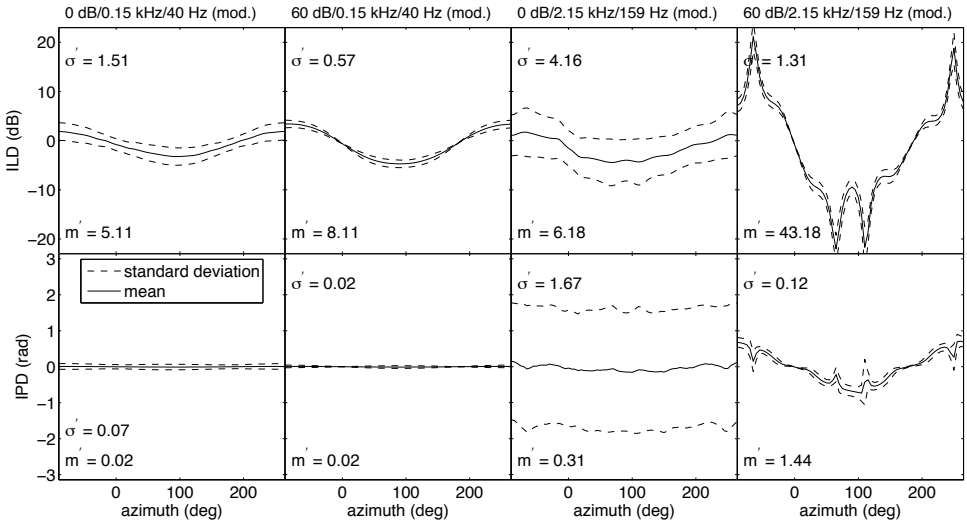

Figure 3.7: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based ILD and IPD parameter as a function of source direction of the Aachen head and band combination. The titles denote the SNR condition, the carrier band centre frequency and the modulation band centre frequency (*mod*).*

quency of speech (F0) starts from approximately 70 Hz, it seems unusual to choose a first modulation band with a centre frequency as low as 40 Hz. The reason is that

algorithm ELT employs interaural parameters in the entire range of carrier and modulation DFT bins up to a maximum F0 frequency of approx. 400 Hz. In addition, many $d_{\mathrm{cb}}/m_{\mathrm{cb}}$ combinations reflect modulation frequencies that are higher than half the bandwidth of the carrier band. The present research consequently encompasses these frequency combinations. Our observations are summarized:

(1) In free-field conditions, we first observe meaningful ILDs of the envelope at a centre frequency of 150 Hz and unambiguous IPDs of the envelope-based IPD at a centre frequency of 2.15 kHz. Hence, as compared to the binaural parameters of the fine-structure, the envelope-based parameters are shown to be applicable in a broader frequency range.

(2) For a low-frequency carrier, here the band with a centre frequency of 150 Hz, the shift of the envelope-based ILD towards the midline is smaller at an SNR of 0 dB, than it has been observed for its fine-structure equivalent. A quantitative inspection of the fine-structure ILD in Figure 3.4 shows an maximum parameter range $m' = 2.02$ dB at 150 Hz in the 0 dB condition, whereas the envelope ILD offers a $m' = 5.11$ dB in the identical setup. Therefore the envelope ILD appears to be a more robust cue of direction at low-frequency carrier bands.

(3) The shift towards the median plane of the envelope ILD is shown to be greater at higher carrier bands. This is illustrated for the $d_{\mathrm{cb}}/m_{\mathrm{cb}}$ combination at 2.15 kHz/159 Hz. The ILD reduces from $m' = 43.18$ dB in the free-field condition to $m' = 6.18$ dB in the 0 dB noise condition. The fine-structure ILD direction-dependent range of the mean with $m' = 5.66$ dB in Figure 3.4 is not much smaller. Presumably interference phenomena at the contralateral ear lead to changes of the modulation frequency and cause the huge interaural difference for some lateral source positions observed in the free-field condition.

(4) Due to the increase of the standard deviation along with the reduction of m', the envelope ILD is considerably weakened as an indicator of direction in the diffuse noise and 0 dB SNR condition. The similarity among the envelope- and fine-structure-based ILD at high frequencies in noise gives a justification for the general approach to calculate the ILD from the fine-structure. Differences are, however, likely to occur in the presence of a concurrent speaker. Then the separation based on different fundamental speech frequencies should theoretically lead to an advantage of the envelope-based ILD processing. In Appendix B Figure B.3, a more exhaustive analysis of the envelope ILD cue in the canteen background at 0 dB is given.

(5) The mean standard deviation σ' of the envelope ILD in the free-field conditions is slightly higher than found with the fine-structure ILD, which is a consequence of the aforementioned fluctuation of the ILD due to the short-term

analysis. Because algorithm ELT is based on a twofold short-time window analysis, the ILD estimation error is higher than in algorithm CLP, which is based on single analysis-window processing.

(6) The mean standard deviation σ' of the envelope IPD is comparable with the fine-structure IPD in a noise-free environment. A dependence between envelope ILD and envelope IPD can be inferred by the fact that the IPD standard deviation is increased at high envelope differences. As the envelope IPD is calculated from an intensity weighted average of phase differences across DFT bins, a strong ILD fluctuation at a particular DFT bin is shown to result in an outlier of the IPD parameter.

(7) The indication of direction formed by the envelope IPD is a function of the low-frequency modulation and, therefore, the phase differences are small for lateral sources. For this reason, at low-frequency modulation bands no meaningful phase difference can be computed, here shown in the modulation band with a centre frequency of 40 Hz. In Appendix B Figure B.2, a more exhaustive analysis of the envelope IPD cue in the free-field conditon is given. Therein it can be observed that it takes a modulation band with a centre frequency of about 100 Hz across all carrier bands to obtain a meaningful phase difference at lateral positions.

(8) Although the envelope IPD circumvents the spatial Nyquist limit at high carrier frequencies, the standard deviation at low SNRs is high and comparable to the fine-structure IPD in similar conditions. In addition, the mean envelope IPD of lateral positions is subject to a shift to the median plane of about the same order as the envelope ILD cue. This behavior is in contradiction with the fine-structure IPD and strongly reduces the distinctness of the envelope IPD in diffuse background conditions at low SNRs. Rakerd and Hartmann (2010) support this result with the finding that the envelope IPD requires a much higher coherence to be a decisive cue of direction, than found for the fine-structure IPD. In Appendix B Figure B.4, a broader analysis of $d_{\mathrm{cb}}/m_{\mathrm{cb}}$ combinations of the envelope-based IPD in the canteen background at an SNR of 0 dB is given.

(9) The range m' of the envelope ILD only slightly decreases towards higher modulation frequencies, as demonstrated in Figure B.1 in Appendix B. This is an important result because it shows that the modulation frequency may be greater than half the bandwidth of the carrier frequency band to calculate valid ILD values. Dau et al. (1997) observed a similar phenomenon in psychoacoustical tests and model simulations. The authors relate it to the leaking of modulation energy among broadly tuned modulation bands and interaural fine-scale differences due to envelope fluctuations, which can be exploited. Correspondingly, we explain this finding by interaural differences of the carrier which persist as a residual in the envelope fluctuations. Regarding speech

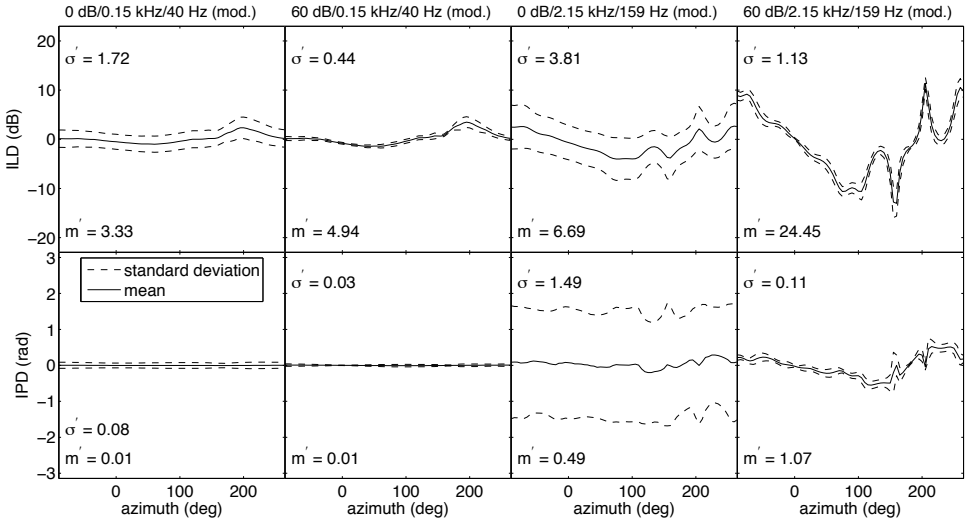

Figure 3.8: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based ILD and IPD parameter as a function of source direction of the HG (low directivity) and band combination. The titles denote the SNR condition, the carrier band centre frequency and the modulation band centre frequency (*mod*).*

enhancement with a DFT approach, the observed phenomenon allows to calculate IPDs across a great many of DFT bins that hold modulation frequencies higher than half the bandwidth of their associated DFT-based carrier bins, as done in Kollmeier and Koch (1994) .

Interaural envelope cues of the HG (low directivity) in a diffuse background

The previous experiment was repeated at the output of HG (low directivity). The results are given in Figure 3.8. Our observations are listed:

(1) In free-field conditions, the envelope ILD of the HG (low directivity) shows a smaller maximum deflection than the envelope ILD of the Aachen head. Thus, the range m' of the envelope ILD in the frontal plane is shown to be compressed. However at an SNR of 0 dB, the parameter range is of approximately the same order as found at the output of the Aachen head. Further combinations of carrier and modulation frequency bands in the free-field condition at the output of the HG (low directivity) are presented in Appendix B Figure B.5.

(2) The envelope ILD of the HG (low directivity) is a poor cue of direction at

low $d_{\mathrm{cb}}/m_{\mathrm{cb}}$ combinations and more so at low SNRs, which demonstrates a disadvantage with respect to the Aachen head.

(3) The envelope IPD at the output of the HG (low directivity) is compressed too and, hence, offers hardly any distinction in the frontal hemisphere in free-field conditions. In the rear, the parameter shows fluctuations. At an SNR of 0 dB, the deteriorated statistical properties of the envelope IPD render any application useless, as it was previously observed with the Aachen head too.

(4) Similar to the statistical findings of the binaural fine-structure parameter, no symmetry around the frontal plane is found, which, as a consequence of the directional processing, gives rise to less front-back errors. The aforementioned fluctuations in the rear hemisphere might, however, lead to confusion with non-symmetrical positions behind the head.

With respect to the intended source separation, an interim summary of the findings above yields disappointing results for interaural envelope parameters in diffuse noise conditions. The envelope IPD has shown to be strongly distorted and cannot deliver an indication of direction. The envelope ILD is less affected by a diffuse sound field at an SNR of 0 dB. Yet, an envelope ILD-based source separation is likely to be strongly hampered by the statistical properties of the short-time parameter too. Figure B.3 and Figure B.4 in Appendix B give further statistical insights under these conditions, for the natural envelope ILD and the natural envelope IPD, respectively.

The comparison between the Aachen head and the HG (low directivity) shows a disadvantage of the binaural envelope statistics for the hearing aids. As it has been shown, the leveling of the directional pressure differences in the frontal plane of the HG (low directivity) has a detrimental effect on both envelope cues. The question that arises at this point is whether the directional front-end will have a disadvantage in the source separation process. Chapter 5.4 will treat this problem.

In preparation for a successful source separation, the calculation of the binaural envelope parameters is altered in the following to compensate partly for the observed susceptibility to noise. As the psycho-physical nature of these parameters cannot fundamentally be changed, the basic aim is to facilitate the usage of the envelope ILD parameter in diffuse noise conditions, and to use the envelope time difference parameter for the suppression of coherent interference. Subsequent to the change of the calculation method, we return to the statistical analysis of the interaural envelope parameters in coherent interference and reverberation, as previously done with the binaural fine-structure parameters.

Magnification of the envelope ILD

The limited potential of short-time interaural envelope cues to distinguish sources at low SNRs and under diffuse conditions suggests the magnification of the binaural

envelope parameters. Interaural magnification (as well as the term) was first introduced by Durlach and Pang (1986) to enhance spatial perception. Later Peissig (1992) analyzed this processing with an algorithm similar to the algorithm of Gaik and Lindemann (1986), i.e. a variant of algorithm CLP of the present work.

We adopt this strategy and introduce a nonlinear magnification of the envelope ILDs by squaring the auto power modulation spectra before calculating the ILD:

$$\overline{\Delta\tilde{L}(d_{\mathrm{cb}}, m_{\mathrm{cb}}, o)} = 10\log_{10}\left(\left[\frac{\overline{\mathring{\phi}_{\mathrm{ll}}(d_{\mathrm{cb}}, m_{\mathrm{cb}}, o)}}{\overline{\mathring{\phi}_{\mathrm{rr}}(d_{\mathrm{cb}}, m_{\mathrm{cb}}, o)}}\right]^2\right), \tag{3.3.7}$$

where the overscore denotes the critical band averaging. The resulting envelope ILDs of this approach are given in Figure 3.9 for the Aachen head in Row A and the HG (low directivity) in Row B.

The plots indicate an amplification of the range m' by a factor of approximately

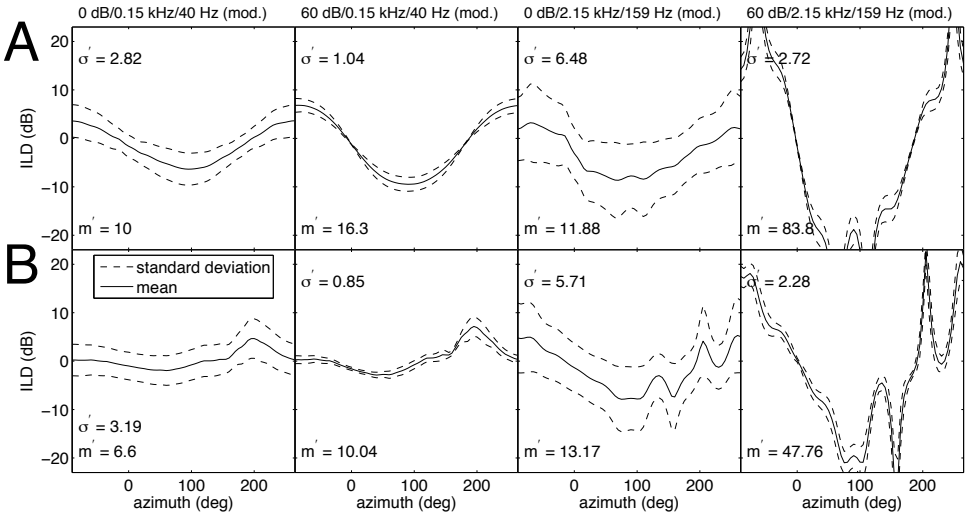

Figure 3.9: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based and magnified ILD as a function of source direction of the Aachen head in Row* A *and the HG (low directivity) in Row* B*. The titles denote the SNR condition, the carrier band centre frequency and the modulation band centre frequency (*mod*).*

two. In a similar manner, the mean standard deviation σ' increases by a factor of approximately two. What appears to offer no benefit on the whole, reveals a compression of the standard deviation at low ILD, i.e. around the median plane. This

is also verifiable for the HG (low directivity). While the envelope ILD is originally compressed around the midline in the frontal hemisphere (see Figure 3.8), the magnification introduces a higher range, i.e. a maximum variation of the mean, as a function of direction. Additionally, a smaller σ' in the frontal direction is observed after application of Equation (3.3.7).
In Appendix B, a more comprehensive overview of this approach in combination with the HG (low directivity) is given. Overall, the advantage resulting from magnification is for most directions offset by an increased standard deviation. Nevertheless, we favour a higher range and a lower standard deviation of the level difference in the median plane, over the original approach. Moreover, since the disjointness of sources in the modulation spectra might be obscured by the compression of the envelope ILD parameter at the output of the HG (low directivity) in the binaural domain, an advantage can be expected in coherent or moderately coherent interference conditions. Consequently, the envelope ILD magnification approach should result in a better source separation under these circumstances.

On the choice of envelope ITD in source separation

Preparatory work on noise suppression using the binaural temporal differences in free-field conditions showed an advantage of the envelope ITD over the envelope IPD. The envelope ITD is calculated analog to Equation (2.4.42), however, per centre and modulation critical band combination, i.e $\overline{\Delta\mathring{t}(d_{\mathrm{cb}}, m_{\mathrm{cb}}, o)}$. Subsequently, PDFs in the range of -3.5 ms to $+3.5$ ms with 500 bins were generated for the following statistical analysis.

As can be seen at low-frequency modulation bands in Figure 3.10, small variations of the IPD lead to high variations of the ITD. As a result of that, the standard deviation of lateral sources is much higher than for sources in the median plane.
Due to the same effect, another benefit of the envelope ITD over the envelope IPD is the reduced frequency at which clear directional labeling is possible. As the comparison between Figure 3.10 and 3.11 shows, the envelope ITD is an indicator of direction, above the modulation band with a centre frequency of 100 Hz upward (under free-field conditions). In the same band, the envelope IPD is less responsive to direction. Yet its standard deviation for lateral sources is lower. The ratio m'/σ' is approximately the same for the envelope ITD and the envelope IPD, however beyond this modulation band the ratio increases in favour of the former.
Finally, the envelope ITD is approximately independent of frequency above approximately a modulation band with a centre frequency of 126 Hz. If, for example, a lateral sound source is to be identified, the envelope ITD approaches a common maximum around 0.7 ms for lateral sources. This represents an advantage in the algorithmic clustering of sources across carrier and modulation frequencies. For these reasons, in this work, the method of calculating temporal differences from the binaural envelope signal is chosen as the standard approach in algorithm ELT.

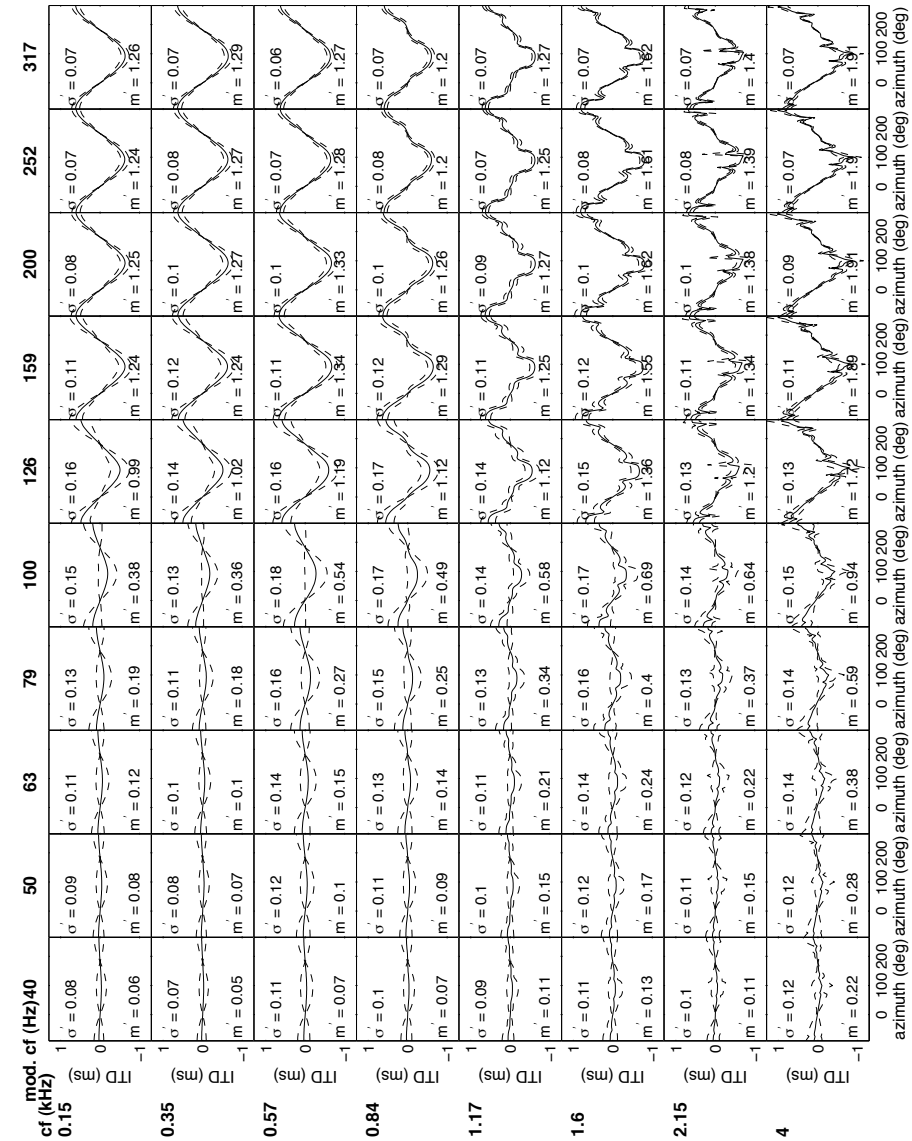

Figure 3.10: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based ITD parameter at the Aachen head as a function of source direction, analyzed in carrier and modulation band combinations that are centred at the specified frequencies (*cf *and* mod. cf*, resp.). The SNR was set to 60 dB.*

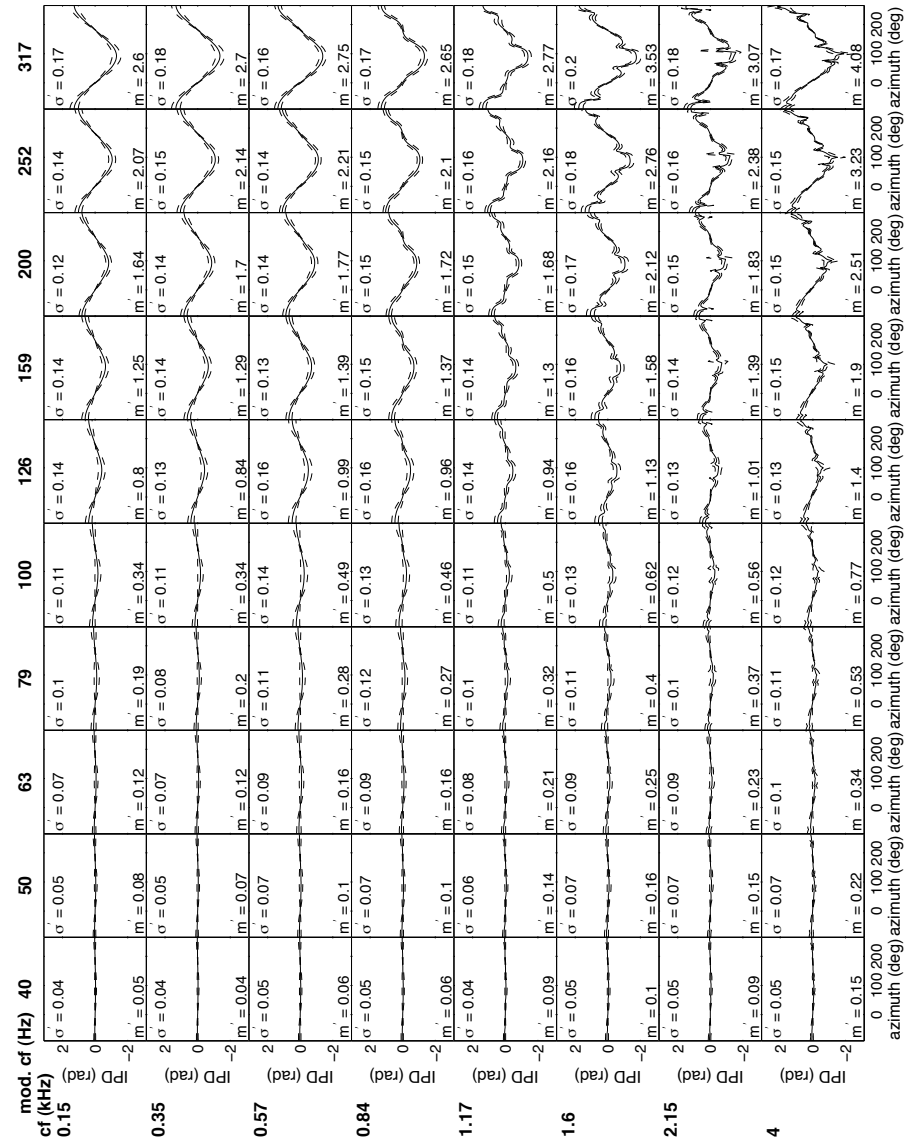

Figure 3.11: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based IPD parameter at the Aachen head as a function of source direction, analyzed in carrier and modulation band combinations that are centred at the specified frequencies (*cf *and* mod. cf*, resp.). The SNR was set to 60 dB.*

An analysis of the envelope ITD at the output of the HG (low directivity), shown in Figure B.8 in Appendix B, reveals an increased standard deviation, in particular at medium carrier frequencies and behind the head. Nevertheless, the interaural analysis at the output of the HG (low directivity) equally benefitted in preliminary source separation exercises when using the envelope ITD.

So far we chose the envelope ITD as a parameter for source separation under free-field conditions. For the sake of completeness, we conclude this study with an analysis of the envelope IPD and the envelope ITD parameter at the output of the Aachen head in the diffuse canteen situation at an SNR of 0 dB. The statistics of these interaural parameter distributions are given in Appendix B in Figure B.4 and B.9. As it is expected from previous inspection in this chapter, both figures demonstrate that neither the envelope ITD nor the envelope IPD are suited to determine direction under these conditions.

Coherent interference in free-field and reverberant conditions

An analysis of two coherent speech sources under free-field and reverberant conditions in the binaural envelope domain, is given. In view of the overall aim of noise suppression, the previously introduced envelope ITD and the magnified envelope ILD are used for generating the PDFs. Despite the different algorithmic approach, the experimentation setups are identical to the setups under which the analysis of the binaural fine-structure cues in free-field and reverberant conditions was carried out.

The results of the analysis are presented in Figure 3.12, where the letter A in the titles refers to the HG (low directivity) and the letter B to the Aachen head. To highlight the most important characteristics, only a subset of carrier and modulation frequency band combinations is depicted. The left-hand side plots represent histograms of modulation bands, of which the centre frequencies are within half the bandwidth of the respective carrier frequency bands. The right-hand side plots show histograms at the centre frequency of the highest modulation filter of 317 Hz used, which is outside each of the employed carrier bands. Our observations of the analysis are:

(1) No disjointness of two sources in the binaural domain of the envelope ITD is observed at combinations of low carrier and modulation frequency bands. The separation of the traces of concurrent speakers increases in higher carrier or modulation bands and for a combination of both. In that sense, the two front-ends show a comparable behavior.

(2) As found in a previous experiment, the envelope ILD is a clear indicator of direction even in the low-frequency carrier band that resides at 150 Hz. The

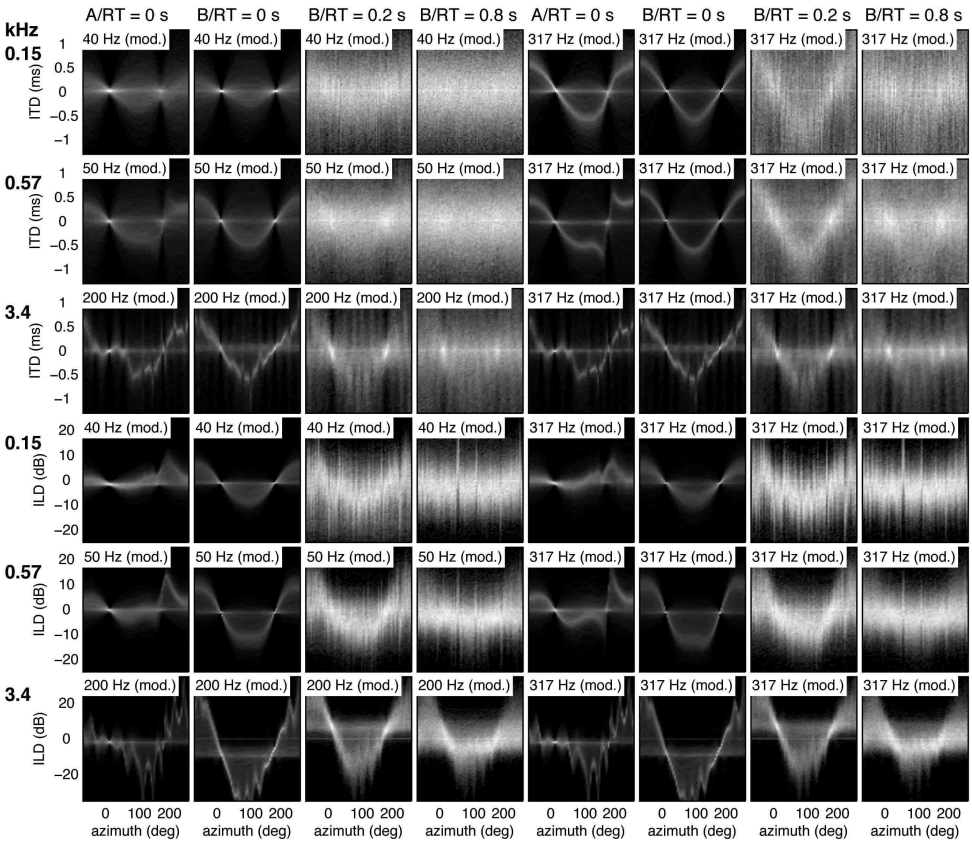

Figure 3.12: *The PDFs of the envelope-based ITD and ILD parameter for two speakers (one at fixed position in frontal direction and one rotating horizontally around the head) are given as a function of azimuth and at combinations of carrier and modulation (*mod*) frequency bands (the centre frequencies are specified for each row). The SNR has been adjusted to 0 dB under free-field and reverberant conditions. Column* A *refers to the HG (low directivity) and the Columns* B *refer to the Aachen head. The simulated reverberation conditions are distinguished through the reverberation time* RT *in the titles.*

traces of the two sources are well separated at the output of the Aachen head. Overall, the disjointness in the binaural domain is less distinct for the HG (low directivity), in particular in low carrier frequency bands.

(3) Generally, no directional information can be obtained from the envelope ITD parameter in mild reverberation. Merely at high carrier and modulation frequency bands, the envelope ITD exhibits an azimuth-dependent distribution.

In spite of that fact, the parameter is severely affected by a high standard deviation of its distribution.

(4) A similar conclusion can be drawn from ILD histograms in reverberation. Whereas these are shown to be rather stable, with respect to their mean, the distributions are wide at all band combinations. This finding corresponds to our observations on the fine-structure ILD and fine-structure IPD: the envelope ILD is shown to be less affected by reverberation than the envelope ITD. Again, the results have to be qualified as reverberation has been simulated with an MISM approach.

(5) The reason for the offset that is observed in some plots could not be identified. It is likely that the envelope along the high frequency carrier is difficult to compute numerically and therefore sensitive to a small but constant mismatch between the left and the right channel.

Conclusions

In the second part of this statistical study, interaural parameters of the binaural envelope signal have been analyzed at combinations of auditory-based carrier and modulation frequency bands. Our findings correspond to psycho-acoustical data in several points.

In the first place, the envelope ITD (or equivalently the envelope IPD) contains directional information in free-field conditions above the F0 modulation frequency of approximately 100 Hz, in our implementation. This parameter of temporal interaural difference is much more susceptible to noise than the fine-structure equivalent and, therefore, it has been observed that no information of direction can be deduced at an SNR of 0 dB, under the diffuse canteen condition analyzed.
This conclusion is supported by the study of Rakerd and Hartmann (2010), who found that a top-level binaural envelope coherence is needed to yield a distinct envelope-based interaural temporal difference cue. In addition, Rakerd and Hartmann (2010) demonstrated that the envelope ITD never achieves the accuracy of the fine-structure ITD. The latter cue has shown to offer a maximum resolution of 1 deg, whereas the envelope ITD offers a resolution limit of about 6 deg. This is approximately the resolution that is provided by the ILD cue in free-field experiments (Stern et al., 2006).
Our analysis shows a certain amount of fuzziness of the interaural envelope parameters, especially for lateral source positions, as expressed through a considerable standard deviation of the short-time parameters in free-field conditions.[12] Finally, the main advantage of the envelope ITD parameter over its fine-structure counterpart is, of course, its applicability above the spatial Nyquist limit under free-field

[12] As it has been already pointed out, the reason for the fuzziness of envelope cues is assumed to originate in the short-time processing of algorithm ELT, especially in the assessment of interaural level differences.

conditions, with and without coherent interference.

With respect to the applicability of the interaural differences of the envelope in the noise suppression task of algorithm ELT, we propose the utilization of the envelope ITD instead of the envelope IPD. The envelope ITD parameter was found to be a better directional indicator at low modulation frequencies as well as to be independent of frequency at higher modulation frequency bands. Furthermore, the envelope ITD does not show side maxima, which is a limitation for the applicability of the fine-structure ITD (Nix and Hohmann, 2006).

The ILD, when calculated from the envelope, corresponds to the fine-structure results with the exception that it remains a parameter of direction at low frequency bands with an m' (maximum range of the mean) that is twice as high, even at an SNR of 0 dB in the analyzed diffuse canteen condition. Hence, in these low SNR and diffuse noise conditions, the ILD has shown to be the only interaural envelope parameter that offers a moderate indication of direction over all frequencies. Therefore, and under these conditions, it can be considered the dominant interaural envelope parameter, which was confirmed in psycho-acoustical cue trading experiments (Rakerd and Hartmann, 2010).

Similar to its fine-structure equivalent, a disadvantage for the envelope ILD parameter is found at the output of the beamforming front-end of the HG (low directivity). Due to the directional weighting, the envelope ILD is shown to be pulled to the midline, a characteristic which is best observed in anechoic and coherent interference situations. As the ILD parameter gains importance in complex noise fields for a successful cocktail party processing in the auditory system, the absence of binaural level differences at the output of a hearing aid will likely have a detrimental effect on speech intelligibility. Therefore, the front-end and the post-processor should at least compensate this disadvantage in terms of an overall speech intelligibility gain. To counteract the compressive nature of the ILD parameter in the frontal hemisphere at the output of the HG (low directivity), a technique known as magnification of interaural parameters has been introduced for a cue expansion (i.e an improved cue range m'). By such means, we aim to improve the binaural processing of algorithm ELT as well as the auditory system.

In summary, we find that the interaural temporal parameters of the envelope in the range of F0 modulations are less robust in noise than it was initially expected during the development of algorithm ELT (Kollmeier and Koch, 1994). Our observations are corroborated by Blauert (1997, p. 333) who summarizes: “It apprears to have been conclusively proven that the mechanism that evaluates interaural envelope time differences is significantly more susceptible to noise than the mechanism that evaluates interaural time differences in the fine-structure of the signal.”
As the binaural resolution deteriorates, the hearing system will discount equivocal cues of binaural difference, in favour of timbre and modulations in the range of the syllabic rate of speech (Barker, 2006). Also the envelope-based F0 and its harmonics, based on quasi-invariant speech sections, are likely a more stable cue in noise

than its short-term derived binaural instants. To date, these model processes are, however, not yet successfully combined in a single CASA-based speech processor. As a step towards a pattern-based application of signal features, the following section introduces a method to employ optimally binaural parameters in the speech enhancement task under different noise conditions.

3.3.3 Pattern-driven source separation

This subsection covers the establishment of pattern-driven weighting functions, which are used in the CASA algorithms ELT and CLP, for the separation of the target speech and noise. The introduction of the chapter already referred to the shortcomings of bottom-up CASA methods in the binding process of a speech stream. Barker (2006) concluded that bottom-up cues merely activate neural structures from which top-down schemata (or hypotheses) are bootstrapped. A simplified analogy of such a top-down schema-driven process in the domain of interaural parameters is presented in the following, based on a Bayesian classification method.

The approach of a principled and robust binaural source separation technique was introduced by Harding et al. (2005) for the design of a CASA front-end in automatic speech recognition systems. The method generates time-frequency soft-masks, based on the probability that the interaural parameter of a certain time-frequency bin is caused by the target source. The strategy is similar to the data-driven mask generation of Madhu (2009b) and Boone et al. (2010); however, as the target source is considered to be fixed at zero azimuth, the processing is much simplified and the probabilities are directly derived from training data, rather than a parametric model.

To calculate soft-masks with a Bayesian a posteriori estimator, the short-time binaural parameters are considered to be a stochastic process with $\vec{\Delta}$ being a time-variant feature vector of binaural parameters at d DFT bins and realizations in, for example, the range of $\pm\pi$ for the IPD. The basis for the stochastic nature of binaural parameters is given by the algorithmic short-time processing and the varying and nonlinear superposition of multiple sources in the binaural domain, as it was identified in the first two parts of the current chapter.

Based on this assumption, it is possible to determine the a posteriori probability of the presence of the target source, if the underlying stochastic process is well approximated by means of an empirical estimation process. This estimation process has to generate a priori knowledge in form of possible feature distributions, i.e. the distributions of binaural parameters of sound from all directions $p(\vec{\Delta})$ and the distributions when only the target $p(\theta_\mathrm{t})$ is present, with θ_t being the target direction. The application these a priori distributions for an optimal (a posteriori) soft-decision

rule can be calculated with the conditional probability:

$$p(\theta_\mathrm{t}|\vec{\Delta}) = \frac{p(\vec{\Delta} \cap \theta_\mathrm{t})}{p(\vec{\Delta})}. \tag{3.3.8}$$

With the expansion of the total probability in the denominator, the equation is rewritten in the familiar form of the Bayesian a posteriori calculation method:

$$p(\theta_\mathrm{t}|\vec{\Delta}) = \frac{p(\vec{\Delta}|\theta_\mathrm{t})p(\theta_\mathrm{t})}{\sum_\theta p(\vec{\Delta}|\theta)p(\theta)}. \tag{3.3.9}$$

The posterior distribution $p(\theta_\mathrm{t}|\vec{\Delta})$ can be directly read at the sample observation $\vec{\Delta}$. To simplify the notation, we denote the numerator of Equation (3.3.9), which is dominated by target signal, $\mathrm{PDF_t}$, and indicate the denominator, which is reflecting binaural parameters of all sources, $\mathrm{PDF_a}$.

Introducing the probabilistic approach to the binaural weighting process greatly simplifies the algorithm and, more importantly, improves the performance of the binaural source separation. While Gaik and Lindemann (1986) and Bodden (1993) already accounted for some of the interactions between competing sources in the binaural domain, the majority of binaural weighting approaches use a clean binaural parameter reference for the establishment of weighting functions in noise (Kollmeier and Koch, 1994; Wittkop and Hohmann, 2003). As previously mentioned, the success of these approaches is severely hindered in complex sound fields.

Binaural parameter statistics have also been applied to sound localization by using a maximum a posteriori (MAP) approach (Nix and Hohmann, 2006). In a side study of the present work (Ketwaru, 2010), this localization approach has been reassessed and applied to different binaural front-ends with and without directivity. In a comparative study with auditory-based anechoic reference or no-reference models (e.g., Albani et al., 1996; Liu et al., 2000; Elzinga, 2010; Opdam, 2010), the statistical approach demonstrated to be superior to every other binaural localization algorithm and, in addition, allowed further improvement in localization accuracy through the application of a beamforming front-end.

As afore-mentioned, in the field of speech intelligibility enhancement, Madhu (2009b) applied a parametric a posteriori approach to establish a data-dependent time-frequency mask with the information delivered from an array-based localizer. In the work of Boone et al. (2010) this approach has been adopted and is applied to a binaural output of the HG (low directivity). Therein an implementation of the localization processor of Albani et al. (1996) is combined with algorithm ELT, which is widely based on the algorithmic details as given in Kollmeier and Koch (1994).[13] Essentially, Boone et al. (2010) applied the localizer to define the aperture and

[13] At the time of the development of this algorithm, we were not aware of the implications the algorithmic details in (Kollmeier and Koch, 1994), which are discussed and revised in Chapter 2.4.3. The algorithm was nevertheless able to increase speech intelligibility (i.e., the Better Ear I3,

subsequently accessed clean binaural reference parameters to perform a directional filtering. In effect, the approach successfully mitigates the loss in speech quality at a given intelligibility gain, based on the scene adaptive procedure that adjusted the aperture of the binaural algorithm to the interference. Nevertheless, binaural parameters obtained by this clean reference approach are not optimal.
The approach taken here constitutes a leap in terms of simplicity and efficiency by matching the binaural differences with a real-world reference of these parameters in a particular situation. The questions that arise with respect to a future application in hearing aids are: first, will it be possible to classify scenes sufficiently accurately in order to determine which reference maps to use, and, secondly, does a particular binaural speech processor adapt to scenes if no perfect matching between a real-world scene and a reference lookup table exists. The second question is addressed in Chapter 5.

In the following, the probabilistic pattern-driven source separation approach is examined by means of sound scene examples. Therefore, we first analyze the classification approach with algorithm CLP, i.e. the binaural carrier-based speech processor. The algorithmic details have been given in Table 2.2. An analysis of binaural differences is performed at 257 DFT bins. To that end, the fine-structure IPD and ILD are calculated and, as a result, the feature vector $\vec{\Delta}$ accounts for $2d$, i.e. 514 variables. Thus, for the two binaural parameters at their respective DFT bins, a priori histograms have to be generated. Consequently, a priori PDFs are constructed with an empirical approximation, by measuring the time series of the short-time binaural parameters in a particular sound scene. Specifically, the time sample values $\vec{\Delta}$ of each DFT coefficient d is binned to the PDFs (i.e. $\mathrm{PDF_t}$ and $\mathrm{PDF_a}$) with 500 bins in the range of $\pm\pi$ for the fine-structure IPD parameter and with 500 bins in the range of ± 40 dB for the fine-structure ILD parameter.

Prior to the estimation process, connected discourse with a length of two minutes was composed of concatenated utterances, separately spoken by both sexes. Male speakers comprised two thirds of the speech material. The sentence material of the Dutch language is taken from the Multilingual TNO Human Factors database (TNO, 2000). The mixtures were composed in twelve spatial configurations. While the target was fixed at 0 deg in all mixtures, the interfering speaker alternately resided at -90, -50, -30, -20, -10, -5, 5, 10, 20, 30, 50 and 90 deg, and the global SNR was set to 0 dB.
When a lookup table in a particular background, e.g. the canteen ambiance, was generated, a mix with one coherent interferer was generated, and the global SNR was set to 0 dB between the scene and the interferer. In this fashion, twelve augmented scenes were generated and subsequently mixed with the target speaker. That way, we aimed to decrease moderately the diffusion of binaural parameters, which would

see Chapter 4), mainly in sustained vowel-sections, which corroborates our assumption of the dominance of the statistical penalty weight in this setting (see Chapter 2.4.3). In informal listening, we considered the algorithmic distortions to be lower than found with the implementation of algorithm ELT, proposed in the present work.

result from a plain mix with a background scene, while keeping the global SNR between the total interference and the target signal at 0 dB.

Although it is as an artificial concept, we continued to determine the SNR at the ear-level, including the linear average between the ears. In view of the findings in the first parts of this chapter, the rationale behind this approach is given by the strong correlation between the statistics of binaural parameters and the SNR at the ear-level. This finding, furthermore, gave rise to the assessment of the binaural speech processors using the ear-level SNR, later in the present work. In spite of the fact that this approach excludes the directional level differences of beamforming front-ends, it allows for a better comparison with omni-directional receivers and between CASA post-processors.
Notwithstanding the fact that the idea of directional level independence has no basis in reality, we equalize this property to isolate characteristics that would otherwise not be easily identifiable. Hearing aids using the approach proposed here should, of course, reflect the directional level dependence of the binaural front-end in the a priori PDFs.

In order to identify binaural cues of the target in the presence of an interferer, Harding et al. (2005) proposed the application of an IBM, as introduced in Equation (2.3.15), which, in the following application, includes and excludes STFT bins like a logical operator prior to the binning process. IBMs were preferred over soft-masks, as they weight the target signal close to 0 dB at the local time-frequency SNR equally strong as high energy portions. Thereby the convolution-like mixtures in the binaural domain are unravelled in a ground-truth based fashion, i.e. a pattern, instead of assigning soft-mask weights and, hence, emphasizing sparsely scattered bins of dominant target speech, which will be tending towards binaural parameter values of free-field conditions. For the utilization of algorithm CLP, the local criterion ε of Equation (2.3.15) has been set to an SNR of 0 dB. The global mixing SNR is set to an SNR of 0 dB. This choice should lead to an approximate equal energy distribution for the target speaker and one interfering speaker in anechoic conditions. However, using the global mixing SNR, much less bins will be labeled with the value one in, for example, the canteen situation (cf. Figure 2.6).

Figure 3.13 shows the a priori histograms of the fine-structure IPD parameter (letter A and B denote the $\mathrm{PDF_t}$ and $\mathrm{PDF_a}$, respectively) and the resulting probability lookup histogram (letter C), for which a certain histogram threshold ϵ_{hist} (cf. Table 2.2) is chosen to prevent a division by zero and to adjust the aperture of the binaural filter. This particular threshold value is held constant in the subsequent illustration of the lookup PDFs of the algorithms CLP and ELT in the Figures 3.13 and 3.15. At a later point in this work, the threshold ϵ_{hist} is optimized to achieve a maximum speech intelligibility in a particular situation.
Returning to the upper plots in Figure 3.13, the anechoic mixture PDFs expose a symmetrical pattern with respect to the midline for both the Aachen head and the HG (low directivity). The traces are well resolved in those histograms, implying

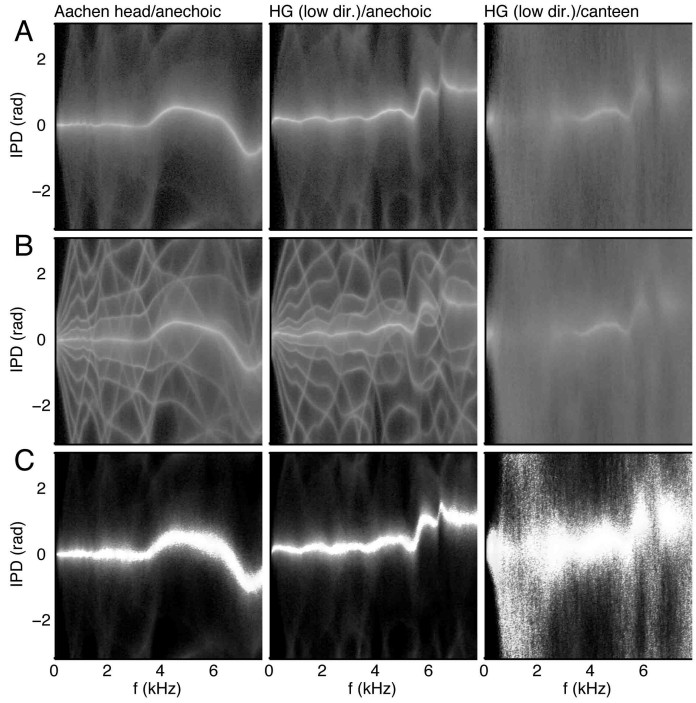

Figure 3.13: *PDFs and weighting lookup tables of the fine-structure IPD as a function of centre frequency, which are used in algorithm CLP for pattern-based source segregation. The letters* A, B *and* C *left to the subplots denote the* PDF_{t}, PDF_{a} *and the lookup table, respectively. Values close to zero are coded in black, whereas higher probability or weighting values are coded in white.*

a strong disjointness in the time-frequency plane. Darker shades of grey indicate mixtures at the time-frequency bin level. Hence, the probability lookup tables in the lower row are not zero off the midline.

Lookup tables of IPD target probabilities may facilitate the usage of IPD cues beyond the spatial Nyquist limit in controlled acoustic setups. However, natural head movements, as they normally occur when facing a target speaker, will probably lead to distortions due to small wavelengths when using the IPD weighting at higher frequencies.

The right-hand side column of Figure 3.13 shows the PDFs in the (augmented, see above) canteen situation. As can be seen, the pattern of traces dissolves, only the midline response is vaguely maintained, resulting in a fuzzy midline weighting.

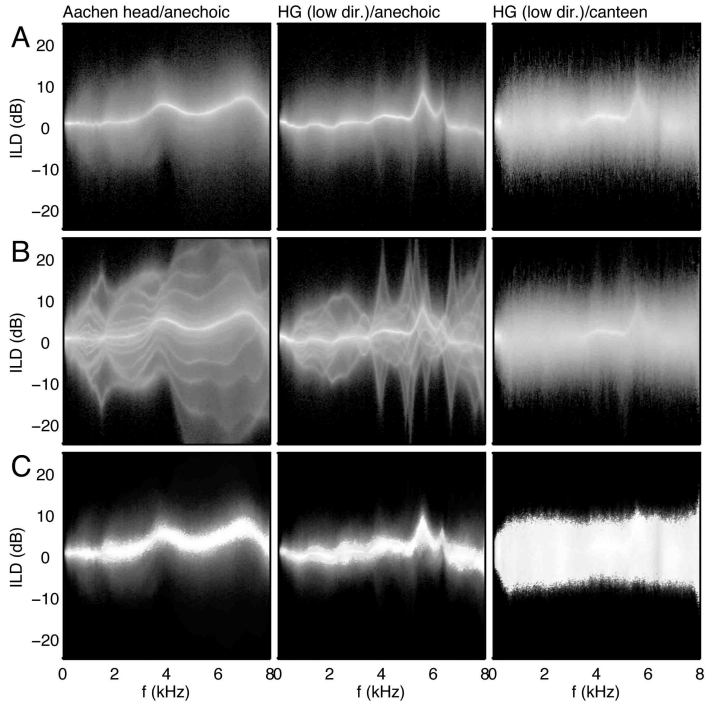

Figure 3.14: *PDFs and weighting lookup tables of the fine-structure ILD as a function of centre frequency, which are used in algorithm CLP for pattern-based source segregation. The letters* A, B *and* C *left to the subplots denote the* PDF_t, PDF_a *and the lookup table, respectively. Values close to zero are coded in black, whereas higher probability or weighting values are coded in white.*

Although the last result is clearly demonstrating the limitation imposed by the fine-structure IPD weighting in unfavourable situations, it also represents the best result that can be gained in this separation process with the help of the here applied probabilistic approach.[14]

Figure 3.14 presents the PDFs and probability lookup table results for the fine-structure ILD. The outliers of the interaural transfer function of the HG (low directivity) above 4 kHz expose a strong difference with the Aachen head. In spite

[14]It has to be considered that the composition of the a priori distributions followed a set of heuristic rules. Their influence on the optimality of the classification approach has not been fully tested yet.

of that, the a posteriori lookup table (Row C) accounts for this characteristic and shows a rather narrow midline weighting. In the canteen situation, the right-hand side column of Figure 3.14, the ILD-based weighting can only suppress binaural level differences of great magnitude, which is expected from the statistical analysis of interaural cues of the fine-structure, presented earlier.

In the following, the probabilistic weighting method for noise suppression is analyzed in view of utilization in algorithm ELT. As it has been introduced, the main difference to algorithm CLP lies in the binaural rendering of the classification and weighting method in a joint carrier and modulation domain. Therefore, based on the algorithmic details in Table 2.3, vector $\vec{\Delta}$ in Equation (3.3.9) contains (basically, see below) for each binaural envelope cue, combinations of 64 carrier and 27 modulation frequency bins. This results into 3456 variables, or features, for which the a posteriori probability is calculated during the weighting process every 8 ms.

Although the implementation of the classification method is conceptually similar to algorithm CLP, it differs in a couple of details. The a priori PDFs have only 150 bins and span the range of ± 50 dB and ± 2 ms for the envelope ILD and the envelope ITD, respectively. The details of algorithm ELT are conform the details given in Table 2.3, including the binaural level magnification approach described above in this chapter. While we provide for this ILD increase with a wider PDF range, the coarser PDF increments are to reduce the computational cost. Furthermore, the local criterion ε in the IBM process is adjusted to an SNR of 5 dB. By preliminary inspection, we discovered an improved suppression of interference with this setting.

Figure 3.15 shows the PDFs of Equation (3.3.9) at three different modulation frequency bins, corresponding to 31, 141 and 297 Hz. In comparison to the PDFs of algorithm CLP, less discrimination of the binaural parameter traces is found for the envelope ILD as well as for the envelope ITD parameter. The PDFs of the envelope ILD cue are rather compact in the anechoic situation, whereas the PDFs of the envelope ITD cue show a high standard deviation, as it has been previously found. Nonetheless, compact midline weighting functions are generated in anechoic interference conditions with both parameters.
A breakdown of the ITD-based midline weighting is observed in the (augmented, see above) canteen situation. The PDF counts of the envelope ITD are at these frequency bins below the applied threshold value ϵ_{hist}, which was introduced to prevent a division by a small number in Equation (3.3.9). During the optimization of this parameter, shown in Chapter 5.2, the threshold is adjusted over a range that allows for the prevention of such a breakdown of the cue-based weighting.
Another drawback of the ITD-based weighting is seen in the plots of the anechoic situations at the low modulation frequency of 31 Hz. At this frequency the envelope ITD parameter shows no lateralization for sources from the side. To repair this, it has been chosen to apply the envelope ITD weighting only above the DFT bin $m_{\text{xo}} = 6$, which is associated with a modulation band centred at a frequency of 78 Hz (see Chapter 2.4.3).

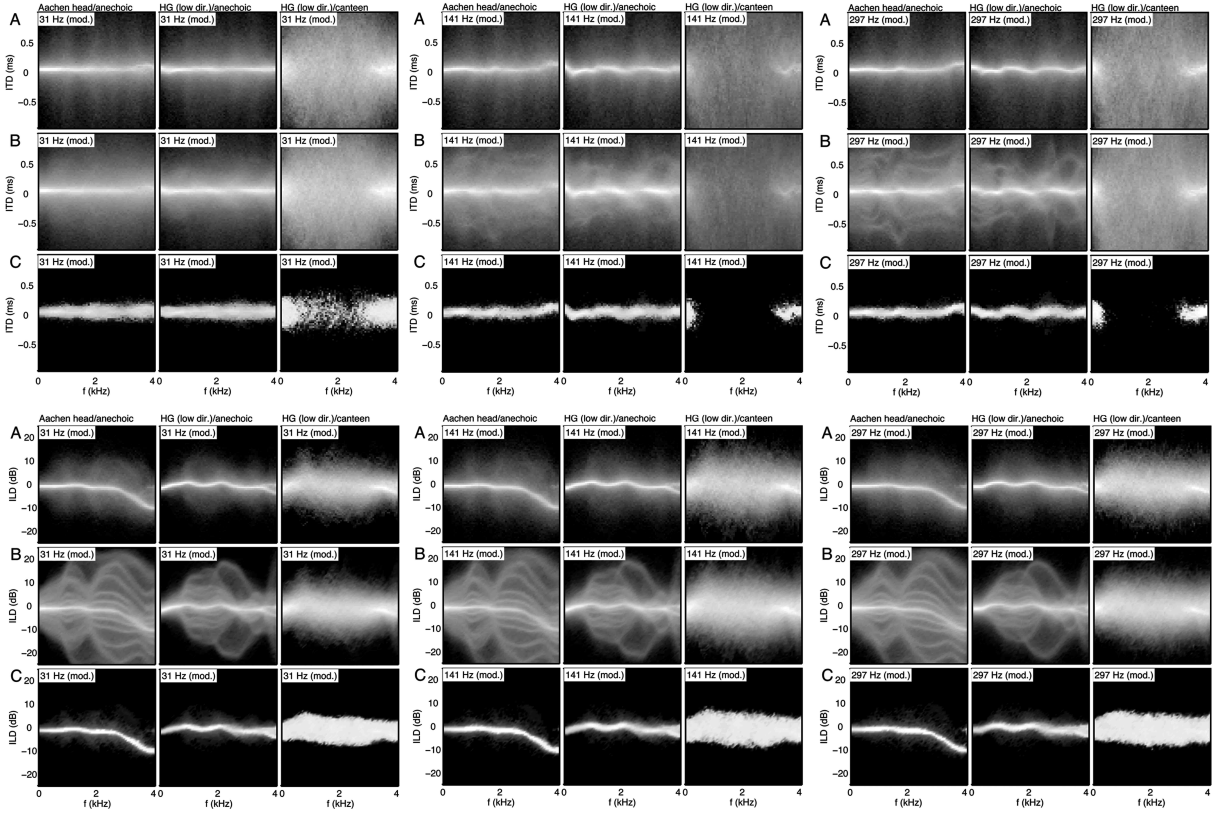

Figure 3.15: *PDFs and weighting lookup tables of the envelope-based ITD (upper plots) and ILD (lower plots) as a function of carrier and modulation (*mod*) frequency. Therein the 3 by 3 plots give a subset of the PDFs and weighting functions at 27 different modulation frequencies up to 400 Hz. The letters* A*,* B *and* C *left to the subplots denote the* PDF_t*,* PDF_a *and the lookup table, respectively. Values close to zero are coded in black, whereas higher probability or weighting values are coded in white.*

Conclusion

In this chapter the statistics of binaural parameters of the fine-structure and the envelope have been studied over a wide range of acoustical conditions. In this context, the binaural fine-structure parameters of different beamforming front-ends of hearing aids have been compared to an omnidirectional hearing aid under free-field conditions. In addition, the similarities and dissimilarities with the model-based binaural parameters of an artificial head have been identified throughout all presented noise conditions.

Our findings are supported by psycho-acoustical studies, in which binaural cues and thereby localization were found to be strongly affected in the event of interference. The physiologically motivated decomposition of the envelope in modulation spectra and a subsequent binaural comparison has not delivered the breakthrough in the problem of reliably isolating speech from noise, in particular in diffuse interference. As a temporal interaural disparity cue, the IPD of the fine-structure has shown to be more robust in diffuse noise, yet is by no means a sufficient indicator of direction in these conditions. In incoherent noise conditions, both the fine-structure and the envelope-based ILD parameter should gain importance during the classification by possessing statistical properties that are less affected. Nevertheless, as it has been shown, even optimal classification reduces to a coarse differentiation between target and noise under diffuse conditions.
Overall, the statistical study explains why binaural speech processors that employ binaural parameters for noise estimation generally fail in diffuse sound fields.

In order to achieve the best possible application of binaural parameters in changing acoustics, a Bayesian classification approach has been applied to their a priori distributions in different noise conditions. The method allows for the establishment of soft-masks based on the probability that a certain analyzed binaural parameter was caused by the target source. To that effect, the approach replaces former heuristic strategies of soft-mask formation by a methodical approach that accounts in an optimal and elegant way for the utilization of a pattern-based source separation at the output of a particular hearing aid.

So far the algorithmic principles of the binaural speech processors of the present work have been defined. Prior to the application of the processors in speech intelligibility enhancement, an optimization of their algorithmic parameters is necessary. To face this challenge with appropriate means, an objective measure of speech intelligibility for binaurally and nonlinearly processed speech needs to be specified. The following chapter summarizes our efforts in finding such an objective measure. Chapter 5 subsequently returns to the optimization and assessment of the binaural speech processors.

4

The instrumental evaluation of speech intelligibility

Together with the, here ignored, amplification and compression tasks in hearing aids, there exists a multitude of parameters and boundary conditions whose complex interdependency and effect on speech intelligibility exceeds the manageability of the development and optimization of binaural speech processors. Therefore, the instrumental evaluation of speech intelligibility is a requirement for the development of binaural speech processors. Chapter 2.4 introduced three binaural CASA filters. Each algorithm contains a set of parameters, which have to be defined for a range of acoustical surroundings, in a manual or algorithmic optimization procedure. In the longer term, individual hearing thresholds have to be included, to tailor CASA techniques properly for the hearing impaired.
The current chapter summarizes our efforts in finding an algorithmic measure that calculates speech intelligibility at the output of binaural nonlinear speech processors. The chapter is divided into two parts. The first part presents and evaluates a binaural and speech-based Speech Transmission Index (STI), which describes the binaural advantage, but fails in the prediction of the effect of nonlinear noise suppression on speech intelligibility. The second part of this chapter is dedicated to this particular problem of nonlinear speech intelligibility enhancement. Finally, a coherence-based and level-weighted Speech Intelligibility Index (SII) is defined as a better ear measure, i.e. taking the ear offering the best intelligibility per critical band, and evaluated to guide the optimization processes and assessment tasks in the remainder of this work.

4.1 A speech-based and binaural Speech Transmission Index[1]

A speech-based and binaural STI is presented and evaluated in a variety of acoustical complexities and spatial conditions. The proposed method facilitates the assessment of speech intelligibility in classical room acoustics and electro-acoustics by simply comparing a binaural speech recording in unfavourable conditions with its clean original. Both the binaural processing stage and the speech-based STI method have effective and computationally fast realizations. The central part of the binaural processor forms a cross-correlation stage that is designed to replicate psycho-acoustic data of binaural interaction. Supplemented with the head shadow effect, which is generated in a better-ear fashion, a fair amount of the binaural advantage in speech intelligibility is modeled.
An evaluation of the method was performed in a suite of listening tests. These tests incorporated different disturbances, such as stationary noise and fluctuating noise, a set of nonlinear signal alterations, including a CASA post-filter, a multitude of spatial configurations with different room acoustics, and with up to four interferers. As a result, the objective method offers a stable prediction of the subjective results in binaural speech intelligibility under most of the linear disturbances. In spite of this, the full amount of the binaural advantage is not achieved by the current implementation of the method, which suggests further research.

4.1.1 Introduction

Fair speech intelligibility (SI) can be considered as the main acoustical requirement in enclosed spaces. There are several acoustical measures that predict SI in rooms. Most used in practice is the reverberation time, RT. Although robust in an isotropic sound field, RT cannot be directly linked to SI. A measure that was developed for this purpose is the energy measure Definition, D50. It is based on the room impulse response, and it is calculated as the ratio of the direct sound and supporting early reflections, arriving within 50 ms after the direct sound, and the late impairing reverberant sound that follows after 50 ms. The D50 measure is prone to fluctuations due to the early, anisotropic part of the sound field. Caused by the interference of early reflections, D50 has shown to fluctuate by a factor of two when altering the recording position only slightly (De Vries et al., 2001).
A robust measure of SI is the Speech Transmission Index (STI). The STI calculates the reduction in modulation depth of a signal that is sent over a channel, e.g. a room (Steeneken and Houtgast, 1980). In the classical approach, the STI uses a set of ar-

[1] Most parts of this chapter were already published in *Schlesinger, A., Ramirez, J.-P., Van Dorp-Schuitman, J. and Boone, M. M., "Report on a binaural extension of the Speech Transmission Index method for nonlinear systems and narrowband interference", International Symposium on Auditory and Audiological Research, 2009, Marienlyst, Denmark* and in *Schlesinger, A., Ramirez, J.-P. and Boone, M. M., "Evaluation of a speech-based and binaural Speech Transmission Index", Proceedings of the AES 40th Conference on Spatial Audio, 2010, Tokyo, Japan.*

tificially modulated noises as stimuli. Recent methods of the STI are speech-based. These measures are partly capable of taking nonlinear disturbances, the influence of fluctuating noise and the influence of speaking style on a phoneme-level into account (Goldsworthy and Greenberg, 2004; Payton et al., 2002; Payton and Shrestha, 2008). Thereby the general speech-based calculation method simply compares the clean signal with the degraded signal. For this reason, it is an intrusive measure. A survey on current STI techniques has been given by Goldsworthy and Greenberg (2004).

Predicting SI monaurally lacks the strong impact of binaural hearing on intelligibility. In spatial configurations, as e.g. S_0N_{120} (the abbreviation stands for a speaker S at 0 deg and a noise source N at 120 deg), the binaural advantage of unmasking the target signal can be as high as 12 dB at the 50% intelligibility level, in continuous noise and in anechoic conditions (Bronkhorst, 2000). Two binaural effects contribute to the advantage. These are the binaural interaction process, which is based on temporal difference cues, and the head shadow effect, that increases the SNR at the contralateral ear with respect to the noise source.
The gain obtained from binaural unmasking, generally known as the release from (spatial) masking, is basically independent of the speech material and the long-term spectrum of the masker. However, the advantage is somewhat diminished in conditions of interfering speech and speech modulated maskers as compared to continuous masking noise (Bronkhorst, 2000).
The contributions of the binaural interaction process and the head shadow effect were found not to be additive. Additionally, a tradeoff as a function of frequency between these two effects is observed. At low frequencies, the main portion of binaural unmasking is caused by binaural interaction and is in the order of 7 dB. The lower portion that amounts to the total benefit is allocated to the head shadow effect. At higher frequencies, the ratio inverts (Bronkhorst, 2000).
In real-life conditions of reverberation and diffuse background noise, the binaural advantage diminishes. However, tests in reverberation revealed a small but significant binaural advantage that, in comparison with anechoic conditions, is mainly determined by the speaker and masker distances (Bronkhorst, 2000).

The two leading models that answer the question of how the auditory system performs binaural interaction, are the coincidence model by Jeffress (1948), which can be formulated as a cross-correlation process, and the Equalization-Cancelation (EC) model by Durlach (1960). Both approaches can be mathematically related (Verhey, 2008) and have shown to capture most of the psycho-acoustic effects that are associated with the binaural interaction process.
In recent years these models, with variations, have been incorporated in techniques of SI prediction. Widespread attention found the SI prediction method by Beutelmann and Brand (2006), who combined the EC model with the Speech Intelligibility Index (SII), a spectral SI measure. Recently, their model was revised and newly evaluated in a variety of acoustical situations (Beutelmann et al., 2010). In order to predict SI in reverberation, a disturbance that rather acts on the time evolution of the wave-

form than on the spectrum, the EC stage has been also linked to an extended SII, which incorporates the modulation transfer function (Rennies et al., 2010).
Another method of binaural SI prediction was developed by Van Wijngaarden and Drullman (2008), who combined a cross-correlation stage for the binaural interaction process (along with the head shadow processing) with the classical STI method. Their method demonstrated a high accuracy in SI prediction in rooms, the realm of the classical STI.

When it comes to the effect of nonlinear distortions on SI, neither of these methods will be successful. The SII method, as implemented in Beutelmann and Brand (2006), requires the speech and the noise to be separated after the EC stage, in order to calculate the SNR. The classical STI method suffers from the problem of intermodulations, either in nonlinear channels or due to modulated interferers, and the very general stimulus approach that suffices for the SI prediction in room acoustics but is not adequate in speech processors. In principle, a speech-based version of the STI offers a solution to these problems. Based on the comparison of the deteriorated speech with its original, intermodulations will not have a detrimental effect on the prediction quality of the measure. However, as it turned out, the determination of nonlinear distortions is not a trivial problem (Christiansen et al., 2010; Taal et al., 2010; Schlesinger and Boone, 2010). Although much work has been devoted to extend the STI method to the measurement of nonlinear distortions (Ludvigsen et al., 1990; Goldsworthy and Greenberg, 2004), we believe that the success will remain limited, as long as mere signal-based approaches are pursued which do not include a top-down context constitutive weighting on a phoneme or sub-phoneme level (see the following Subsection 4.2).
Our position is based on a comparative analysis[2] of a speech-based STI using the revised envelope regression method of Goldsworthy and Greenberg (2004), a coherence-based SII (Kates and Arehart, 2005b) and the Spectro-Temporal Modulation Index (STMI), similar to Elhilali et al. (2003), but using the revised envelope regression method of Goldsworthy and Greenberg (2004) for calculating the similarity between the spectro-temporal response fields (STRFs) of the clean signal and the STRFs of the degraded signal (Schlesinger and Boone, 2010).[3] The outcomes show for none of these objective measures a single functional relationship between subjective and objective results for a collection of linear and nonlinear envelope-thresholding distortions.[4] As the STMI is based on a high-level model of the STRFs in the auditory cortex, our findings support the need for the inclusion of hypothesis driven processes when modeling SI at the output of mask-based, i.e. nonlinear, speech enhancement processors.

Preliminary to these findings, the speech-based STI in the revised envelope-regression

[2]See Appendix C for a reprint of the results.
[3]The Neural Systems Laboratoy toolbox (URL *http://www.isr.umd.edu/Labs/NSL/*) was used for this purpose.
[4]Envelope-thresholding, also termed centre-clipping, is a nonlinear distortion of speech enhancement processors using a varying gain function. See Appendix D.

version of Goldsworthy and Greenberg (2004) originally had been chosen by the author for its computational efficiency, its quality in a multitude of linear distortions (including reverberation) and the (at that time anticipated) possibility for the prediction of nonlinear mask-based distortions. In the implementation presented here, the speech-based STI is linked to a binaural processing stage, which is based on the coincidence model of Jeffress (1948). Although Van de Par et al. (2001) showed that the EC model is physiologically more plausible, the coincidence model is more practical when implemented as an efficient cross-correlation process. In order to assess narrow-band distortions and to prepare for the incorporation of elevated thresholds of the hearing impaired (Holube and Kollmeier, 1996), a Gammatone filterbank with a filter channel density of approximately one on the ERB scale is used for the peripheral frequency analysis.

The aim of the following sections is to give a more elaborate introduction and evaluation of the speech-based and binaural STI. Thereby the evaluation relates not only to the binaural processing of the method, it also accounts for the analysis in a multitude of nonlinear disturbances, i.e. peak clipping, envelope thresholding and phase jitter. We will also consider a binaural hearing aid algorithm for speech enhancement and different kinds of maskers, e.g. a fluctuating masker.
We continue with the introduction of the algorithm. Subsequently, the binaural STI method is evaluated, discussed and conclusions are drawn.

4.1.2 Algorithm

The fundamental processing structure of the algorithm is shown in Figure 4.1. The inputs to the algorithm are the binaural, time aligned clean and deteriorated speech samples, which are analyzed in blocks. The central part of the algorithm consists of the linkage of a binaural processor and a speech-based STI envelope-regression method. Hence, the envelope regression method analyses the modulation depth across the internal binaural representation of the input and selects the trace as a function of the interaural delay that offers the highest modulation depth. In the frequency range of the head shadow effect, the processing is effectively reduced to the comparison and the maximization of the modulation depth at both ears. The proposed method can be subdivided in the following steps.

(1) The applied binaural speech material is low-pass filtered at 9.5 kHz and sampled at 22.05 kHz. Silent gaps, defined as the level of -50 dB in frames of 10 ms, are determined in the clean signal with a VAD procedure, and are subsequently discarded in the clean and the degraded signal at equal time positions.

(2) A peripheral frequency analysis is performed with a Gammatone filter bank of 4th order using 30 ERB bands with centre frequencies ranging approximately

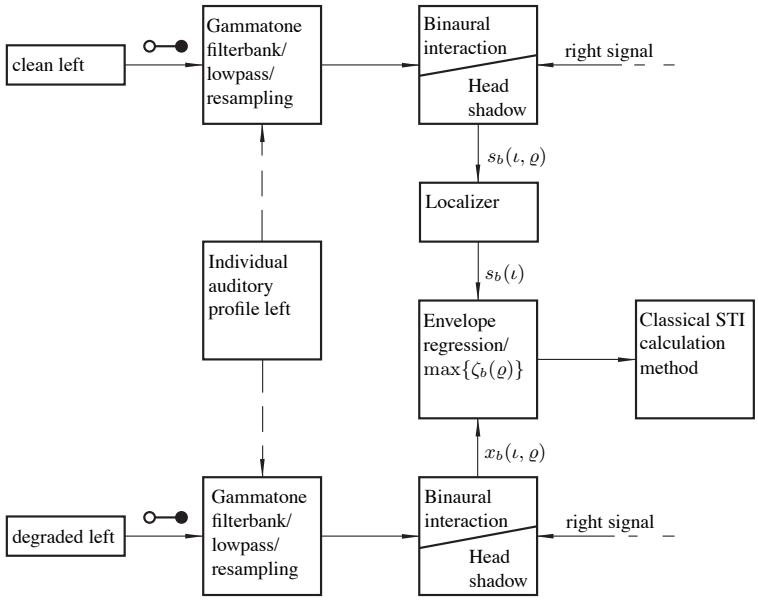

Figure 4.1: *Block diagram of the speech-based and binaural STI method. The dashed lines at the left-hand peripheral stage indicate that no hearing thresholds are modeled in the current implementation.*

logarithmically from 0.1 to 8 kHz on the linear frequency axis, i.e., linearly on the ERB scale. The implementation of the filters is taken from Wang and Brown (2006). No middle ear filter model is included in the calculation.

(3) A Hamming window of 10 ms length is convolved with the squared output of the Gammatone filter bank to yield the intensity envelope below 50 Hz.

(4) Prior to the binaural processing, the intensity envelope is downsampled by the factor of 7 and subsequently partitioned into samples of 30 ms length using the cross-correlation and auto-correlation in the binaural processing stage (see below). The windows have an overlap of 75 %, which results in a sampling frequency of 132 Hz.

(5) The central binaural processing stage was motivated by the work of Van Wijngaarden and Drullman (2008). Therein, the binaural advantage is split into the contribution from the binaural interaction processing, which is calculated between 0.5 to 2 kHz by the cross-correlation of the binaural signals at increments of 0.1 ms between -0.8 and 0.8 ms (ϱ), and the contribution from the head shadow effect, realized in a better ear fashion below 0.5 kHz

and above 2 kHz. Since the mean square value of the cross-correlation and the auto-correlation (the latter is performed to yield a common sampling for the two binaural effects) is taken to sample the time evolution of internal binaural envelope, the square root has to be applied to the output of the binaural processing stage in order to recover the intensity representation. The choice of $\varrho = 0.1$ ms is in the range of the model-based coincidence intervals of approximately 50 to 150 μs (Blauert, 1997, p. 342.).

(6) The location of the target source is calculated every 0.25 s from the clean binaural input through a band-wise maximum search across the binaural correlation at all increments of ϱ. In the frequency range of the head shadow effect, the direction is inferred from the intensity maximum in the same time interval.

(7) As a modulation metric, the stochastic reformulation of the envelope regression method of Goldsworthy and Greenberg (2004) is chosen. Assuming $s_b(\iota)$ to be the binaural clean and $x_b(\iota)$ the binaural contaminated intensity envelope of 0.5 s length, where ι is a discrete time sample and ω is the short-time window index, which is dropped for notational convenience in the following. Subsequently, the modulation ζ is calculated in each band b and for a set of lateral differences (ϱ) through:

$$\zeta_b(\varrho) = \frac{\mu_{sb}}{\mu_{sb} + \mu_{zb}} \cdot \frac{\mathrm{E}\{(s_b(\iota) - \mu_{sb})(x_b(\iota, \varrho) - \mu_{xb})\}}{\mathrm{E}\{(s_b(\iota) - \mu_{sb})^2\}}, \tag{4.1.1}$$

where μ_{sb}, μ_{xb} and μ_{zb} are the intensity means, and $z_b(\iota) = |x_b(\iota, \varrho) - s_b(\iota)|$. The Equation is a linear regression of the contamined envelope onto the clean envelope. It is based on an MMSE criterion and uses estimates of the means and variances. The stochastic formulation makes the method preferable over other speech-based methods, as it relies only on running averages in windows of short duration. Goldsworthy and Greenberg (2004) expanded the normalization fraction in Equation (4.1.1) to account for nonlinear operations—in particular for cases when the modulation depth is abnormally increased.
$\zeta(\varrho)$ is calculated in analysis frames with index ω of 0.5 s, which overlap 50 %. The threshold of perceptible changes is therefore 250 ms, which is about 50 to 150 ms higher than found in psycho-acoustic experiments (Blauert, 1997, p. 323). For each window, the maximal modulation is calculated by searching for the maximum modulation depth across the increments:

$$\zeta_b = \mathrm{argmax}_{\varrho \in [-0.8\ \mathrm{ms}, 0.8\ \mathrm{ms}]}\{\zeta_b(\varrho)\}. \tag{4.1.2}$$

The better-ear effect is calculated similarly, by choosing the ear offering the highest modulation depth.

(8) Subsequently, the band-wise modulation metric is related to the apparent SNR in bands:

$$\mathrm{aSNR}_b = 10 \log_{10} \frac{\zeta_b}{1 - \zeta_b}. \tag{4.1.3}$$

The band-wise results are further processed according to the classical STI method (Steeneken and Houtgast, 1980), i.e. clipped below -15 dB and above 15 dB, transformed to the transmission indices, which are weighted with an adapted band importance function of Pavlovic (1987) for average speech, and summed to the binaural STI. Finally, the STI values are averaged across the analysis frames ω.

The challenge in designing the binaural stage above revolves around finding an optimal sampling and frame length ω that allows for an accurate regression-based modulation transfer calculation from cross-correlation based mean square values. To demonstrate that the quality is equivalent to the calculation of the modulation-transfer of the unprocessed intensity envelopes, both methods are compared to the theoretical RMS based modulation-transfer:

$$\hat{\zeta}_{\omega,b} = \left(1 + 10^{\frac{-\mathrm{SNR}_{\omega,b}}{10}}\right)^{-1}. \tag{4.1.4}$$

Therefore 50 sentences of the Semantically Unpredictable Sentence (SUS) test corpus of Ramirez et al. (2009) were used and the $\mathrm{SNR}_{\omega,b}$ was set to 0 dB. A continuous masker was generated with the long-term speech spectrum of the SUS corpus as the noise signal. Standard envelope regression analysis was performed on windows of 0.5 s with intensity envelopes sampled at 150 Hz. The modulation-transfer at the output of the binaural stage was simulated in a simplified manner with an auto-correlation computation (and a subsequent calculation of the square root, see above), using the parameters of item 4 in the list above.

Figure 4.2 shows the results of a regression analysis with the r^2 measure, indicating the amount of variance that is modeled by the envelope regression method. Good r^2 values are observed from about 500 Hz, where the bands are associated with the maximum perceptual weighting for SI (Pavlovic, 1987; Yoo et al., 2007). This excludes the low frequency channels of moderate r^2 quality and, overall, leads to a high predictive power of the envelope regression based STI, as it was shown by Payton and Shrestha (2008). The regression analysis is subsequently repeated with the envelope regression of the auto-correlated signal. As it is shown in Figure 4.3, the regression analysis of the processed signals is comparable to the previous results. Consequently, it can be concluded that no lowering of the predictive power is introduced by the binaural stage.

Returning to Figure 4.1, the diagram of the STI method also shows the inclusion of the left and right ear hearing threshold. This method offers the possibility to predict SI for the hearing impaired, whose ILD differences are expected to deviate from natural ILDs (Durlach and Colburn, 1978). Alternatively, one could also model the hearing loss at a central level of the STI procedure, i.e. after the binaural stage, similarly to the classical approach (Steeneken and Houtgast, 2002). Holube and Kollmeier (1996) calculated a speech-based STI with a bank of 22 critical filters for hearing impaired people and showed that the provision for the elevated threshold

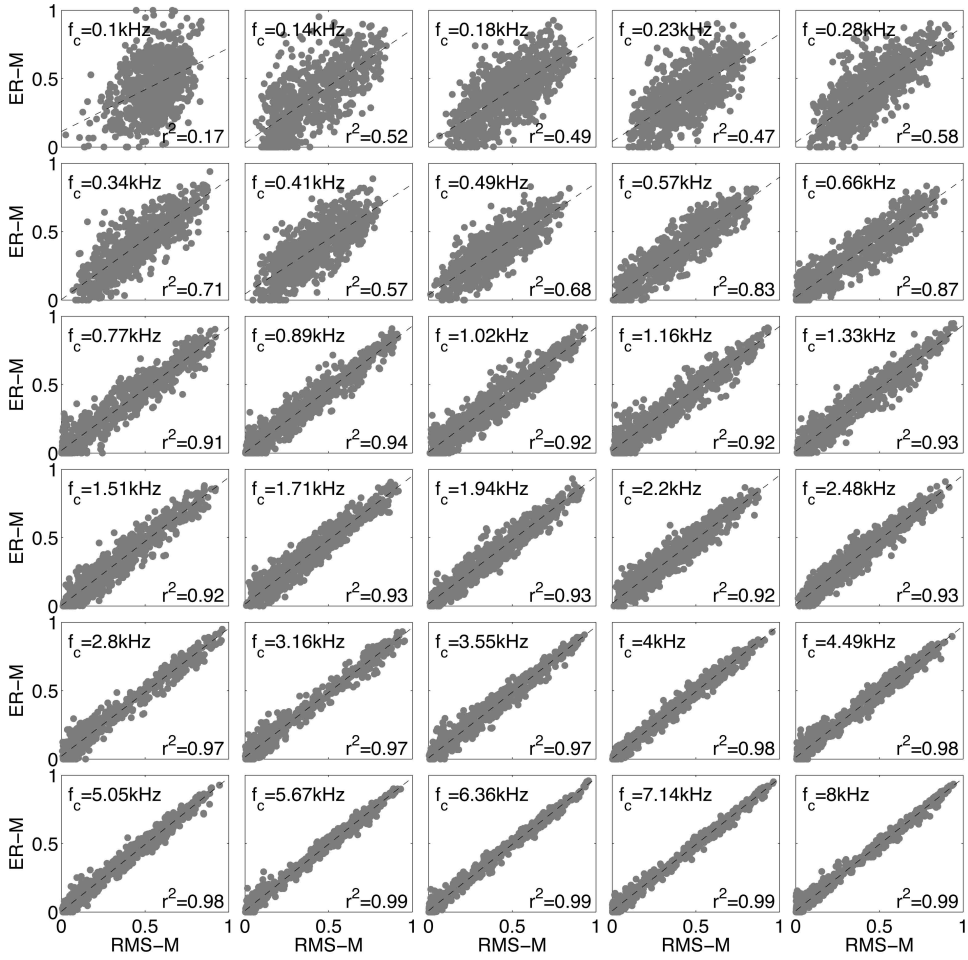

Figure 4.2: *A regression analysis, using the r^2 measure, of the modulation transfer in 30 ERBs, centred at the frequencies given in the upper left-hand side corner of each subplot.* $\mathsf{ER}-\mathsf{M}$ *denotes the envelope regression modulation transfer calculation method of Goldsworthy and Greenberg (2004) and* $\mathsf{RMS}-\mathsf{M}$ *is the theoretical modulation transfer, given in Equation (4.1.4). The window length for the calculation of the short-time modulation metrics was adjusted to 0.5 s.*

leads to good proficiency in modeling the hearing loss, compared to other audiological measures. Elevated hearing thresholds have not been included in the present implementation, since the method has only been evaluated against normal hearing

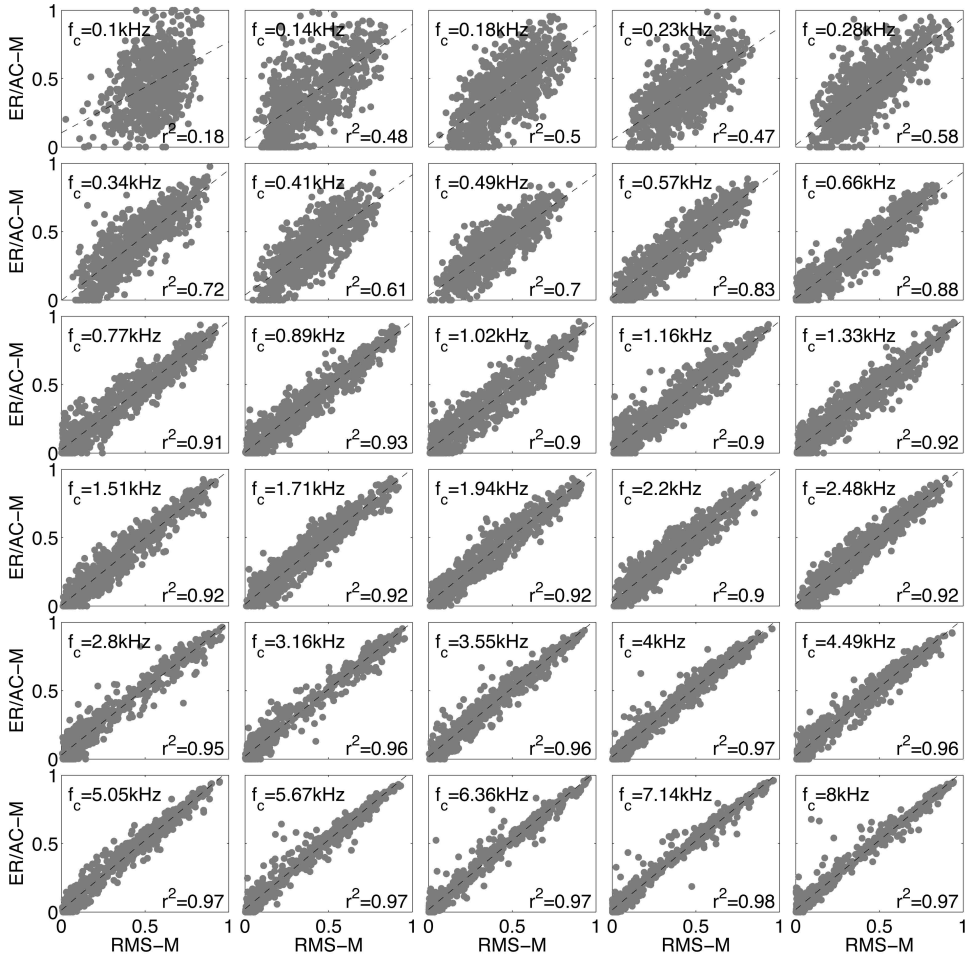

Figure 4.3: *A regression analysis, using the r^2 measure, of the modulation transfer in 30 ERBs, centred at the frequencies given in the upper left-hand side corner of each subplot.* $\mathsf{ER/AC-M}$ *denotes the auto-correlated envelope regression modulation transfer, corresponding to Equation (4.1.1), and* $\mathsf{RMS-M}$ *is the theoretical modulation transfer, given in Equation (4.1.4). The window length for the calculation of the short-time modulation metrics was adjusted to 0.5 s.*

people. The following section gives the results of the evaluation.

4.1.3 Evaluation

The proposed method for a speech-based and binaural STI had been evaluated in an earlier work by the author and colleagues (Schlesinger et al., 2009). In the present work the analysis is extended to a diverse series of monaural and binaural conditions, which were subjectively evaluated in four listening tests. Throughout the evaluation, the algorithmic parameters of the STI method were kept constant to ensure comparability among the conditions and with the results of the previous study of Schlesinger et al. (2009).
The SUS corpus of Ramirez et al. (2009) was used in the subjective tests. The sentences were composed of four main key words in German in a syntactically coherent frame, but with no semantic coherence. Consequently, the predictability of the message to be retrieved by the listener is minimized. Throughout the binaural tests, the SUS samples were convolved with the HRTFs of an artificial head (Schmitz, 1995) and corrected for the headphones (type Sennheister HMD 46-3-6) that were used in the listening test. The sampling frequency of the SUS set is 44.1 kHz and the presentation level was adjusted to 70 dB (A) SPL. The recordings were stored for the analysis with the proposed STI method. Four listening tests were performed to draw a comprehensive picture of the proposed STI method.

(1) A percent-correct score test was conducted to analyze the operation in constant background noise, fluctuating background noise and on nonlinearly processed speech. Table 4.1 lists the conditions. Eight students of normal hearing with a hearing threshold below 15 dB HL took part in the test. The monaural stimuli were presented to the right ear.

Figure 4.4 (A-F) gives the results of the test. The conditions cover up to about 90 % of SI. In order to link all tests presented with their different conditions to a common reference, a second order polynomial is fitted to the linear disturbances (condition 1 to 26) in Figure 4.4 A. The verification of a narrow distribution of monaural and diotic conditions along a common SUS word score - STI curve is a feasible approach to evaluate the quality of a binaural STI method (Van Wijngaarden and Drullman, 2008).[5] As a figure of merit of the polynomial fit, the r^2 measure indicates a fair predictability for the STI method in long-term as well as short-term stationary noise and across differing spectral characteristics of the maskers. The r^2 measure could be improved when excluding the fluctuating noise conditions, or when setting the STI transmission indices of fluctuating noise according to a flatter psychometric function. The second approach has been demonstrated by Rhebergen and Versfeld (2005) with the Speech Transmission Index (SII) as a means to assess the effect of fluctuating noise. Herein, we exclude a modification of the transmission indices in favour of a universal measure for different kinds of disturbances.
Figure 4.4 B gives the results of the nonlinear conditions 27 to 41. The envelope

[5] With respect to a monaural word score - STI relation, Van Wijngaarden and Drullman (2008) showed that dichotic conditions coherently shift to lower SI values, if the binaural advantage is not calculated.

Table 4.1: *Conditions of speech distortion as evaluated in the percent-correct score test 1. The abbreviation SSN (i.e., Speech Shaped Noise) refers to the applied noise, which is spectrally shaped with the long-term average speech spectrum of the target speaker. The amplitude modulated noise, i.e. the short-term fluctuating noise, was generated by extracting the speech envelope of arbitrary speech samples from the target speaker and by modulating the long-term spectrum of the target speaker with this envelope. The CASA post-filter is a binaural mask-based speech processor, operating on the waveform fine-structure, similar to algorithm CLP of this work. The abbreviation bw. indicates the bandwidth employed in the stimuli. Wide-band (wb.) represents a frequency range from 0.05 to 7 kHz and narrow-band (nb.) comprises a range from 0.35 to 3.4 kHz.*

cond.	pres.	bw.	noise type	SNR (dB)
1	diotic	wb.		inf
2—10	diotic	wb.	amplitude modulated noise	[-4,-3,-2,-1,0,1,2,4,6]
11—14	diotic	nb.	SSN male identical to speaker	[-5, - 3, -1, 1]
15—18	diotic	wb.	SSN male identical to speaker	[-5, - 3, -1, 1]
19—22	diotic	wb.	SSN female	[-5, - 3, -1, 1]
23—26	mon.	wb.	SSN male identical to speaker	[-5, - 3, -1, 1]
27—31	mon.	wb.	envelope thresholding	[60, 75, 80, 85, 90]
32—34	mon.	wb.	peak clipping	[60, 70, 80]
35—37	mon.	wb.	phase jitter	[0.2, 0.3, 0.4]
38—39	dicho.	wb.	SSN/S_0N_{90}	[-12, -9]
40—41	dicho.	wb.	SSN/S_0N_{90} and CASA post-filter	[-12, -9]

thresholding (also known as centre clipping) and peak clipping conditions were generated as described in Kates and Arehart (2005b), and the phase jitter conditions were simulated as described in Elhilali et al. (2003) (see Appendix D). As can be seen, for all disturbances, the STI method correctly assigns the trend according to the severity of the distortion. While peak clipping conditions are lying close to the common SUS-STI curve, the phase jitter conditions are less close, and envelope thresholding conditions are clearly off the curve due to an overestimation of SI by the STI method. To define the nonlinear processing of a mask-based binaural hearing aid algorithm for SI enhancement, which is conceptually similar to the binaural speech processor of Gaik and Lindemann (1986), two binaural conditions were tested in this setting.[6] Condition 38 and 39 are S_0N_{90} arrangements and feature SNRs of -12 and -9 dB, respectively. Conditions 40 to 41 are the same arrangements but were

[6]The algorithm has been applied with clean binaural reference maps and with a heuristic tuning of the algorithmic parameters. In addition, a heuristically adjusted cepstral smoothing post-filter has been connected downstream to suppress musical noise.

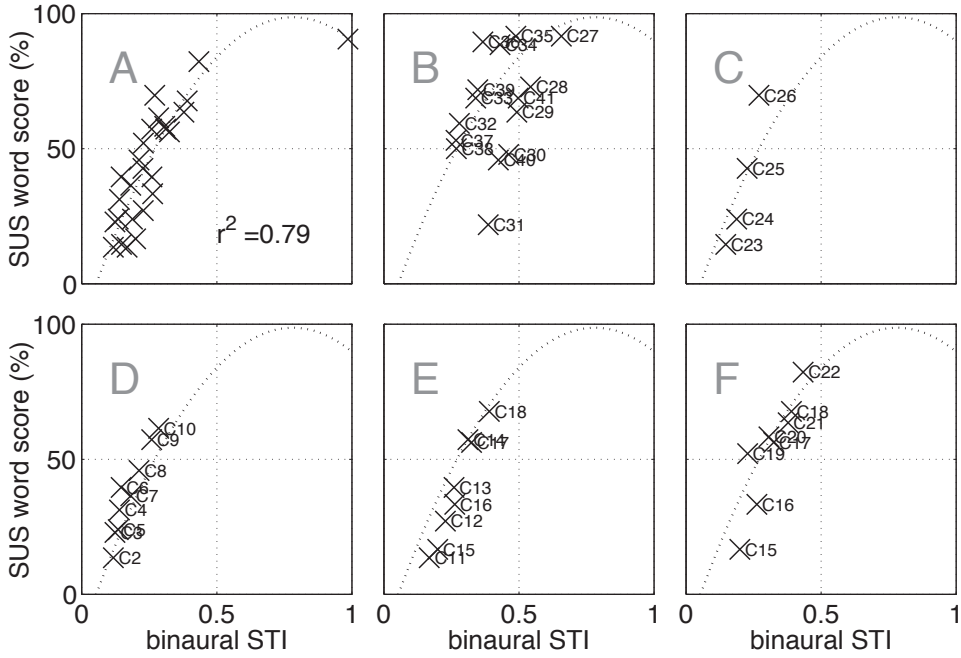

Figure 4.4: *Results of the listening test 1.* A*: linear degradation of monaural and diotic signals and a second order polynomial fit, that serves as a common reference for the quality of the proposed method throughout all SUS-STI evaluations,* B*: nonlinear conditions,* C*: pure monaural conditions,* D*: fluctuating noise conditions,* E*: narrow- and broad-band conditions and* F*: male and female SSN masker conditions. See Table 4.1 for a description of the test conditions.*

processed by the SI enhancement algorithm. Spatialization had been performed by convolving the sources with the respective HRTFs of the Aachen head (see Chapter 3.2).

No improvement, and no considerable decline, of SI is yielded by the binaural processor, probably due to its conservative implementation and the rather difficult conditions that an SSN masker constitutes at very low SNRs. As expected, the SI enhancement is overestimated by the STI method and the nonlinear processing clearly reveals its envelope thresholding nature.

Figure 4.4 C shows how well the monaural conditions 23 to 26 fit the polynomial. The conditions only slightly deviate from the polynomial with an increased slope. In general, the release from binaural masking is smaller than 1 dB for diotic presentations with respect to monaural listening conditions (Bronkhorst, 2000). The difference in slope is mainly due to the fluctuating noise conditions, which result in

an inclination of the polynomial.
Figure 4.4 D shows the quality of the STI method to assess the impact of a fluctuating masker on SI. Despite the afore-mentioned slight underestimation of SI, the conditions lay closely to the polynomial.
A comparison between wide band and narrow band (see Table 4.1 for the information on bandwidths) is shown in Figure 4.4 E. At last, Figure 4.4 F contrasts subjective and objective results for SSN conditions using either the spectrum of male or the spectrum of female speech. As can be seen, all of these conditions scatter narrowly along the polynomial.

(2) In order to evaluate the binaural processing of the proposed STI method, a percent correct-score test with the target speaker and one long-term stationary interferer at mutually permutated positions was conducted. The spatial arrangements consisted of free-field conditions and simulated room-acoustics with the reverberation times of 0.4 and 1.5 s, corresponding to reverberation radii of 8.5 and 4.4 m, respectively. The MISM of Van Dorp Schuitman (2009) was used for the room simulation. The room size was set to 20x30x15 m. The sources were located at a radius of 5 m around the receiver, which was located slightly off the centre of the room and at a height of 2 m. The first and second order reflections were modeled with the MISM technique, whereas higher order reflections were appended to the IR using statistical modeling. The virtual sources were omnidirectional.
Eight students with normal hearing participated in the test. Figure 4.5 gives the results and an explanation of the individual test setups. For each spatial SN configuration, the SUS word score at an averaged SRT ± a certain margin,[7] i.e. a pair of conditions, was measured and subsequently averaged over all participants. The method had been adopted from Brand and Kollmeier (2002).
The analysis of the word-scores at the approximated 25% and 75% intelligibility levels in Figure 4.5 shows on the whole correctly adjusted SNR levels. However, at the S_0N_{60} and S_0N_{120} conditions the influence of reverberation on spatial unmasking was underestimated and therefore most conditions were unintelligible. What can be noticed with respect to a possible failure in the measurement procedure (we assume an annotation failure), is the fact that the S_0N_{120} condition with a RT of 1.5 s scored higher than in the condition with a RT of 0.4 s, although the SNRs were equal. Despite that, we display these conditions, since the STI method follows the same trend.
Ideally, the binaural STI method should place all binaural conditions along the monaural/diotic polynomial. This is observed for most of the conditions to a fair degree. At a closer look, an overestimation at low SI scores and an underestimation at high SI scores of the objective method can be observed with respect to the polynomial. This result is in line with the outcomes of monaural and diotic conditions in Figure 4.4 C, E and F, which feature long-term SSN, while the polynomial was fitted to a mix of fluctuating and non-fluctuating SSN conditions. Consequently, (non-parallel, see next paragraph) deviations of the condition pairs from the poly-

[7] The SRT was measured individually and averaged.

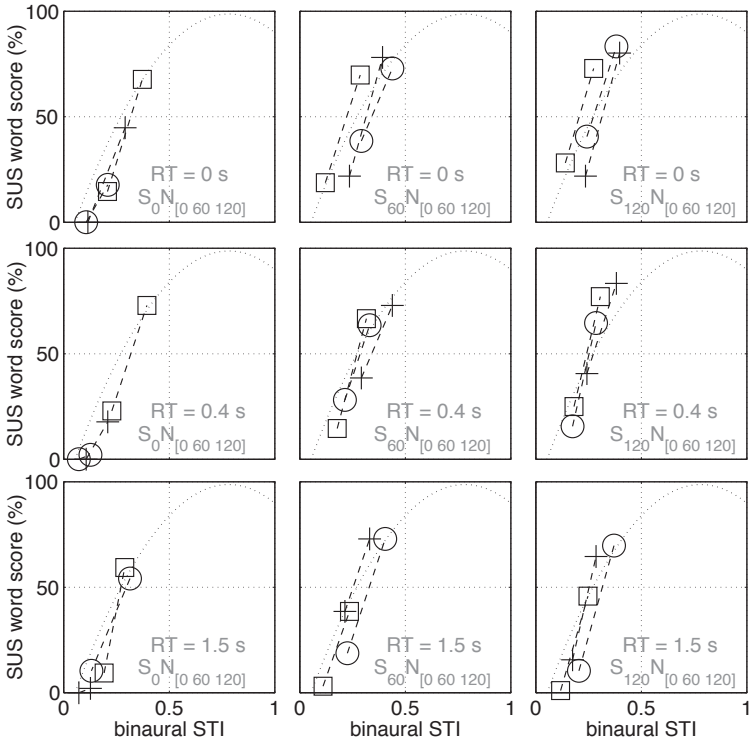

Figure 4.5: *Results of the listening test 2. The first row of the plots gives the results for estimated SRTs ± 2 dB in anechoic playback conditions. The second row gives the results for estimated SRTs +0 and +4 dB in reverberation with RT = 0.4 s. The last row of plots features the results of conditions using averaged SRT +0 and +4 dB with RT = 1.5 s. In each subplot, the noise source* N *rotated over the discrete angles of 0° (square), 60° (plus sign) and 120° (circle).* S *refers to the target speaker angle (see the annotations). The distracter* N *throughout all conditions was composed of the long-term SSN of a male voice.*

nomial might be attributed to the nature of the noise rather than the binaural processing.

Furthermore, we observe a correctly predicted trend of SI for the pairs of each particular spatial condition. However, when comparing the word-score STI relation over all conditions, especially in the anechoic setup, the trends are not always correctly reflected by the STI. This indicates a lack in binaural processing, specifically, because the slopes of condition pairs are in most of the cases horizontally shifted (Van Wijngaarden and Drullman, 2008). Thereby, this horizontal shift is subject to the spatial configuration, which, again, indicates most probably a lack of binaural

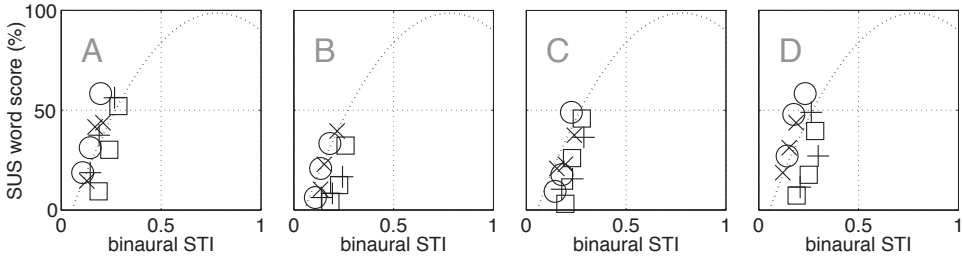

Figure 4.6: *Results of the listening test 3.* A*: 2 distracters and RT* $= 0$ *s,* B*: 2 distracters and RT* $= 0.4$ *s,* C*: 3 distracters and RT* $= 0.4$ *s and* D*: 4 distracters and RT* $= 0$ *s. In every subset, the target rotated over the discrete angles of 0° (square), 60° (plus sign), 90° (cross) and 120° (circle). The distracters throughout all conditions were composed of the long-term SSN of different male voices.*

processing.

(3) In a third binaural test, multi distractor scenarios were analyzed. The target speaker was placed at different angles in the mixtures of equally distributed distracters and the reverberation time was either set to 0 or 0.4 s, using the simulated setup of listening test 2. The target speaker was spatialized at 0, 60, 90 and 120 degrees. In the two distractor scenario, distractor one and two were spatialized at the target angle -120 and $+120$ degrees, respectively. In the three distractor scenario, distractor one, two and three were spatialized at the target angle -90, $+90$ and $+180$ degrees, respectively. Lastly, in the four distractor scenario, the distractor one, two, three and four were spatialized at the target angle -144, -72, $+72$ and $+144$ degrees, respectively. The SNRs of each spatial test condition were adjusted after estimating the average SRT. The averaged SRT was subsequently defined as a condition and an SNR margin of ± 2 dB was added for generating the second and third condition in each spatial setup.
Again, eight students with normal hearing participated in the test. See Figure 4.6 for the results. Although the test did not fully sample the whole intelligibility scale, the results give an insight into the behavior of the method in these complex conditions. With respect to the SRT, the overall reduction in scores across all conditions might be attributed to facilitation effects in the SRT method. As regards the results of the STI method, there, first, appears to be a horizontal shift, i.e. an underestimation of SI similar to the findings in test 2, as the target speaker changes its position from 0 to 120 deg. This probably indicates a shortcoming in binaural processing. Second, the more complex the situation becomes, the more fuzzy the binaural STI method gets. In Figure 4.6 D, for example, the algorithm misses the trend under the N_{60} conditions.

(4) In a final analysis, the proposed instrumental evaluation method is applied to

predict the Binaural Intelligibility Level Difference (BILD) for mutually permutated target/masker conditions (see Blauert, 1997, p. 265ff.). Therefore, the SRTs across a wide range of spatial configurations of 24 normal hearing students were measured in the standard way (Plomp and Mimpen, 1979). To calculate the objective BILD,

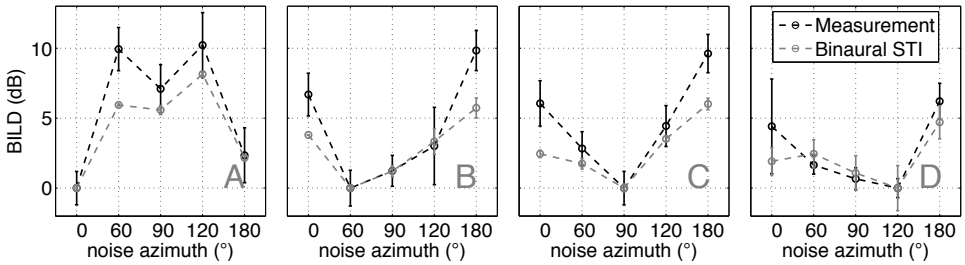

Figure 4.7: *Results of the listening test 4 of BILD measurements and predictions at* A*:* $S_0N_{[0\ 60\ 90\ 120]}$*,* B*:* $S_{60}N_{[0\ 60\ 90\ 120\ 180]}$*,* C*:* $S_{90}N_{[0\ 60\ 90\ 120\ 180]}$ *and* D*:* $S_{120}N_{[0\ 60\ 90\ 120\ 180]}$*. The distracter throughout all conditions was the long-term SSN of male speech. The confidence intervals equal two times the standard deviation.*

we proceeded in the following way: Based on a maximum likelihood fit, the psychometric function of the applied SUS corpus is approximated with:

$$p_{\mathrm{s}}(\mathrm{SNR}) = \left(\frac{1}{1 + e^{(\Omega - \mathrm{SNR})/\Psi}}\right), \tag{4.1.5}$$

where Ω is the SRT, Ψ the steepness[8] and p the probability of the correct response, which depends mainly on Ω in free-field conditions. When identifying p with a parametric SUS-STI approximation for $\mathrm{S_0N_0}$ conditions (in here the above introduced polynomial fitted through the monaural/diotic conditions) and when setting $\Omega = 0$, the SNR in Equation (4.1.5) predicts the BILD. Figure 4.7 gives the subjective and objective results. For all conditions analyzed, the objective method follows the trend. However, an underestimation of the binaural STI is observed the more orthogonal the target/masker angle gets.

■ 4.1.4 Monaural and binaural intelligibility in rooms

To evaluate differences between the monaural and binaural STI on various positions in an acoustic environment, a virtual room was simulated with the MISM of Van Dorp Schuitman (2009). The room is shoebox-shaped with dimensions W x L x H

[8]The parameter Ψ of the applied corpus equals 2.1, corresponding to a slope of 12% per dB for stationary SSN.

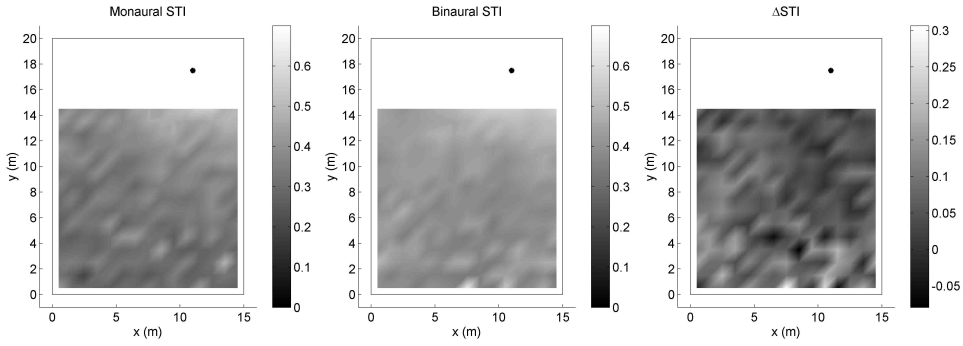

Figure 4.8: *The monaural (left-hand plot), binaural (centre plot) STI values for all receiver positions in the virtual room. The difference between binaural and monaural STI values, relative to the monaural STI is given in the right-hand plot. The source is denoted by a closed dot. The simulated room has a reverberation time of 1.25 s at low to mid frequencies.*

= 15 x 20 x 5 m (V= 1500 m^3). A virtual source (omnidirectional) was positioned in the front of the room, off centre and at a height of 2 m (see Figure 4.8). A total of 225 receivers were located in the room at the same height as the source, spaced 1 meter apart in the x- and y-directions. For each receiver location two impulse responses were simulated: First, a monaural impulse response, as it would have been measured in a real room using an omnidirectional microphone and second, a binaural impulse response, as it would have been measured using the Aachen head.

The configuration was such that image source modeling was used for reflections up to the second order. For the simulation of higher order reflections statistical modeling was performed. The binaural room impulse responses were generated by convolving the sound field with HRTFs of the artificial head as measured in an anechoic room. It was chosen to set a uniform absorption coefficient for the room boundaries, leading to a reverberation time of 1.25 s in low to mid frequencies and a reverberation radius of approximately 2 m. Above mid frequencies, a roll-off at higher frequency bands is modeled due to sound absorption through air. For each receiver position both the monaural and binaural STI values were determined with the envelope regression method. The results are shown in Figure 4.8.

As expected, the STI values are highest for receiver positions close to the source. Further away from the source, the late reverberation of the room can mask the direct sound, leading to a decrease in speech intelligibility. However, the human auditory system is capable of suppressing sound coming from directions other than that of the source, as discussed earlier. Therefore the monaural STI will underestimate the perceived intelligibility. As can be seen in Figure 4.8 the binaural STI indeed is generally higher, because it takes this suppression effect into account. The difference

between the monaural and binaural STI values is most apparent at locations far from the source, where the room reverberation is considerably higher than the direct sound level. To demonstrate this, the difference $\Delta\text{STI} = \text{STI}_{\text{bin}} - \text{STI}_{\text{mon}}$ is shown in Figure 4.8 (right-hand side plot). In diffuse fields, the difference can be as high as 50%, and sometimes even higher.

4.1.5 Discussion

In the first section of this chapter on assessing SI, we presented and evaluated a speech-based and binaural STI method. The STI method uses a Gammatone filterbank as an approximation of the peripheral auditory frequency decomposition. Subsequently a coincidence model of the binaural interaction process is applied along with head shadow processing, to establish an interaural representation. Based on an internal representation of the binaural target envelope and the degraded binaural envelope of the target in a sound scene that is to be analyzed, an efficient envelope regression method calculates the modulation transfer across different spatial directions. Ultimately, the direction offering the highest modulation transfer is selected for calculating SI in the standard way.

The method is capable of predicting monaural and binaural speech intelligibility under all of the here tested linear conditions with fair to good quality. Even fluctuating noise, a masker that has been addressed with a much more elaborated monaural model of SI by Rhebergen and Versfeld (2005), can be assessed with good accuracy. This capability is for the main part a consequence of the STI intrinsic time domain analysis, as opposed to a purely spectral measure like the SII (Payton et al., 2002). Despite that, the STI approach lacks fundamental stages of the auditory system. E.g. the modulation filterbank of the lower auditory system that is responsible for the TMTF is not modeled. The STRF analysis is not included either. These deficiencies are expected to lead to an underestimation of SI in e.g. co-modulation masking conditions (see Moore, 2003, p. 100) and to an overestimation if the carrier under an intact envelope is destroyed (Elhilali et al., 2003). Accordingly, we may find a slight but constant underestimation of the binaural STI in fluctuating masker conditions because of these simplifications of the STI method (see Figure 4.4 D).

The evaluation of nonlinearly processed speech revealed that the proposed STI method is insufficient in describing SI for this signal degradation, in particular in envelope thresholding conditions that typically occur in time-varying filter functions. This finding was expected, because, as previously mentioned, also high level SI models like the STMI fail in these conditions (Schlesinger and Boone, 2010). In Figure 4.4 B, the similarity between artificial envelope thresholding distortions and the CASA post-filter distortion was demonstrated. The observed slight deterioration of subjective speech intelligibility with the CASA processed scene is a consequence of the low SNR and a rather conservative implementation of the algorithm of Gaik and Lindemann (1986), who did not incorporate the probability based weighting

approach, as introduced in the previous chapter.
The second part of this chapter discusses the idea of using a high-level informational weighting based on the contextual importance of sub-phoneme bits on SI. In addition to an improved rendering of nonlinear distortions, this approach might also solve the very same problem described by Payton and Shrestha (2008) with the speech-based STI: "... the metric values are driven by voiced sounds in words such as vowels and, despite strong vowels, some key words have low probability of correct identification." A statistical approach to include context related information (on a sub-phoneme level) was developed by Kates and Arehart (2005b), whose method labels the importance of 32 ms long speech frames to SI according to their relative SPL and implements a three-level time domain weighting of a coherence based SII. Unfortunately, such a short-time labeling is questionable for the STI, since it analyzes the intensity envelope of the syllabic rate, which is in the range of 4 Hz.

The polynomial of the second order that was fitted to the linear conditions reminds us of the polynomials that were usually fitted to word-score/STI curves (Steeneken and Houtgast, 2002). In view of recent work on the problem, a psychometric function could be used to transform the results to a linear relation between subjective and objective results (see e.g., Christiansen et al., 2010). This approach is adopted in the following section.

Throughout all binaural listening conditions, the model shows a fair agreement with subjective results. However, the predictions are not free of errors across different spatial configurations in anechoic conditions, reverberation and multi-speaker scenarios, as it has been shown in listening tests two and three. The quality of the binaural STI is especially apparent in the prediction of the BILD of test four. The model renders the trend of the evaluation, although the predicted binaural unmasking underestimates the psychoacoustical results when the speaker and masker are spatialized with 90° separation. There may be several reasons for this observation. First, we find the results to be dependent on the fitting polynomial (of test one) that was used in calculating the BILD. Changing this polynomial, by for example excluding the fluctuating noise conditions, leads to a moderate nonlinear increase of the predicted BILD values. Furthermore, the binaural stage in the STI method is a simplified approach of the underlying neural process. As an additive superposition of the contributions that result in the overall release from masking, i.e. temporal disparity and the head shadow effect, is not observed in perceptual studies (Bronkhorst, 2000) and, generally, as the underlying neural mechanisms are not fully understood, the present binaural model will have difficulties with obeying all perceptual phenomena. Further research should analyze the influence of a non-uniform arrangement of increments in the cross-correlation stage. For example, the model of Dietz et al. (2009) accounts for the angular-dependent distribution of 'best IPDs' (to which neurons are tuned) that show a maximum around $\pm$ 60° with respect to the median plane. A better approximation of the binaural STI model prediction to the evaluation results might be efficiently accomplished with an optimization routine and a cost function, as e.g. the r^2 measure, as it is done in the following section.

What is missing in the evaluation of the binaural listening conditions with the binaural STI is a juxtaposition with the monaural STI, similar to the approach of Van Wijngaarden and Drullman (2008). Despite that, the advantage of the binaural STI over the monaural STI in unmasking, was qualitatively demonstrated by calculating monaural and binaural SI in a simulated room. The outcome reveals a deviation of SI in favour of binaural listening, even in the absence of competing sources. This deviation increases the more diffuse the sound field becomes, i.e. in the far-field of the source, as shown in the simulation of a large room.

Conclusion

A speech-based and binaural STI method has been presented and evaluated. The evaluation shows fair model predictions in the majority of the tested conditions, including conditions of fluctuating noise, reverberation and a multitude of spatial configurations. The speech-based and binaural STI is therefore appropriate in describing a wide range of speech intelligibility reductions. Although the functional relationship between subjective and objective assessment is not valid for combinations of linear and nonlinear disturbances, the measure indicates generally for each particular SI reduction a correct trend and can, therefore, be practical in the SI prediction of separate signal distortions.
Binaural listening can provide good improvement of SI through spatial unmasking of the target speaker. The results of the evaluation validate the conceptual approach of the method, although the full binaural advantage cannot be provided by the model. In spite of the fact that the model presented is not suited to reflect SI consistently across all conceivable distortions, future research should improve the accuracy of the method in the fields of room acoustics and linear distortions. These fields are the classical domains of the STI method. The linkage with a binaural processing stage may provide a thorough SI measure for natural binaural listening conditions. In addition, the application of a speech-based version may simplify the measurement procedure in room acoustics and simulation.

In the remainder of this work the STI method is left behind because of its mentioned shortcomings in predicting mask-based envelope-thresholding distortions. The following section attempts to formulate an algorithm that reflects SI of linear and nonlinear distortions in accordance with subjective speech perception.

4.2 The characterization of the relative information content by spectral features for the objective intelligibility assessment of nonlinearly processed speech[9]

The objective intelligibility assessment of nonlinearly enhanced speech is a very general problem. Nonlinear speech enhancement processors operate primarily on the low-level and transient components of speech. As these sections contain important acoustic cues as well as context-constitutive information, they dominate speech intelligibility. For that reason, short-time intelligibility measures at low-level and transient components are weighted by their contribution to the overall intelligibility. In this section, spectral features are calculated from auditory sub-bands and are used to label these sections of high information content. A genetic optimization is performed to adapt the spectral feature measures to the linearly and nonlinearly processed speech material of a listening test. The results demonstrate the general capability of the approach; however, the method is not as good as the I3 measure of Kates and Arehart (2005b). In addition, it will be shown, that this level based method can be further improved for the applied speech distortion corpus.

4.2.1 Introduction

One of the important questions in nonlinear speech enhancement asks for the balance between the SNR improvement and the introduced distortion introduced that is most beneficial for speech intelligibility (Kjems et al., 2009). While linear algorithms are often applied in speech processors, such as hearing aids, they generally represent a suboptimal solution in constantly changing acoustics. Nonlinear processors aim at approximating the MMSE, i.e. the optimal filter, in changing acoustics and can be realized as time-frequency mask-based approaches (see Chapter 2.3.1). It is yet the nonlinearity that constitutes a challenge for the subjective and objective evaluation of the audiological benefit.
In the subjective evaluation of speech processors, SRT tests are preferred over percent-correct scores, which show a limited basis for generalization of the audiological benefit (Greenberg and Zurek, 2001). SRT tests, on the other hand, require an invariant SNR improvement in order to quantify the algorithm's effect on speech intelligibility. Nonlinear algorithms do not satisfy this condition (Greenberg and Zurek, 2001). Moreover, subjective listening tests are laborious tasks and are not feasible during the development of complex algorithms, which are equipped with parameter sets that depend on the acoustic scene and, when considering hearing aids, on the particular auditory performance of a hearing impaired person.

[9] Apart from minor changes, the content of the present section was published in *Schlesinger, A. and Boone, M. M., "The characterization of the relative information content by spectral features for the objective intelligibility assessment of nonlinearly processed speech," Proceedings of the Interspeech conference 2010, Tokyo, Japan.*

Objectively, speech intelligibility can be calculated from the sum of the audible contributions in different frequency bands. This forms the empirical basis of the well-known Articulation Index (AI) theory. The Speech Intelligibility Index (SII) and the Speech Transmission Index (STI) are based on this concept and are successful in predicting intelligibility for a large number of linear distortions (ANSI/ASA, 2007; Steeneken and Houtgast, 2002). In spite of this, the classical SII and the classical STI are not adequate to assess disturbances due to nonlinear speech enhancement. Nonlinearity violates the principle of superposition and, therefore, the requirement of the classical SII approach to calculate the SNR from the isolated spectra of speech and noise cannot be met. The classical STI employs artificially modulated tones and calculates the SNR in bands, from the modulation depth of the intensity envelope. The principle of superposition would again be needed to derive an overall result from these test signals. Furthermore, nonlinear speech processors tend to increase the modulation depth abnormally. The classical STI mistakes this deterioration for an increase in speech intelligibility.

Intrusive measures have been developed to deal with nonlinear distortions. These are measures that relate degraded speech to its clean reference by calculating the difference or the correlation (Kates and Arehart, 2010). Some evolve along the same line as the methods of SII (Kates and Arehart, 2005b) and STI (Goldsworthy and Greenberg, 2004), others build on sophisticated perceptual models, as such the Perceptual Evaluation of Speech Quality (PESQ) model adapted for SI (Beerends et al., 2009) or the Dau model, as applied in Christiansen et al. (2010). Most of these methods correctly predict the perceptual trends for gradual changes of additive noise and nonlinear distortions. However, for the comparison between the input and output of a nonlinear speech enhancement algorithm, a functional relationship for these different kinds of distortion between subjective perception and objective prediction is essential.

The application of complex perceptual models could not solve this problem as long as the time-variant nonlinear processing was not taken into account. As it has been demonstrated in the previous section, speech enhancement is usually associated with an envelope thresholding distortion and merely modifies the low-level and transitional components of speech. Yoo et al. (2007) found that these portions hold only 2 % of the energy of the original speech but are almost equally intelligible. In order to compensate for this characteristic, Kates and Arehart (2005b) developed a coherence based SII (CSII) that includes a time-domain weighting based on the short-time RMS level. They showed that short-time sequences of -10 to 0 dB with respect to the overall RMS have a major contribution to speech intelligibility.

On the contrary, Taal et al. (2011a) recently made progress with a purely speech-based approach of an intrusive short-time measure for mask weighted noisy speech. Their model is based on equally contributing transitional intelligibility scores, calculated in windows of 400 ms length. As the analysis windows overlap by half, this particular window-length is in the order of the syllabic rate of speech. In predicting speech intelligibility of ideally binary masked (IBM) speech, the objective intelligibility measure of Taal et al. (2010) demonstrated to be superior to other recent

models for this problem (Christiansen et al., 2010; Boldt and Ellis, 2009).

In this section, it is investigated whether spectral feature measures can be applied to identify transitional components of speech and to what extend these are suitable to weight short-time intelligibility predictions with respect to the varying, i.e. relative information content. The utilization of spectral features in speech processing is not new. Spectral features are, by way of example, successfully used to improve automatic speech recognition tasks (You et al., 2004; Housseinzadeh and Krishnan, 2007).

We proceed now in the following way. First, four spectral feature measures for the separation of speech into voiced and unvoiced segments are presented. Adapted to the occurance of relative high information content, these short-time measures form feature vectors that are applied as weighting to the time-course of the short-time CSII method of Kates and Arehart (2005b). Thereafter, a parameter optimization of the measures to psycho-acoustical data is given and a comparison to the results of state-of-the-art speech intelligibility measures is provided. Finally, we draw a conclusion and select an objective measure, which will be applied in the binaural speech processor optimization and evaluation of the following chapter.

4.2.2 Algorithms

The algorithmic approach is divided into two parts. In the first part, the extraction and adaptation of the spectral features is examined. In the second part, the coherence based SII method is presented.

Extraction of source information with spectral features

The source-filter model is a widely applied method in speech processing, to separate the excitation characteristics of the vocal chords from the resonator characteristics of the mouth. In this model, speech is assumed to be short-time stationary in the range of 10 to 30 ms. Accordingly, the model can be formulated as a linear convolution:

$$s(\iota) = \upsilon(\iota) * \varsigma(\iota), \tag{4.2.6}$$

where $\upsilon(\iota)$ is the source and $\varsigma(\iota)$ is the filter that forms the speech signal $s(\iota)$ as a function of discrete time sample number ι. Linear prediction and cepstrum techniques are often used to separate $\varsigma(\iota)$ from $\upsilon(\iota)$. In here, the main interest lies simply on the differentiation whether the glottis produces white noise for unvoiced speech or a periodic stimulation for voiced speech. This feature can be directly calculated from $s(\iota)$. In a preliminary study, four spectral measures were identified that correlate with this change of articulation. These are the Renyi Entropy (RE), the Shannon Entropy (SE), the Spectral Band Energy (SBE) (Housseinzadeh and Krishnan, 2007) and the Madhu Flatness Measure (MF) (Madhu, 2009c). As the

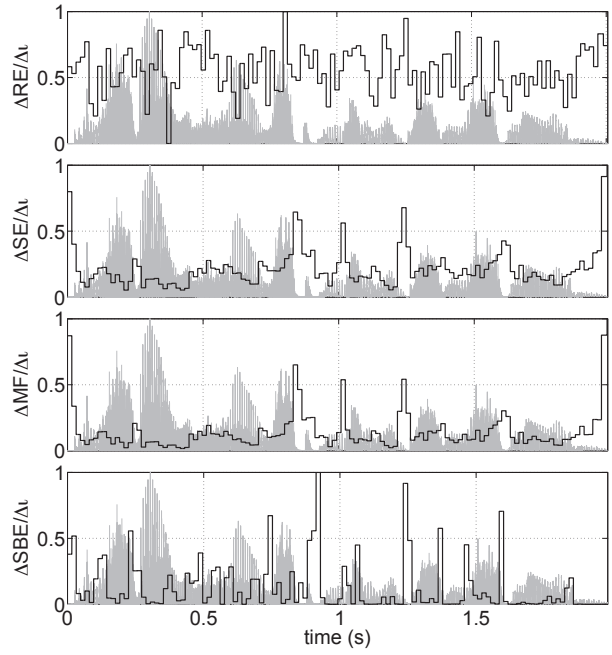

Figure 4.9: *Graphs of the differentiated spectral feature measures, exemplarily applied to label the transitional parts of a sentence, whose positive waveform is plotted in the background (grey shaded).*

transitions between phonemes and formants, i.e. the low-level and transient components of speech, are essential for intelligibility, the first order derivative with respect to the short-time frames was taken to flag these segments of speech. In Figure 4.9 the four differentiated spectral feature measures are shown. For comparability, the waveform of the analyzed sentence is plotted in the background (grey shaded). In detail, the measures were calculated from the clean speech waveform that was sampled at 22.05 kHz. An analysis window of 706 samples (32 ms) was Hanning weighted and padded with zeros prior to a 1024-point DFT. With a window-overlap of 50 %, a new spectrum was calculated every 16 ms. The STFT spectra with frequency coefficient d were filtered block-wise with centre frequencies and bandwidths of the auditory critical bandwidth filters, as given in Table I of ANSI/ASA (2007). Prior to the calculation of the entropy measures the sub-band amplitude spectrum was normalized to obtain the probability density functions:

$$\hat{s}_{\iota,b}(d) = \frac{|s_{\iota,b}(d)|}{\sum_{d=d_\mathrm{l}}^{d_\mathrm{u}} |s_{\iota,b}(d)|}, \tag{4.2.7}$$

where $s_{\imath,b}(d)$ is the STFT representation of frame $\imath$ that was subdivided in non-overlapping critical bands b, with a lower DFT coefficient frequency bound d_l and an upper coefficient frequency bound d_u. RE was then calculated with:

$$\mathrm{RE}_{\imath,b} = \frac{1}{1-\beta} \log_2 \left(\sum_{d=d_\mathrm{l}}^{d_\mathrm{u}} \left[\hat{s}_{\imath,b}(d) \right]^\beta \right). \tag{4.2.8}$$

The order β was set to 3 throughout this study. SE was calculated as:

$$\mathrm{SE}_{\imath,b} = -\sum_{d=d_\mathrm{l}}^{d_\mathrm{u}} \hat{s}_{\imath,b}(d) \log_2 \hat{s}_{\imath,b}(d), \tag{4.2.9}$$

and the MF measure as:

$$\log_2(\mathrm{MF}_{\imath,b}+1) = -\frac{1}{\log_2(N_{\sum \mathrm{d}})} \sum_{d=d_\mathrm{l}}^{d_\mathrm{u}} \hat{s}_{\imath,b}(d) \log_2 \hat{s}_{\imath,b}(d), \tag{4.2.10}$$

where $N_{\sum \mathrm{d}}$ is the amount of DFT frequency bins in frame b. The SBE, which is no information theoretic measure, displays the relative energy distribution, is calculated as:

$$\mathrm{SBE}_{\imath,b} = \frac{\sum_{d=d_\mathrm{l}}^{d_\mathrm{u}} |s_\imath(d)|^2}{\sum_d |s_\imath(d)|^2}. \tag{4.2.11}$$

Thereafter, the results in sub-bands were weighted with the band importance function for average speech (Pavlovic, 1987) and summed. This resulted in short-time feature vectors.

To adapt these to the transitions of high information content in an optimization routine, the feature vectors were either compressed, or expanded by an exponent e. Subsequent to the differentiation, a first order recursive smoothing was applied to allow the feature vectors to align with the phonemic transitions. With $\rho_\imath \equiv [\Delta \mathrm{RE}_\imath^e/\Delta\imath, \Delta \mathrm{SE}_\imath^e/\Delta\imath, \Delta \mathrm{MF}_\imath^e/\Delta\imath, \Delta \mathrm{SBE}_\imath^e/\Delta\imath]$, the smoothed output was calculated as:

$$\rho_\imath = (1-\ddot{\alpha})\rho_\imath + \ddot{\alpha}\rho_{\imath-1}, \tag{4.2.12}$$

where $\ddot{\alpha} = e^{-\Delta T/\breve{\tau}}$, ΔT is the frame shift and $\breve{\tau}$ is the time constant. This procedure of contrast enhancement was finalized with the introduction of a threshold value according to:

$$\rho_\imath = \begin{cases} \rho_\imath, \rho_\imath > \epsilon_\mathrm{v} \\ 0, \rho_\imath \leq \epsilon_\mathrm{v} \end{cases}. \tag{4.2.13}$$

The coherence SII

The intrusive CSII predicts the segmental SNR, i.e. the $\mathrm{SNR}_\mathrm{seg}$ between the input $s(\iota)$ and the output $x(\iota)$ of a system by calculating the signal power fraction that

is linear related. This is possible through the magnitude squared coherence (MSC) function:

$$|\Delta\gamma(d)|^2 = \frac{|\sum_{\imath=1}^{M_\imath} s_\imath(d)x_\imath^*(d)|^2}{\sum_{\imath=1}^{M_\imath}|s_\imath(d)|^2 \sum_{\imath=1}^{M_\imath}|x_\imath(d)|^2}, \tag{4.2.14}$$

where $s_\imath(d)$ and $x_\imath(d)$ are the STFT representation of the input and the output signal, respectively, and the asterisk denotes the complex conjugate. Subsequently, the $\mathrm{SNR}_{\mathrm{seq}}$[10] per critical band b can be calculated:

$$\mathrm{SNR}_{\mathrm{seg},b} = \frac{\sum_{d=1}^{N_\mathrm{d}} \mathbf{\Pi}_b(d)|\Delta\gamma(d)|^2 \sum_{\imath=1}^{M_\imath}|x_\imath(d)|^2}{\sum_{d=1}^{N_\mathrm{d}} \mathbf{\Pi}_b(d)(1-|\Delta\gamma(d)|^2) \sum_{\imath=1}^{M_\imath}|x_\imath(d)|^2}, \tag{4.2.15}$$

where $\mathbf{\Pi}_b$ is a matrix of rounded-exponential filters that are described in Kates and Arehart (2005b). To include the weighting with the feature vectors $\rho_\imath$, the summation terms over $M_\imath$ frames in the Equations (4.2.14) and (4.2.15) were replaced with $\sum_{\imath=1}^{M_\imath} \star\rho_\imath$, where $\star$ is either the cross-spectral density or the auto-spectral density. After the calculation of the $\mathrm{SNR}_{\mathrm{seg},b}$ in Equation (4.2.15), the CSII index calculation is conform the standard ANSI/ASA (2007).

4.2.3 Fitting of feature vectors to subjective data

In order to adapt the weighting of the feature vectors to subjective intelligibility scores, a nonlinear optimization was performed. The listening test set comprised five conditions of additive speech shaped noise and five conditions of envelope thresholding, which is a nonlinear distortion that occurs in speech enhancement processors (see the previous section). Details on the implementation of the envelope thresholding conditions are found in Kates and Arehart (2005b) and in Appendix D. Eight people with normal hearing (< 15 dB HL) participated in a percent-correct score test that uses the SUS corpus in German of Ramirez et al. (2009). The subjects had to respond to three versions of each condition, which were presented to the right ear. The participants were paid for their services and trained. The entire set of conditions was subsequently used for the objective index calculation. To account for the nonlinear relation between subjective and objective results, a logistic function was applied to the objective results Υ:

$$\Lambda(\Upsilon) = \frac{1}{1+e^{\psi+\varphi\Upsilon}}. \tag{4.2.16}$$

The parameters ψ and φ together with the parameters of the feature vectors $\ddot{\alpha}$, ϵ_v and e were optimized with the r^2 measure as a cost-function of a linear regression. A genetic algorithm was applied to the non-monotonic optimization problem (Houck et al., 1995). The optimized parameters are given in Table 4.2.

To compare the results with existing measures, the Short Time Objective Intelligi-

[10] Kates and Arehart (2005b) used the term speech distortion measure and showed the equivalence of this metric with the RMS-based SNR.

Table 4.2: *Results of the parameter optimization for adjusting the spectral feature measures to the applied speech material and subjective scores.*

feature	ψ	φ	ϵ_{v}	e	$\ddot{\alpha}$
SBE	4.1	6.2	0.33	1.6	0.69
SE	3.6	5.3	0.27	1.5	0.79
MF	3.8	5.1	0.33	1.17	0.79
RE	2.4	9.1	0.39	3.45	0.02

bility (STOI) measure of Taal et al. (2010), the CSII of Kates and Arehart (2005b), the three RMS level weighted CSII of Kates and Arehart (2005b) and an optimized one RMS level CSII were included in the evaluation. The optimization of the one RMS level CSII was additionally executed with a genetic algorithm. In this procedure, two level ranges and their weighting in a logistic function were provided for the algorithm (in total six parameters). As the combination of two level ranges and their logistic weighting can be expressed as a linear combination, the algorithm converged to one optimal level with an upper dB bound of -4.9 dB and a lower dB bound equal to -7.7 dB (for $\psi = 1$ and $\varphi = 8$).

The results of the optimization are shown in Figure 4.10. As figures of merit, the r^2 measure and Kendall's τ, a rank statistic, are given. As can be seen, none of the spectral feature measures yield a frame-weighting that improves the correlation, compared to the existing measures. Only RE shows some improvement over the unweighted CSII, at the cost of a higher standard deviation. This result was confirmed by others, who found an RE weighting to be beneficial for enhancing the performance in automatic speaker identification tasks (Housseinzadeh and Krishnan, 2007). Good results were achieved with the STOI measure, the RMS three level CSII and the one RMS level optimized CSII. However, before applying the one RMS level CSII, the measure has to be studied in more exhaustive listening tests before a general judgment on its quality can be made. The STOI measure already experienced comprehensive testing in Taal et al. (2010) and showed a correlation coefficient of 0.95 on IBM processed speech data.

■ 4.2.4 Discussion

The application of the analyzed spectral measures did not yield an advantage over existing methods in labeling speech sections of high information content. In spite of that, the spectral feature measures presented, showed from their first inspection to their adaptation in the short-time frame weighting of speech intelligibility, a fair degree of conformity. In particular the differentiated SBE, MF and SE measures show similar patterns, responding strongly to changes from voiced to unvoiced speech (see

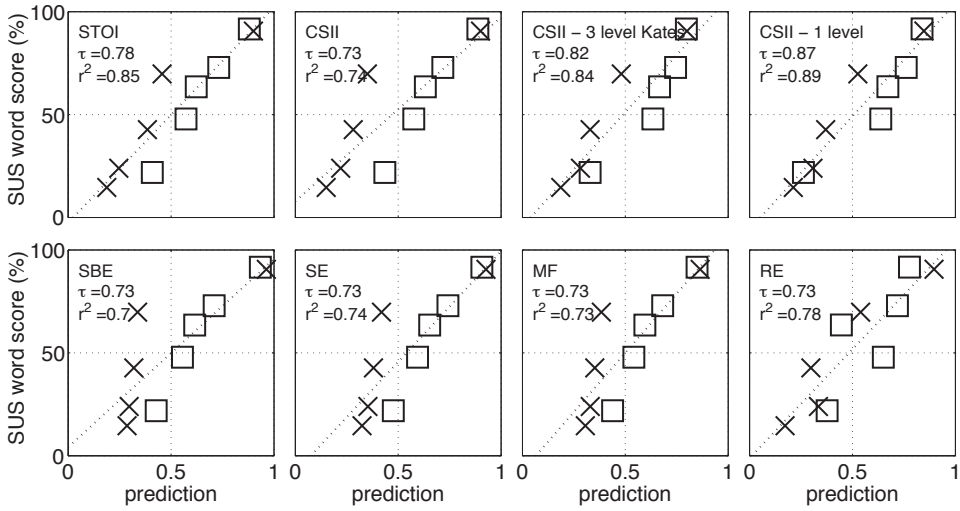

Figure 4.10: *The first row shows the word-score/prediction relation of the short-time measures* STOI *and* CSII*, as well as the short-time weighted measures* CSII − 3 level Kates*, also known as the I3 measure. The upper right-hand plot gives the results of the optimized* CSII − 1 level*. The lower row shows the results of herein analyzed short-time weighting CSII methods, which are based on differentiated spectral feature measures* SBE*,* SE*,* MF *and* RE*.* □ *denotes envelope thresholding conditions and* × *the additive noise conditions (SSN). A regression line was fitted to the data. The* r^2 *measure and Kendall's* τ *are given to assess the quality of the SI models.*

Figure 4.9). Their sensitivity to these changes is confirmed by low threshold values, $\epsilon_v < 0.33$, and small expansion values, $e < 1.6$, as found in the optimization (see Table. 4.2). Their first order low pass filter time constants $\breve{\tau}$ are in the range of 40 to 70 ms. Apparently, the smoothing corrected consistently for the misalignment of the feature vectors with the transitional parts in speech. Even though no improvement over the unweighted CSII was achieved, the spectral measures proved to label speech information fairly well. This is obvious when remembering that envelope thresholding acts mainly on low level regions. If a spectral feature measure would consistently miss these speech sections, speech intelligibility would be strongly overestimated.

RE is less sensitive to the changes of voiced and unvoiced speech. Nevertheless, the differentiated RE reveals steep slopes in speech sections of high information content, which is expressed through a small filter coefficient in Table 4.2. The absence of a clear feature pattern, however, leads to a high threshold and expansion exponent, which results into a low r^2 value due to the increased deviation from the regression line.

In view of the good performance of the level-weighted CSII measure, it can be stated that the observed feature measures are less adequate predictors of the relative speech information content. Since the feature measures were simply extracted from auditory critical filters, it is questionable whether better results can be achieved by applying these to the source signal, or the filter signal alone, in Equation (4.2.6). If, e.g. entropy measures are calculated from the slow changing transfer-function of the vocal tract, an increased local stability and a better identification of transitional speech parts might be yielded.

Here entropy was only used to discriminate between voiced and unvoiced passages. Leijon (2007) makes the observation that the acoustic speech information rate and the performance of speech perception are coupled. He further showed that there is no direct relation between information rate in frequency bands and the empirical additivity-concept of audible contributions in different frequency bands of the AI theory.

In terms of the CSII, our findings are different from the results reported by Taal et al. (2011b). There, a STOI-like magnitude spectral correlation coefficient outperforms a series of recent objective measures, including the level-weighted CSII. The reason lies for the most part in the difference of the speech data that were used, and the manner in which the objective measures respond to the phase of the signal. While our test material comprises envelope thresholding distortions, which set the waveform to zero once the envelope of the signal falls below a certain threshold, Taal et al. (2011b) used IBM processed speech with strong implications for the phase of the target speech. In IBM processed speech the original phase information may be largely, if not completely, discarded and replaced with a different uniformly distributed phase. Under these circumstances the magnitude squared coherence function, on which the CSII is based, is biased to zero, although the IBM processed speech may be fully intelligible (Taal et al., 2011b). A striking example is the IBM separation of a mixture at -60 dB, i.e. essentially pure noise, that shows a speech intelligibility of 100 % (Kjems et al., 2009). Consequently, the results of this section are only valid for speech material that is similar to the one assessed here.

In the context of the present work, we still have to evaluate the objective intelligibility measures on CASA enhanced speech. These speech processors generally establish a soft-mask weighting and approximate the Wiener-filter. Furthermore, these filters optimally operate at favourable SNR ratios, i.e. when the target signal has significant energy. Therefore, the original phase of the target signal is substantially included in the distribution of sources across time-frequency bins. For this reason the previously assessed envelope thresholding conditions and the soft-mask approaches in terms of conserving the original phase, are approximately compatible. Therefore, preservation of the original target phase at the output of the here analyzed binaural CASA processors provides scope for the application of the level based CSII in the remainder of this work.

For binaural processing, a better ear three RMS level based CSII for speech intelligi-

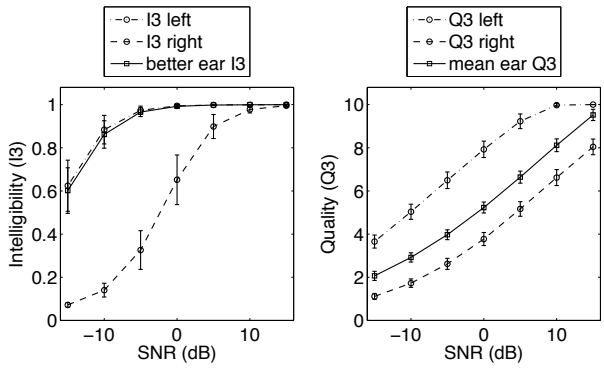

Figure 4.11: *The intelligibility index I3 and quality index Q3 of (Kates and Arehart, 2005a), extended with a "better ear" decision stage for the I3 index and a "mean ear" averaging for the Q3 index. The working principles are shown for a range of SNR conditions in white Gaussian noise. The confidence intervals equal two times the standard deviation.*

bility and a mean ear three RMS level based CSII for speech quality were developed. Following the development of their monaural counterparts by Kates and Arehart (2005a), these measures are denoted 'Better Ear I3' and 'Mean Ear Q3', respectively. For the calculation of the mean ear and better ear effect, Equations (4.2.14) and (4.2.15) are expanded such that the segmental SNR is calculated per critical band b and RMS weighted MSC values at both ears, using the same window size for the time course weighting of 32 ms. As the MSC function of Equation (4.2.14) tends to be faulty (cf. Chapter 2.4) when calculated from a single STFT bin across the clean and the degraded signal, a longer time series (a window of 1.7 s length is used for calculating the mean of the MSC in the present implementation) is necessary to maintain the accuracy of the monaural index. Subsequently, the maximum segmental SNR per critical band b and per window for calculating the mean MSC, using a 50 % frame shift of 850 ms, is chosen for calculating the binaural advantage in terms of the better ear effect. Hence, the binaural image is updated every 850 ms. As regards the mean ear effect, a binaural averaging is performed per critical band b and per window of the mean MSC calculation. A final averaging of the segmental SNRs is executed for both instrumental measures to calculate the overall indexes.

The similarity between the monaural and the binaural measure is demonstrated through the standard deviation in Figure 4.11 for a range of mixing SNRs. For the purpose of this experiment, 100 SUS sentences of the Ramirez et al. (2009) corpus were spatialized with a frequency-independent ILD of -14 dB to generate these plots. As can be seen, the Better Ear I3 is equal to the monaural I3 at the left ear. The loss in binaural intelligibility unmasking is considered to be in the order of 2 to 3 dB, when excluding the effect of temporal differences, i.e. binaural interaction

(Bronkhorst, 2000; Van Wijngaarden and Drullman, 2008).
Due to the nonlinear three RMS level weighting, the result of the Mean Ear Q3 is tending towards the quality of the ipsilateral ear, for the observed signal degradation. Although data on binaural quality perception seems to be missing in literature, our choice of an equal averaging across the ears for binaural algorithms is in line with the binaural quality assessment of Rohdenburg (2008).
As mentioned above, the three level RMS weights of Kates and Arehart (2005b) have been applied in previous experiments and are applied in the remainder of this thesis. Although the one-level weighting found in this work outperforms the method of Kates and Arehart (2005b) for the speech material used here, we prefer generalizability and comparability as well as a well established measure, over a possible moderate gain in accuracy.

Finally, a clarifying remark to a question that may arise at this point. The simplification of the binaural stage with respect to the binaural stage of the proposed binaural STI model is due to incompatibility with the CSII measure and the limited amount of time the author could spend on the problem of objective speech intelligibility enhancement in the context of the overall aim of this work, i.e. the enhancement of speech intelligibility. Furthermore, the fact that the SII is intrinsically not suitable for the assessment of reverberated speech (Schlesinger and Boone, 2010), is circumvented in the majority of acoustical situations, assessed in this work. That is, most mixtures are an additive superposition of clean target speech and a real-world background or another coherent interferer.
If, however, the target speech is reverberated, the better ear decision stage will be applied to the segmental SNR, as previously described, however, without a time-course weighting. The instrumental measure is subjected to an intelligibility weighting, with the band-importance function for critical bands given in ANSI/ASA (2007), and will be referred to as the Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ in the remainder of this work. Although it is a central message of the present work that an SNR measure that is indifferent to articulatory features in speech, is inappropriate to assess nonlinearly processed speech, we return to this general manner of objective assessment in case of reverberation, since we are lacking better means.

Conclusion

This second section of the objective assessment of speech intelligibility dealt with the objective assessment of linearly and nonlinearly processed speech. It has been analyzed whether spectral feature measures can be utilized to label the relative information content in short-time frames of speech, in order to establish an improved time-course weighting for an SI prediction of nonlinearly processed speech. Although the analyzed spectral feature measures are capable of labeling transitional sections, which substantially affect SI, these measures do not represent an alternative to the existing RMS level-based weighting method of the short-time CSII measure. Further optimization has shown that only a small level range in speech, predominantly tran-

sients and articulatory hubs, contribute to intelligibility. This result corresponds to the findings of Yoo et al. (2007).

For the purpose of predicting binaural speech intelligibility of nonlinearly processed speech, the three RMS level weighted CSII was implemented in a better ear fashion and introduced as the Better Ear I3. Based on the observations made in the present chapter, the Better Ear I3 measure is employed for evaluating and optimizing binaural CASA speech processors in the following chapter.

5

Optimization and assessment

This chapter deals with the optimization and the assessment of binaural CASA speech processors that are connected in series to different binaural front-ends with and without beamforming. Based on the tools and types of enhancement and assessment of speech intelligibility that were presented in the previous chapters, it is the aim to assess objectively the audiological benefit of binaural CASA speech processors, in various acoustical environments. Prior to this assessment, the algorithmic parameters of the three binaural speech processors of this thesis are optimized with a genetic algorithm. Subsequently, the assessment of these processors will address the questions of generalizability of optimized filter solutions in changing acoustics, the influence of the front-end and a performance comparison of the processors. In addition, as the applied binaural speech processors approximate different modes of the binaural processing in the model hearing process, we will pursue the general question which strategy offers the best source separation power in a given sound scene.

5.1 Introduction

In order to define the scope of this chapter, we will briefly recapitulate the relevant findings of the previous chapters and draw conclusions, for the following setups.

The theoretical introduction of this work shows that MVDR beamformers can be combined with CASA-based post-filters to achieve an MMSE solution. Three binaural CASA post-filters have been introduced in Section 2.4. Each of them uses a binaural measure, or a combination of binaural measures, to separate speech from noise. More specifically, algorithm CC is conceptually similar to the algorithm of Allen et al. (1977), and separates speech from noise by a primitive classification scheme based on the binaural waveform coherence. Algorithm CLP is based on the binaural filter of Gaik and Lindemann (1986) and utilizes binaural cues of the fine-structure waveform in a pattern-driven separation processes. The third post-processor, algorithm ELT of Kollmeier and Koch (1994), also applies binaural cues in a pattern-based fashion, to improve speech intelligibility. This algorithm sup-

presses interference in a joint centre and modulation frequency domain and applies the binaural cues of the envelope waveform as a directional classifier.

Binaural cues are considered the only low-level cues that are independent of articulatory and speaker-dependent features. This constitutes a great advantage over monaural speech enhancement approaches, which have to account for these dependencies. Moreover, the presented binaural speech processors offer a binaural output signal, thereby enhancing the audiological benefit, and they can be implemented as real-time processors.

In a statistical attempt to define the problem of speech enhancement using a time-varying filter, in Chapter 2.3.2, it has been shown that the SNR of the acoustic waveform is an incomplete measure for defining the problem of speech intelligibility enhancement. The local SNR at time-frequency units has been found to be a better means for this task. However, to facilitate the comparability with the majority of similar work (e.g. Wittkop et al., 1997), we have consistently referenced and will reference the speech-in-noise problems to a waveform-based SNR.
In order to limit the scope of the following study, the optimization will be performed at an SNR of 0 dB through different acoustic conditions, and subsequently be assessed over a varying SNR range. An SNR of 0 dB is considered to be typical in a cocktail-party situation (Bronkhorst, 2000). This situation determines the baseline environment in terms of the here applied binaural CASA processors without a beamforming front-end. For binaural speech processors that are serially connected to bilaterally applied front-ends, the source level SNR can be much lower. However, because the statistics of the binaural parameters have shown to be mainly determined by the ear-level and binaurally averaged SNR, the present study on optimization and assessment is further confined to an ear-level SNR and binaurally averaged setup definition.
Chapter 3 shed light on the characteristics of binaural parameters of the waveform fine-structure as well as of the corresponding envelope in distorted wave fields. The results reflect psycho-acoustical findings, stating that binaural temporal difference cues of the envelope are more susceptible to noise than the carrier analogue.
Additionally, realistic binaural parameters, measured at the entrance of the ear-channel of a mannequin head, have been compared to the modified binaural parameters at the output of bilaterally applied hearing aids. In the following, it is a central question whether the binaural parameter modifications as a consequence of the hearing aid transfer function, will have an effect on the separation power of post-processors.
The application of commercially available hearing aids in this work allows for the incorporation of manufacturing and fitting imperfections as well as the entire digital processing chain. Consequently, the following study aims at a realistic picture of the audiological benefit, albeit the mixing setup, given in Chapter 3.2, poses a set of restrictions to the simulated realism in the following section.
The probabilistic weighting approach that has been introduced in Chapter 3.3.3, represents an elegant and efficient means to take account of the peculiarities of a

certain binaural front-end and the acoustics it is applied to. The strategy greatly increases the applicability of binaural cues in noise, although, needless to say, it cannot cancel out their physical deterioration in adverse situations. The probabilistic weighting method is used in the following study. To that purpose, the histograms of the lookup tables were built as described in Chapter 3.3.3.

The problem of assessing intelligibility of binaural and nonlinearly processed speech has been covered in Chapter 4. A better ear version of the I3 measure of Kates and Arehart (2005a) has been developed for the optimization and the assessment of the binaural and non-linear CASA processors. Alternatively and where necessary, use will be made of the here derived mean ear version of the quality measure Q3 of Kates and Arehart (2005a), and of the underlying coherence-based and intelligibility-weighted Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ measure.
As these objective metrics are based on speech segments of 32 ms length (with 50 % overlap), these are considered to be capable of the assessment of fluctuating speech (Eneman et al., 2008, p. 446). Accordingly, fluctuating speech (without low-level sections using a simple VAD method) will be applied during coherent interference situations. To demonstrate this possibility, a linear regression of the I3 measure against the SUS word scores for fluctuating and stationary noise was performed. For the test procedure and test conditions, see listening test one in Section 4.1.3 and Table 4.1. The resulting regression analysis is given in Figure 5.1. The r^2 statistic indicates that 70 % of the variance in the data is expressed by the I3 measure.

The level for understanding speech by 50 % in noise is generally determined with the

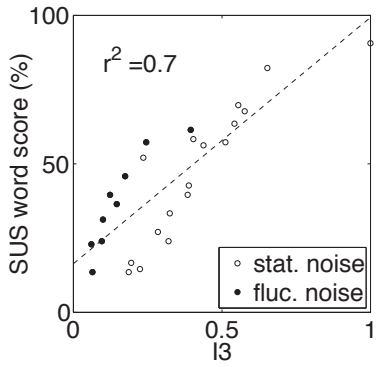

Figure 5.1: *Regression analysis of the percent-correct scores versus the I3 measure for stationary and fluctuating noise conditions, which were introduced in listening test one in Section 4.1.3.*

SRT method. For people with normal hearing in the presence of stationary noise,

this threshold is usually found at −4 to −5 dB (monaural presentation). The SII of the standard ANSI/ASA (2007) predicts this threshold for stationary noise at an SII value of 0.33 (George, 2007, p. 117). However, as it can be seen in Figure 5.1, this I3 threshold at the SRT depends on the character of the noise. For example, in fluctuating noise the threshold of 50 % intelligibility is found at around an I3 of 0.25. For non-fluctuating maskers, on the other hand, the I3 is inclined towards higher values. This deviation has to be considered in the analysis of different acoustical situations.

The fusion of the algorithmic results of the previous chapters forms the basis of the current chapter. A universal procedure is consequently established, in which binaural CASA processors are combined with specific front-ends with and without beamforming, and in which these are optimized and analyzed in realistic conditions with an objective measure of speech intelligibility.

The following chapter is split into two parts. In the first part, the CASA post-processors are optimized in order to adapt the parameters of the algorithms to a specific binaural front-end as well as to a specific acoustic scene. In the second part, the optimized CASA algorithms are assessed in changing acoustic environments.

5.2 Parameter optimization of post-filters

A unifying problem to many fields of science is the optimization of algorithms or processes. In acoustics, the speech processing branch faces an increasing algorithmic complexity which makes the optimization challenging. In fact, a deterministic search for an optimum performance is often not possible, as an exhaustive enumeration of a multidimensional search space demands, even for relatively small problems, an impractical computational effort. At the expense of accuracy, stochastic search algorithms reduce the calculation effort. In general this trade-off strongly reduces the time of convergence while yielding a good solution. The problem of optimization is well known in hearing aid design. Influences of the individual pathology and listening preferences, on the performance of particular hearing aid features, like the insertion gain, compression or feedback cancellation pose a difficult optimization problem.
Several approaches of user specific adaptation of hearing aid parameters exist (Dillon, 2001, p. 312 ff.). However, as the algorithms advance in complexity, it is likely that users are streched by the decisions they have to take.
Therefore, stochastic optimization using a genetic algorithm (GA), was introduced in the optimization of acoustic feedback cancelation (Durant et al., 2004) and the optimization of cochlear implants (Baskent et al., 2007). In both of these studies, the GA optimizes algorithmic parameter sets with subjective feedback. The results prove the qualification of a GA in these tasks.

For the current problem of optimizing complex CASA processors, the approach of

using a subjective input as a fitness assignment in the GA procedure is desirable. However, a subjective test would require great efforts to discretize the search spaces according to perceptual scales, as it has been done by Durant et al. (2004). Moreover, in order to keep the optimization manageable, parameter sets would have to be confined to a small number for each GA iteration. Additionally, efficient SRT tests are excluded in the assessment of nonlinear processors and percent-correct score tests are laborious and lack generalizability (Greenberg and Zurek, 2001). The present work, therefore, adopts the idea of the model-based improvement and the model-based assessment of speech intelligibility, which was formulated as a future means for an efficient hearing aid design by Hohmann (2008).
By utilizing a GA in the optimization of CASA processors, this idea is further developed on the grounds of evolutionary model processes. The outcome of the optimization does not only produce an optimum parameter set for each CASA processor in a controlled listening setup, it also shows how the model hearing might utilize the different cues in the best possible way. The validity of this approach, however, has to be qualified. That is, the results of the optimization are subject to the degrees of freedom that are made in the model assumptions, such that these are not unconditionally equivalent to the neural functionalities.

5.2.1 Genetic optimization framework

The parameter optimization using a GA is based on the evolutionary principles of the survival of the fittest strategy. Therefore, a biological individual with a certain sequence of genes (a chromosome) can be abstracted, for our task, as a CASA speech processor with a certain set of algorithmic parameters. The GA selects good and best individuals based on an objective function that defines the respective fitness in a certain domain. Besides the selection of good solution candidates, the GA incorporates the principles of genetic mutation and crossover. Together, these are the three genetic operators that create new solutions. Principally, the optimization can be written as a maximization approach over a search space L. Considering $\aleph : L \rightarrow \mathbb{R}$ to be an objective function that assigns to each solution $\wp \in L$ a fitness value, the GA maximizes this fitness-value with:

$$\aleph(\wp_{\clubsuit}) = \max\{\aleph(\wp) \mid \wp \in L\}, \tag{5.2.1}$$

where $\wp_{\clubsuit}$ is the final parameter solution of a GA run. Here, the objective speech intelligibility measure I3 in a better ear fashion was applied as the objective function (Better Ear I3), unless another one is mentioned.

Many variants of genetic operators exist and often these are tailored to specific problems, as e.g. done by Durant et al. (2004) and Baskent et al. (2007). In the present work, no such predefined GA algorithm was applied. Instead the Genetic Algorithms for Optimization Toolbox by Houck et al. (1995) was used. This toolbox offers a

variety of genetic operator versions and has demonstrated to be efficient in many different optimization problems (Houck et al., 1995). Throughout the optimizations performed in this work, we applied the defaults of the GAOT that were found to perform well for a wide range of optimization tasks (Houck et al., 1995). This additionally accounts for the more natural representation of parameters with floating point numbers, which speeds up the convergence of the optimization by an order of magnitude in terms of CPU time (Houck et al., 1995).
Here, we will not give a thorough analysis of the quality of the GA solutions and the general suitability of the GA procedure with respect to the genetic operators and their functional parameters. To get a general indication of the optimization complexity, we examined the reproducibility of GA solutions, by running several GA optimizations for particular setups. Additionally, the GA-optimized parameter-sets are applied in changing acoustic environments, which gives information about the robustness and generalizability of a certain solution. Whether the best solutions of the GA, found here, are perceptually insignificantly deviating from the optimum solutions, as e.g., shown by Baskent et al. (2007), cannot be inferred—mainly because the optimum solutions are not known.

For the optimization task, the additive mixtures of simulated anechoic recordings with real-world scenes of Chapter 3.2 are utilized. Silent gaps in the clean speech material, defined as a drop of the RMS level by −50 dB relative to the overall RMS level in frames of 10 ms, were excluded with a VAD procedure. To introduce target speaker diversity in the parameter set evaluation, tokens of 5 s length of one female speaker and two male speakers from the TNO (2000) database were randomly selected and concatenated to 15 s of diverse speech material. Interfering speakers acquired their sentence material from the same database (including the same VAD procedure). Care was taken that no overlap with the target speaker existed and no repetition within the 15 s of target speech occurred. In conditions where the influence of reverberation was to be assessed, the MISM setup of Figure 3.1 in Chapter 3.2 was used.
The size of the initial population, that seeds the GA search space with random, but bounded, parameter values, as well as the number of iterations, i.e. the termination condition of the GA, were set according to the complexity of the optimization problem. Thereby the reproducibility of overall GA solutions served as an indicator.
No effort was made to discretize the search-space for integer parameters. Therefore, the floating point numbers of certain parameters were rounded to the nearest integer value.

The following subsections treat the optimization of the three binaural speech processors at the output of different front-ends and in different acoustical situations.

5.2.2 Optimization of the CC algorithm

In Chapter 2.4 the binaural coherence-based post-processor of Allen et al. (1977), in here abbreviated as algorithm CC, was introduced. Four algorithmic parameters were selected to fine-tune the coherence-based separation scheme with respect to the front-end and the acoustic scene, in which the post-processor is applied. The optimization using the introduced GA framework was started with an initial random sampling of parameters, i.e. an initial population of 1000 parameter sets and 300 iterations of the GA. This optimization was performed for a set of front-ends and the real-world scenes that are introduced in Chapter 3.2. Additionally, the post-processor was optimized for quality in the canteen environment, for which the Mean Ear Q3 index served as the objective function in the GA process.
In order to analyze the effect of reverberation, the binaural output of the Aachen head in the above-introduced MISM-simulated room with a reverberation time of 0.2 s was utilized to generate the S_0N_{90} condition for an optimization. The intelligibility weighted Better Ear SNR_{seg} was applied in the GA optimization and the final assessment.[1]

Figure 5.2 gives the results of the optimizations. The outcomes are described below with respect to a set of specified characteristics. This facilitates the comparison with the optimization results of the other binaural speech processors.

Attenuation of diffuse noise: When optimized for the Better Ear I3 measure, the CC processor gains a moderate improvement of speech intelligibility in the canteen, the workshop and the bus environment. As can be seen, in the bus environment, the Aachen head already offers maximum speech intelligibility at an SNR of 0 dB. The reason for this could be a wide dynamic range, a flat omnidirectional frequency-transfer at low frequencies and no nonlinearities in the processing chain of the mannequin. In addition, the low-frequent bus noise might leave enough undistorted speech energy in mid- to high-frequency ranges. It is well-known that speech can be fully intelligible, even if it is audible in a band-limited presentation as narrow as one-third of an octave (see e.g., Barker, 2006, p. 318).
As indicated by the lowest algorithmic results during the GA runs, and as expected by the efficient working range of the MSC function, the CC post-processor appears to work fairly robust in diffuse sound fields, where it offers interference suppression in the mid- to high-frequency range.
Hardly any improvement is seen for the quality optimization with the Mean Ear Q3 measure in the canteen situation, despite the fact that similar sets of

[1] As mentioned in the previous chapter, the I3 measure, which is a version of the SII, is generally an inaccurate measure to assess the impact of reverberation on speech intelligibility (Schlesinger and Boone, 2010). The underlying coherence-based SNR_{seq} in the method of Kates and Arehart (2005b) lacks a high predictive power of speech intelligibility under these conditions too. However, as a remedy for the assessment of reverberated and nonlinearly processed speech, the intelligibility weighted Better Ear SNR_{seq} is used as a measure.

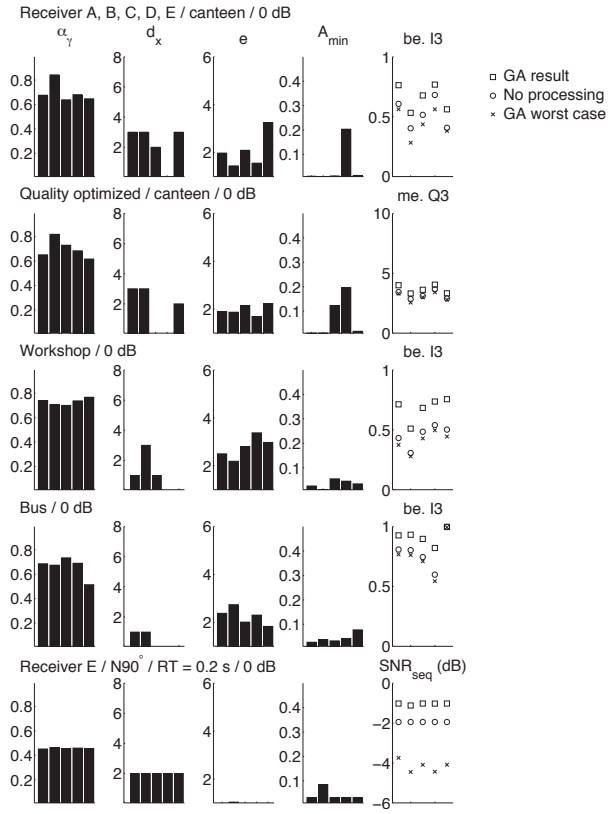

Figure 5.2: *Parameter optimization using a GA of the coherence-based post-processor CC at the output of* A*, the HG (low directivity),* B*, the HG (high directivity),* C*, the BTE (omnidirectional),* D*, the BTE (directional) and* E*, the Aachen head. Sequentially, each bar in the plots refers to the algorithmic parameters of these front-ends, with the exception of the last row of the plots, where five repeated GA optimization runs are plotted for the Aachen head front-end in the simulated room setup with one interfering speaker at 90 deg and RT = 0.2 s. The optimized parameters are the* α_γ *the MSC smoothing constant,* d_x *the lower cutoff frequency of the coherence weighting,* e *the compression/expansion exponent and* $A_{\min}$ *the maximum suppression (for further explanation see Chapter 2.4). The right-hand column of plots gives the results of the applied objective function for each front-end/scene combination. The results of the optimization of the HG modes in the canteen environment have to be considered with care due to a recording failure. See Chapter 3.2 for further information.*

$\wp_{\clubsuit}$ (excluding the BTEs) resulted in comparable Q3 variations between worst and best solutions.

Interestingly, the estimation of the MSC by averaging the auto and the cross power spectra is accomplished with a relatively small value of the time constant α_γ, i.e. $\breve{\tau} = \frac{-\Delta T}{\log_n \alpha_\gamma}$ is in the range from about 10 ms to 40 ms. This might be attributed to the fact that the MSC-weighted mask is eventually applied to the modulus of the STFT representation. An increase of α_γ results in a smoothing of the target speech, which is interpreted as a decline of speech intelligibility by the intrusive speech intelligibility measure.
This observation is in contrast with the algorithmic settings of Peissig (1992), who reported a time averaging α_γ in the range of the syllable and phoneme rate, i.e. up to a value of α_γ which is an order of magnitude higher than found in the present GA optimization. His approach led to a subjective quality improvement in a highly reverberant environment. By informal listening, we confirm the qualitative benefit of a higher smoothing constant. The quality measure Mean Ear Q3 used in the GA optimization, however, produces also a relatively small α_γ, because it is an intrusive measure. Consequently, the quality perception associated with a longer time constant cannot be predicted. See also the discussion at the end of this chapter.

Attenuation of coherent noise: A small improvement in terms of Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ is found for the attenuation of one interfering source at 90 deg in a simulated room with RT = 0.2 s, using the Aachen head as a front-end.
As the STFT components in algorithm CC are calculated per blocks of 16 ms length, the MSC at zero-lag is effectively incapable of distinguishing between coherent signal components from the front and from the sides (see also the examples given in Chapter 2.4). Moreover, since the interfering source is inside the reverberation radius at one meter distance ($r_{\mathrm{RT}} = 1,2$ m), only part of its energy is diffused by reverberation and could consequently be labeled as diffuse through a low MSC value. It stands to reason to exclude algorithm CC from further attempts of improving speech intelligibility in coherent interference conditions in the following. Furthermore, preliminary inspection showed that no speech intelligibility enhancement can be achieved in fully coherent interference conditions. Therefore, this scene setting was already excluded from the GA optimization task presented here.

Comparison of directional and non-directional front-ends: Using the ear-level SNR as a common reference among the front-ends, no clear difference can be observed in the GA selection of optimal algorithmic parameters. The moderate variation of the enhancement results (cf. the lowest scores in the GA solution sampling in the right-hand column of plots) should be an indication of the robustness of the algorithm.
To improve the direct-to-reverberation ratio, hence to increase the direct sound energy of a target speaker, Martin (2001) suggested to employ directional front-ends as a pre-processor to algorithm CC. The anticipated benefit refers to the overall gain, which is not computed here. However, the improvements

are unmistakeably realized, as expressed in the balanced benefits across the different directional front-ends, despite the fact that the SNRs at the source-level are much lower for the directional aids.
The parameter sets are qualitatively comparable. Nevertheless, these are to some extent specific to each binaural front-end across different environments. See, e.g. the parameter sets of the HG (high directivity) front-end across different conditions.

Reproducibility of solutions by the GA: To test the quality of the GA solution, five GA runs were conducted for the coherent interferer condition in mild reverberation and using the output of the Aachen head. The test condition can be considered difficult, because algorithm CC can hardly suppress the interfering source or diffuse signal components in this scene setting. Consequently, the improvement of the Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ is small. In any case, the GA realized the application of algorithm CC without a loss in terms of Better Ear $\mathrm{SNR}_{\mathrm{seg}}$, which supports the necessity for an algorithmic optimization. Finally, the parameter sets of the best solutions indicate by visual inspection that two maxima have been found in the search space. Despite the difficult optimization task, the close similarity of the two referring parameter set solutions indicate a considerable degree of accuracy for the observed GA optimization.

5.2.3 Optimization of the CLP algorithm

This section presents the optimization of the binaural processor of Gaik and Lindemann (1986), denoted algorithm CLP in this work. The working principle of the algorithm was introduced in Chapter 2.4. For each front-end and scene combination a GA optimization was executed with an initial population of 2000 parameter sets and 300 genetic cycle iterations. The results are given in Figure 5.3. Similar to the previous analysis, the observations are subsumed under the previously defined categories.

Attenuation of diffuse noise: In the analyzed canteen and workshop environment the improvement of speech intelligibility is in the range of 20 %.[2] When comparing these results of the canteen environment with algorithm CC, there is slight improvement, nonetheless the gain remains moderate. In the workshop environment, on the other hand, the improvement at the output of the BTE in two directivity modes, is lower than attained with algorithm CC (row three in Figure 5.3).
Noticeable is that the ILD parameter is not used in both the canteen and the workshop environment at the output of omnidirectional receivers, as expressed

[2] In a first approximation, we may speak of a linear relationship between I3 and subjective speech intelligibility.

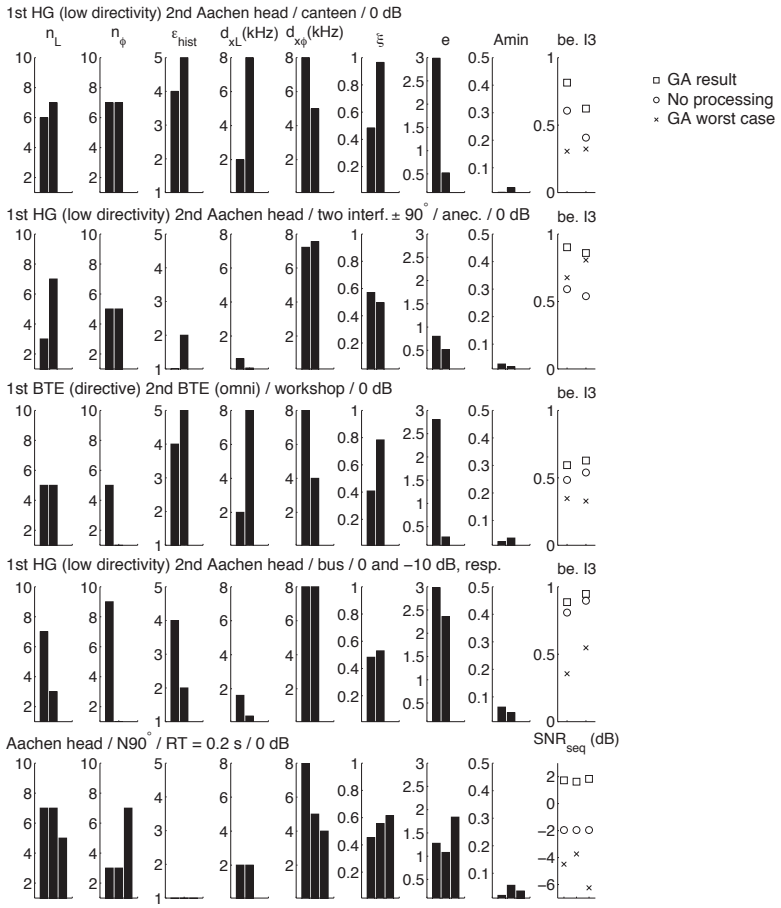

Figure 5.3: *Parameter optimization using a GA of the CLP speech processor. Each bar of the plots refers to a parameter with a certain front-end (see headline for each row). The right column gives the speech intelligibility results of the GA procedure. In the lower row, the reproducibility of the GA solution is checked with the Aachen head in a MISM-simulated room with an interfering speaker at 90 deg and a reverberation time of 0.2 s (assessed with the Better Ear SNR_{seg}, for the reasons mentioned above). The optimized parameters of the algorithm are n_{L} the bin-size of a 2D filter the ILD domain is smoothed with, n_{φ} (n_{ϕ} in the plot) the bin-size of a 2D filter the IPD domain is smoothed with, ϵ_{hist} the threshold of the probability weighting, d_{xL} the lower cutoff frequency of ILD-based weighting, $d_{\mathrm{x}\varphi}$ the upper cutoff frequency of the IPD-based weighting, ξ the ILD/IPD tradeoff in the weighting process, e the compression/expansion exponent of the weighting function and A_{min} the maximum suppression in the weighting process.*

through a high value of parameter d_{xL}. In case of directional front-ends, however, the ILD is used in the weighting process above 2 kHz.
The bus situation has previously shown to be difficult to evaluate across different front-ends. Speech intelligibility at the output of the Aachen head remains high, even at a global-mixing SNR of −10 dB, which was chosen to lower the speech intelligibility for the optimization as a special exception. However, a strong fluctuation of the GA solutions indicates a sensitivity of the algorithm with respect to the parameter sets $\wp$. Finally, an almost negligible improvement is achieved under this condition with either front-ends (fourth row of plots in Figure 5.3). The limited speech enhancement gain is certainly imposed through saturation effects, in terms of noise suppression, at high levels of intelligibility in the unprocessed situation. However, the benefit of the GA optimization can be seen in both optimization setups, in which possible parameter set solutions sample the wide range of speech intelligibility.

Attenuation of coherent noise: The improvement of speech intelligibility in the canteen environment is slightly higher than observed with algorithm CC. This advantage is likely a consequence of the presence of a couple of coherent speech sources in this environment, which are more efficiently suppressed with algorithm CLP.[3]
In the presence of two interfering sources at ±90 deg (second row of the bar charts in Figure 5.3), the algorithm achieves an absolute intelligibility improvement of 40 % with both the HG (low directivity) front-end and the Aachen head. The parameter values show that some sequential and spectral clustering of sound objects is performed in either of the cases in the binaural domain, as expressed with n_{L} and n_{φ}. A histogram threshold ϵ_{hist} of zero and two for the HG (low directivity) and the Aachen head, respectively, indicates resolved (or separated) binaural parameter distributions in the probability weighting function (see Chapter 3.3.3). Both binaural parameters are relevant throughout the entire spectrum, as expressed by a low d_{xL} and a high $d_{\mathrm{x}\varphi}$ for the ILD and IPD, respectively. In both final parameter sets the ILD is included in the weighting process by a ratio of approximately 1/2 with respect to the IPD parameter, see parameter ξ. The upper cut-off frequency of the IPD parameter is possibly resulting from the circumvention of the spatial Nyquist limit with lookup tables.
The attenuation of a coherent source in slight reverberation demonstrates how the algorithm is tuned to a certain default state (fifth row of plots in Figure 5.3). The ILD and IPD parameters are equally smoothed. Furthermore, no smoothing of the histograms is performed (ϵ_{hist} of zero); the ILD parameter

[3] In Figure 2.10 of the introduction, both scenes, the canteen and the workshop environment, were compared using the long-term averaged MSC. The comparison indicated that both environments are equally coherent (or the opposite), when using a common smoothing constant for binaural recordings of 20 s length. Not shown in this figure is the MSC average on shorter spans, which might be (at times) higher (and lower) for the canteen environment because of coherent interferer in the proximity of the recording position.

is used at frequencies higher than 0 or 2 kHz (d_{xL}) and the IPD cue is employed up to at least 4 kHz ($d_{\mathrm{x}\varphi}$). Moreover, both directional cues are equally important in the weighting, as indicated by a $\xi \approx 0.5$. Even though no GA parameter set $\wp_{\clubsuit}$ is found twice in the reverberation setup, the final solutions show no random fluctuation and realize in each case an improvement in the range of 4 dB.

Comparison of directional and non-directional front-ends: No distinct difference can be seen between the directional and non-directional front-ends in terms of speech intelligibility. With respect to the parameter sets, a close qualitative similarity can be found between the first row and the third row in Figure 5.3. Under these similar conditions, the front-end characteristics appear to have an influence on the parameter choice.
Overall, the parameter sets adapt considerably to the peculiarities of the front-ends for all environments and generate a (on average similar) benefit independent of the front-end.

Reproducibility of GA solutions: The search space has eight dimensions in the current optimization. The repetition of three GA runs in the coherent interferer situation in slight reverberation, produced three different parameter set solutions. Nonetheless, the solutions result in similar maximum Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ outcomes. As compared to algorithm CC, the final GA solutions lead to a higher benefit, while the worst GA solutions are as low as found with algorithm CC. Although this is a consequence of a lower algorithmic robustness, it also shows the potential of the algorithm that needs to be activated by scene-dependent algorithmic parameters.

5.2.4 Optimization of the ELT algorithm

Finally, the algorithm of Kollmeier and Koch (1994), in this work named ELT algorithm, is optimized. The algorithm was introduced in Chapter 2.4. As the search space has 12 dimensions, the solution sampling was increased to an initial population of 3000 parameter sets and 500 GA iterations. Besides, three optimization runs were performed for each front-end/environment combination, to minimize the chance of non-optimal GA solutions.
The optimized conditions, that are presented here, were restricted to the front-ends of the HG (low directivity) and the Aachen head. Additionally, the free-field noise condition of coherent interference was simplified such that it possessed only one interfering source at 90 deg, instead of two at ±90 deg, as used for the optimization of algorithm CLP. The choice was made to assure a high disjointness of sources in the carrier and modulation frequency domain of algorithm ELT, which is coarser than the carrier frequency domain of algorithm CLP. In addition, the histograms of the weighting function are coarser in terms of the standard deviation of the binaural

parameters. Therefore, more than two sources show an increased potential to overlap (see Chapter 3.3.2). An overlap, be it in the carrier and modulation frequency domain, or in the binaural representation, would hamper speech enhancement under the free-field test conditions that are optimized here.

To focus this analysis on the most salient effects, the optimizations of the workshop and the bus environment are not given here, as preliminary inspection showed no advantage in terms of speech intelligibility.[4] In addition, the optimization outcomes in the canteen environment are limited to reporting the HG (low directivity) front-end results. The GA optimization of the Aachen head will be compared to the GA results of the HG (low directivity) front-end, in the coherent noise condition only.

The outcomes of the optimizations are given in Figure 5.4. The observations are again summarized with respect to the above introduced four characteristics.

Attenuation of diffuse noise: The threefold optimization attempt of the algorithm in the canteen condition (first row of plots in Figure 5.4) does not result in an improvement of speech intelligibility. In the process no identical parameter sets of $\wp_{\clubsuit}$ are found by the GA. Likely due to the inadequacy of the algorithm in this situation, the search presumably takes place across a rather flat landscape of solutions. At a closer look, however, we observe a sound parameter balancing.

The first two optimization runs, for instance, show a considerable application of the ITD cue in the canteen condition. The clustering across the ITD plane is small, $n_{\mathrm{t}} = 4$, demonstrating the attempt at directional sensing. This purpose appears to be confirmed by an increased threshold of the weighting histograms, $\epsilon_{\mathrm{hist}} = 4$. Moreover, the standard deviation weighting is active, as indicated by a high $e_{\sigma\mathrm{t}}$ exponent. Finally, a high compression of the weighting function $\acute{\mathcal{M}}^{\mathrm{ft}}_{\mathrm{elt}}$ in Equation (2.4.46), expressed by small e values, demonstrates that a rather soft weighting is applied. Hence, the masks are presumably applied to suppress carrier modulation frequency bins that hold outliers across the ITD representation.

Overall, the results, including the GA results for the Aachen head front-end/canteen combination, which are not shown here, suggest that the algorithm does not introduce a deterioration of speech intelligibility in these situations through an adaptation and a reduction of the weighting.

The reduction of the weighting based on binaural cues can change the algorithm's functioning in such a way, that it is only driven by the binaural parameter standard deviation weighting. Based on the algorithmic parameters given by Kollmeier and Koch (1994), we believe that an active standard deviation weighting has been the main reason for the 2 dB SNR improvement

[4] Despite the absence of these diffuse field conditions in the present section on the optimization of algorithm ELT, the following section will provide the overall optimized performance of this speech processor in the workshop environment in Figure 5.7.

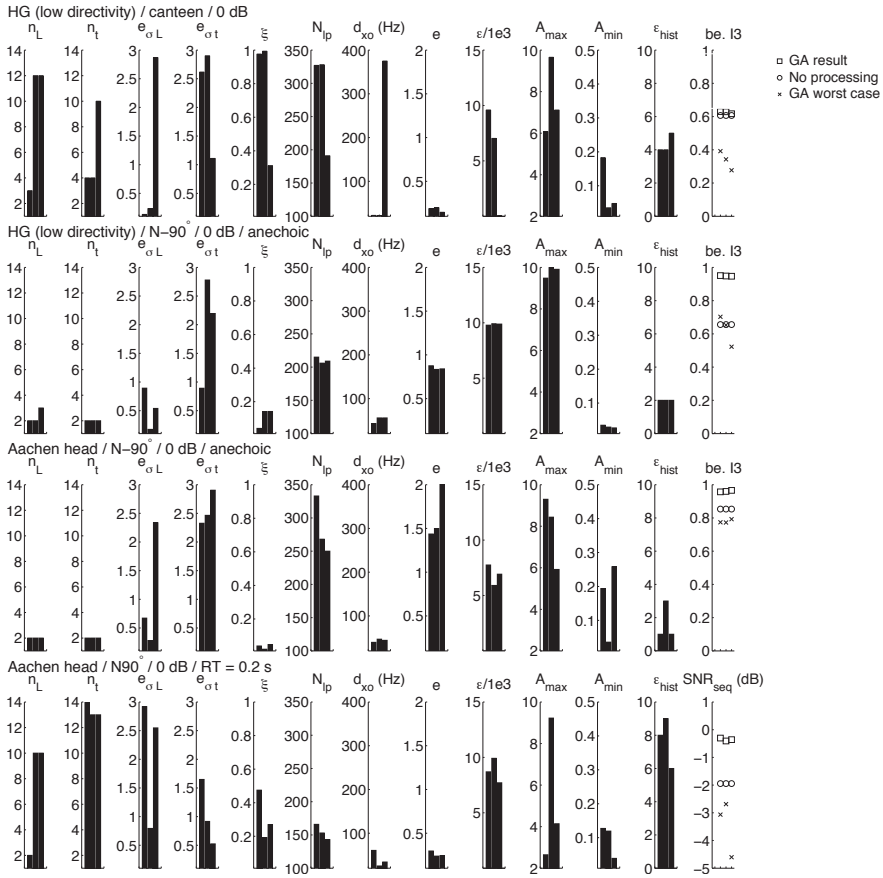

Figure 5.4: *Optimized parameters using a GA of the ELT speech processor for different front-ends and acoustic conditions (see titles in each row). The right column gives the results of the optimization in terms of speech intelligibility and in terms of the Better Ear SNR_{seg} for the reverberation condition. For each environment, the optimization was repeated three times to observe the reproducibility of individual parameters and set solutions. The parameters of the optimization are n_{L} the bin-size of a 2D filter the ILD domain is smoothed with; n_{φ}, the bin-size of a 2D filter the ITD domain is smoothed with; $e_{\sigma\mathrm{L}}$ and $e_{\sigma\mathrm{t}}$ are exponents for the compression/expansion of the ILD and IPD standard deviation masks, respectively. ξ is the ILD/IPD tradeoff in the weighting process; N_{lp} and $d_{\mathrm{x}\varphi}$ are the filter order and the cutoff frequency of an envelope lowpass filter, respectively; e is the compression/expansion exponent of the weighting function and ϵ a small constant for preventing a division by zero; $A_{\max}$ and $A_{\min}$ are the upper and lower bound of the mask $\mathcal{M}_{\mathrm{elt}}^{\mathrm{ft}}$, respectively, and ϵ_{hist} is the histogram threshold of the probability weighting.*

the authors found (see the introduction of the algorithm ELT in Chapter 2.4 and the discussion at the end of this chapter). In the current implementation of the algorithm, this standard deviation weighting is only loosely defined. To tighten the definition of this weighting function, one could, for example, narrow the level and time ranges in Equation (2.4.43) and (2.4.44), respectively. Probably the loosely defined standard-deviation based penalty weighting of our implementation and the negative implications of a less resolved modulation decomposition, result in a lower SNR gain in diffuse conditions, than achieved by Kollmeier and Koch (1994). As previously described, the lower resolution of the modulation spectrum in our implementation of algorithm ELT is required to allow for a directional weighting based on the binaural parameters of the envelope.

Attenuation of coherent noise: Indicated by the efficient suppression of a single noise source in anechoic conditions (second and third row of plots in Figure 5.4), but offering no benefit in slightly reverberating condition (fourth row of plots in Figure 5.4), algorithm ELT is shown to be merely able to suppress highly coherent noise sources. Throughout all GA parameter set solutions, the magnified ILD cue (see Chapter 3.3.2) is preferred over the ITD cue. The choice is in accordance with the observed characteristics of binaural envelope cues in noise in Chapter 3.3.2. As compared to the ITD cue, a lower standard deviation of the magnified ILD cue has been detected in this study.
The repetition of the optimization did not yield quantitatively identical parameter set solutions. However, the parameter sets converge qualitatively and achieve similar intelligibility scores. Low n_{L} and n_{t} values, as well as a low ϵ_{hist} testify that the probability weighting functions show a clear directional pattern.
Whereas the final mask is compressed and as such rather smooth in the canteen enhancement task, a parameter e of approximately 1 for the HG (low directivity) front-end and value around 1.5 to 2 for the Aachen head front-end, shows that the directional mask filtering is highly active in the weighting process.

Comparison of directional and non-directional front-ends: The comparison between the front-ends has been limited to the coherent noise condition, as this is the only condition in which the algorithm achieves an improvement in terms of speech intelligibility. Overall, the GA finds similar solutions for both front-end types. A difference exists for the expansion of the masks. Probably due to the fact that natural binaural parameters show less overlap (and thereby little ambiguity), the weighting function can be aggressively applied at the output of the Aachen head. Additionally, more fluctuation in finding optimal lower and upper bounds of the weighting function (A_{min} and A_{max}, respectively) exists in the GA optimization of the Aachen head output.

Reproducibility of GA solutions: As a consequence of the increased dimensionality of the search space and the search across a likely shapeless optimization

function, the genetic search did not yield identical parameter set solutions for the four conditions tested. This also holds for the coherent noise condition, in which the algorithm offers a clear benefit and should show dominant maxima of speech intelligibility across the parameter space. Nonetheless, the GA search resulted in qualitatively similar results for each front-end in coherent noise conditions.

Discussion

The GA optimization of algorithmic parameter sets produces practical, optimal and logical solutions. While the efficiency of the GA procedure is a consequence of the survival of the fittest strategy, the logic that has emerged from the solutions, is a product of the interplay between CASA algorithms in the improvement and assessment of speech intelligibility in particular environmental arrangements. Moreover, this algorithmic logic validates the underlying model assumption that it compares well with psycho-acoustic findings.
The GA approach offers a great benefit in the tuning of parameter sets, as natural processes are often too complex to be efficiently analyzed and described. Moreover, the GA approach provides solution strategies that may underly natural ranking processes of low-level cues. Of course, the solutions will only reflect what the system allows.

Throughout the optimization of the three CASA speech processors it was found that parameter sets show generally a high variation across possible solutions (as indicated by the lowest GA solution in each optimization run). For this reason, the optimization itself is necessary to gain, in the best case, an improvement of speech intelligibility and, in the worst case, not to introduce deterioration. A scene detector, similar to the ones introduced by Wittkop and Hohmann (2003) and Bach et al. (2011), could be applied in future hearing aids to switch between optimal parameter sets and programmes, or to turn off the processing when no improvement is expected.
With respect to reproducibility we found that smaller optimization problems can be comprehensively calculated with the GA in order to find identical solutions. If the optimization problem becomes more complex and the algorithm lacks suitable means to deal with a scene, a lot of variation in algorithmic parameter sets is observed as a consequence. Nevertheless, no deterioration of speech intelligibility has been introduced by the binaural processors under any condition.
With respect to the applicability of binaural parameters in adverse conditions, we find the observations of Chapter 3.3 on their statistics in noise verified. The temporal difference parameter of the fine-structure (here the IPD) has shown to be the most important parameter in both incoherent and coherent conditions. For front-end/scene combinations in which the carrier ILD is more dominantly included in the directional weighting process, the ILD never gains more than half of the 'attention' used in the total directional decision taking process. Using only binaural envelope

cues, the envelope ILD is favoured over the envelope ITD in coherent interference conditions. In diffuse interfering scenarios, as the here observed canteen environment, the ratio can be the inverse.
In the second part of this chapter, it is observed whether the optimal solutions, found here, lead to robust improvements under changing acoustics. Before we proceed with that question, the present section is completed with the results of the optimization of the smoothing parameters that are used in the cepstral smoothing process.

5.3 Optimization of cepstral smoothing constants

The technique of cepstral smoothing is treated in this thesis to reduce the perceptually annoying phenomenon of musical noise in mask-based approaches. For an introduction to the method see Chapter 2.3.4. The following paragraphs deal with the GA optimization of the time constants and the maximum mask-based suppression constant used in the cepstral smoothing process. We start with a description of the optimization setup.

Prior to the optimization, mixtures with audible artifacts were generated. The target speech material was a 10 s long concatenated series of sentences, spoken by three speakers (one female and two males). Three sentences of the TNO (2000) corpus were selected for every speaker. Pauses in the speech material, defined as a drop of the SNR of -50 dB relative to the overall equalized RMS sentence level in 10 ms frames, were excluded with a VAD technique. Subsequently, the speech material was mixed with the ambiance recording of the canteen (the Aachen head recording was made monaural, without an equalization for the HRTFs) and mixed at SNRs between -15 and 15 dB using increments of 5 dB.
As it has been illustrated in Figure 2.6 in Chapter 2.3.2, the above-mentioned mixture violates the disjointness property of Equation (2.3.18) as the energy of the target signal is not dominating in well-separated time-frequency areas, but rather in a smooth balance with the noise. If speech enhancement with an IBM approach, presented in Equation (2.3.15), is performed for such a mixture, audible artifacts occur as a consequence of: sparsely scattered binary weighting values, the discrepancy between the actual energy proportion between target signal and the interference, and the nonlinear IBM approach. Hence, the generated mixtures and the IBM weighting method offer a suitable approach for inducing the musical noise phenomenon. To counteract musical noise, the Equations (2.3.20) through (2.3.24) present the cepstral smoothing technique used in the present speech enhancement strategy.

To sample a great variety of IBMs, the local criterion ε in Equation (2.3.15) was added to the mixing SNR: $\eta = SNR + \varepsilon$, where ε ranged from -15 to 5 dB with increments of 5 dB. The sum η constitutes a relative criterion that is based on the observation that the IBM is widely unaffected by the covariation between the local criterion and the mixture SNR (Kjems et al., 2009). The IBM as well as the cepstrum method of Chapter 2.3.4, using identical algorithmic parameters, were applied

in the present study. A couple of algorithmic modifications were introduced to yield general results with the present study. The first modification relates to the recursive smoothing time constant, α_{PSD}, for estimating the power spectral densities during the IBM calculation (see Equations (2.3.16) and (2.3.17)). α_{PSD} was set to 8, 20 and 30 ms. Secondly, as it is common practice to average a small set of temporally adjacent cepstral coefficients before extracting the pitch estimate with Equation (2.3.23), we have analyzed the influence of two implementations: one with an averaging over three temporally adjacent cepstral bins and one without averaging.
Accordingly, for 35 SNR/ε combinations per time constant and pitch estimation technique, a GA optimization was performed. The genetic search tries to optimize the time-constants α_{loE}, α_{hiE}, α_{p}, α_{n} and the maximum suppression A_{min}, i.e. the lower bounding of the mask. The initial population was performed on 1000 sets of randomly chosen parameter sets and the GA was terminated after 500 cycles.

Regarding the choice of an objective function for speech intelligibility, we proceeded as follows. With the aim of finding optimal time constants of the cepstral smoothing operation, one should test the processing on two perceptual scales. One dimension should reflect to what extent the perceptual effect of musical noise is reduced. For that purpose Scholz (2008) developed a multidimensional quality measure that predicts the perception of musical noise. On a second dimension, and in the context of the present work, the influence of cepstral smoothing on speech intelligibility has to be analyzed. Since the first scale could not be studied in the scope of this work, we relate the optimization of the cepstral smoothing technique solely to speech intelligibility.
In order to determine the optimal cepstral time constants in terms of speech intelligibility the STOI measure of Taal et al. (2011a) has been applied. The STOI measure shows a high degree of conformity with subjective data of intelligibility for IBM processed speech. A comparison with the I3 measure of Kates and Arehart (2005b) is not given here, as the I3 measure has not been designed for IBM processed speech (see Taal et al. (2011b) and the discussion in Section 4.2 for further information).

The averaged results of the cepstral smoothing optimization for every η are given in Figure 5.5. There, the outcomes are compared to the lower bounded IBM processed speech, specifically $\mathcal{M} = \max[\mathcal{M}_{\mathrm{IBM}}, 0.1]$. Overall, the results indicate no decline in speech intelligibility after the application of cepstral smoothing. The different SNR mixing values produce conformal intelligibility plateaus, with maxima depending on the local criterion. With respect to the cepstral time constants in the range of $\eta = 0$ dB, the optimization shows that only very small quantities of smoothing are allowed if the objective is pure speech intelligibility enhancement.

Concerning the power spectral density averaging constant, α_{PSD}, for estimating the IBM, we find that a higher α_{PSD} results in a small decrease of the pitch smoothing, α_{p}. At the same time, we observe an increase of the α_{hiE} constant towards higher η. No clear dependence of α_{n} on α_{PSD} can be observed and A_{min} is found independent of α_{PSD}.

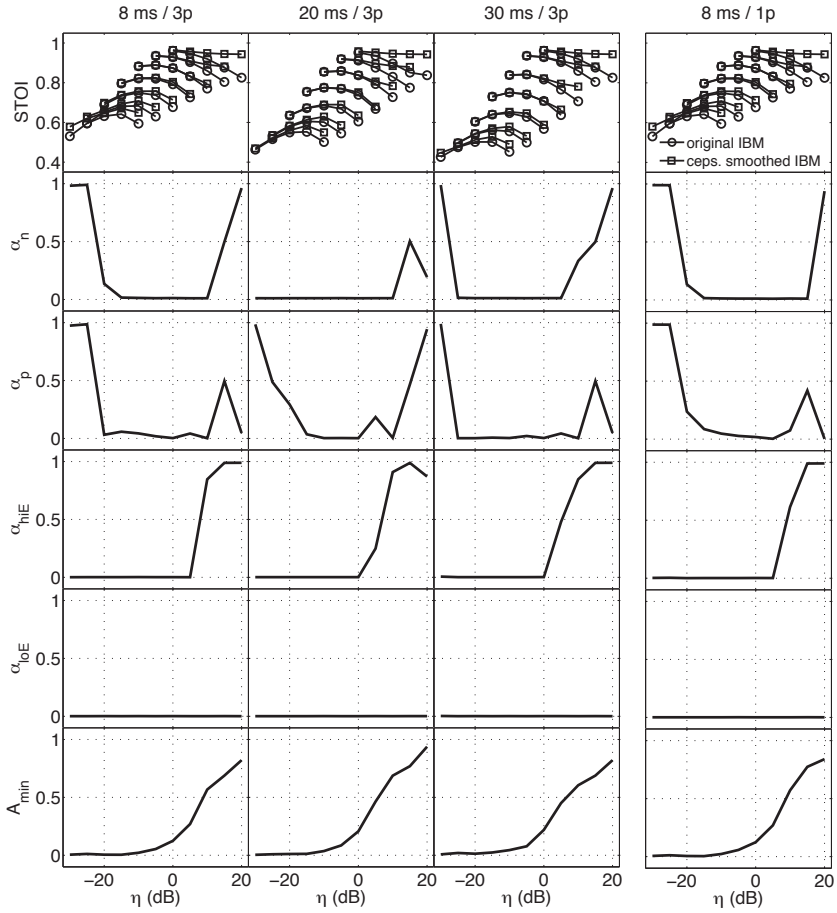

Figure 5.5: *The upper row gives the intelligibility prediction of IBM processed speech-in-noise mixtures in comparison to cepstrally optimized and smoothed IBMs for the same mixtures, using the STOI measure. The plots below present the averaged optimized cepstral smoothing constants for a direct DFT/IDFT analysis/synthesis and IBM speech enhancement approach with different time constants of* α_{PSD} *(8, 20 and 30 ms, see titles). The right-hand column gives the results of the pitch estimation technique without averaging temporal adjacent quefrency bins (*1p*), whereas the three columns at the left-hand side used an averaging over three adjacent bins (*3p*).* η *is a relative SNR criterion used in the calculation of the IBMs.*

The comparison between the different pitch estimation techniques, i.e. using one or three adjacent time bins, is limited to a medium difference between the respec-

tive pitch smoothing constants, α_{p}. Although the average parameter curves are conformal, the averaged pitch smoothing constant is higher if no averaging of the cepstral power density estimates is performed. Omitting the averaging may have a detrimental effect on the cepstral smoothing technique in unvoiced passages. Yet, overall, a negligible quantitative difference is found in terms of the STOI measure. The observation that the pitch estimation technique mainly results in changes of the pitch smoothing constant α_{p}, is an indication of the cepstral separation power of different signal components.
Furthermore, the study shows a couple of global trends. Regarding high SNR/ε combinations, a strong growth of the α_{hiE}, α_{p}, α_{n} and A_{min} is observed. This means that A_{min} is the most critical parameter. It clearly shows the difficulty of the varying gain-based speech intelligibility enhancement task. Only a small suppression of the interference is found to be critical for maintaining a certain level of intelligibility or for achieving a slight improvement. An interesting finding is the increase of the α_{hiE} smoothing constant, which might be an indication for an improved tracking of speech components that form fine-scale articulatory contours. This feature of the cepstral smoothing technique was recently used for the speech intelligibility assessment of nonlinearly processed speech (Schlesinger, 2012).
On the other hand, at low SNR/ε combinations, only a strong growth of the α_{p} and α_{n} parameters results in an improvement of the STOI prediction. At low mixing SNRs, the IBM unity gain strongly deviates from the actual energy proportion of the target signal in the mixture. The fluctuating signal parts at higher quefrencies are smoothed through α_{p} and α_{n}, while the envelope of the target signal is maintained through low values of α_{loE} and α_{hiE}. Consequently, the envelope is shown to be very important for intelligibility, while an increased smoothing in the higher quefrency bins leads to a suppression of the interference.

To summarize, the reduction observed here of the cepstral smoothing constants in the intelligibility optimization problem typifies the discrepancy between quality and speech intelligibility at the outmost extremes of the trade-off. As mentioned before, a quality dimension of musical noise perception should be introduced to balance the opposing aims carefully.

In Section 5.4.5, it will be studied whether cepstral smoothing constants that have been determined heuristically in a quality optimization, generalize to different acoustical setups. Furthermore, it will be analyzed to what extent cepstral smoothing affects speech intelligibility and quality at the output of a binaural speech processor.

5.4 Assessment of binaural speech processors

This section presents the assessment of binaural CASA processors in changing acoustics. We are addressing several questions in this study. First, we want to clarify what the benefit is for speech intelligibility, of a particular front-end/post-processor combination in a certain interference setup. Secondly, we want to analyze whether the

optimized parameter sets of the binaural post processors always guarantee a benefit in changing acoustics. Thirdly, we want to study the influence of the front-end on the enhancement power of the post-processors.

Throughout the following analysis of coherent and incoherent interference conditions, the speech material consisted of 45 s of concatenated sentences, for both the target and coherent noise signals taken from the TNO (2000) corpus. The material was spoken by a female and two male speakers of Dutch origin and in the Dutch language. The male/female proportion was set to 2/3. The sentences were RMS equalized and pauses, defined as the RMS level of −50 dB in frames of 10 ms relative to the overall RMS level, were excluded. The spatialization of the target speaker at zero azimuth and the interferers at different azimuths, as well as the ear-level SNR mixing of those, or alternatively the mixing of real-world background recordings, was performed in accordance with the method presented in Chapter 3.2.
As regards the optimization setups prior to this analysis, (exclusively) GA results at a mixing SNR of 0 dB and the concomitant weighting functions were applied (see Chapter 3.3.3). From several GA optimization runs for a specific front-end/post-filter and scene combination at an SNR of 0 dB, an arbitrary parameter set was selected for further investigation.
Based on preliminary inspection and the results of the GA optimization, a selection of CASA processors for a particular interference environment was made. Whereas all processors were used in diffuse noise conditions, the study of coherent interference conditions was limited to the application of algorithm CLP and ELT. In a preliminary analysis, algorithm CC was found to be applicable only for the suppression of incoherent noise. This is obvious, considering that the binaural waveform coherence at zero lag (i.e., the weighting function of algorithm CC) is a poor indicator to distinguish between frontal and lateral sources.
To limit the scope of the entire study further, no across environment analyses were performed for particular front-end/environment-optimized parameter sets.
As a general aim, acoustic scenes were chosen to analyze the algorithms of speech enhancement across the spectrum of possible interference conditions. Care was taken to omit redundancy. Therefore, only a subset of the scene/front-end combinations that are introduced in Chapter 3.2 are analyzed.

5.4.1 Canteen environment

The canteen environment is the first condition under which the three binaural processors CC, CLP and ELT are applied to the output of several front-ends, in order to enhance speech intelligibility. First, a GA optimization of the parameter sets, at an SNR of 0 dB in the canteen environment of the analyzed front-end/post-processor combinations was executed. The results of this parameter optimization have been partly presented in the previous section. In the following, these respective front-end/post-processor parameter sets were held constant during the assessment, in a level range between −10 and 10 dB with increments of 5 dB. Figure 5.6 gives the

results. Our observations are classified with respect to the above-mentioned questions of this section.

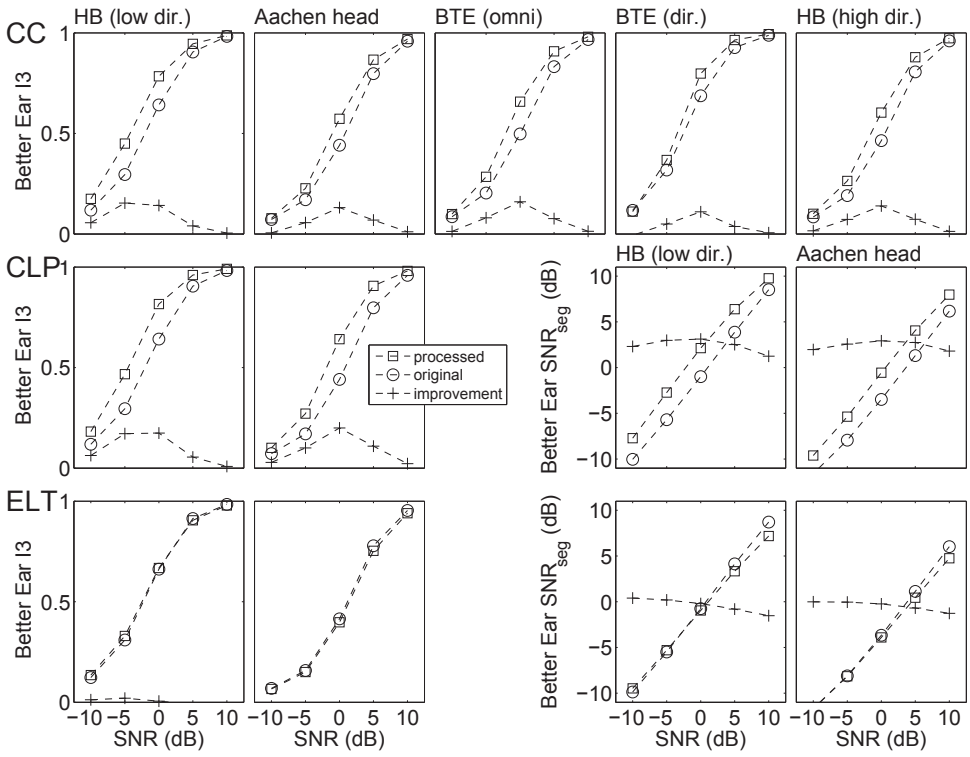

Figure 5.6: *Speech intelligibility enhancement of different front-end/post-processor combinations in the canteen environment at different mixing SNRs, assessed with the Better Ear I3 measure. Additionally, the intelligibility weighted Better Ear* SNR_{seg} *is given for algorithm* CLP *and* ELT*. The target speaker was fixed at 0 deg throughout the assessment.*

Improvement of speech intelligibility: The predicted absolute improvement of speech intelligibility in the canteen situation is in the range of 15 % at the output of algorithm CC. The speech processor CLP slightly increases the improvement to maximally about 20 %, as already concluded from the Figures 5.2 and 5.3. No improvement is obtained by algorithm ELT.

Robustness: Apart from algorithm ELT, each algorithm operates best around an SNR of 0 dB, i.e. the condition for which the post-processors were optimized.

Hence, generally, towards lower and higher SNRs the intelligibility gain declines. A small deviation from this rule is observed for algorithm CC at the output of the HG (low directivity). Algorithm ELT is essentially switched off by a low compression/expansion value (see the optimization results in Figure 5.4) and therefore does not affect the original level of speech intelligibility.

Effect of front-end: None of the particular front-ends show a marked benefit at the ear-level SNR. This includes the comparison between directional and non-directional front-ends. However, we have to recall that the SNR at the source-level is much lower in this comparison when using the directional front-ends. Therefore, the front-ends allow for the successful application of the post-processors under SNR conditions smaller 0 dB, measured at the source-level. With respect to the optimal performance range, a small advantage is seen for the HG (low directivity) front-end, likely due to the more favourable Better Ear I3 values of the unprocessed conditions. Serially connected to this front-end, the processors CLP and CC have their optimal performance over an SNR range of −5 to 0 dB.

5.4.2 Workshop environment

The workshop environment offers the highest diffuseness in this study. Consequently, the performance of the coherence-based algorithm CC should increase.
With respect to the front-end recordings in this environment, a recording failure led to only one channel being recorded with the HG in the low directivity and high directivity mode. The monaural recording was subsequently applied to both channels and made binaural by a temporal decorrelation between the channels. Although this approach does not reconstruct the original spatial scene, preliminary listening gave the impression of a well externalized scene, meaning the spatial events are located outside the head. Therefore, it has been included in the present front-end/post-processor comparison. The results have, of course, to be considered with care.
The optimized parameter sets were chosen for each front-end/post-processor in the workshop condition at an SNR of 0 dB and held constant throughout the assessment.

Figure 5.7 gives the results. What follows are our observations.

Improvement of speech intelligibility: Algorithm CC enhances the speech intelligibility estimation by maximally 25 % in the 0 dB SNR condition. The improvement tapers off towards lower and higher SNRs. Algorithm CLP shows an improvement at the output of the HG (low directivity) of maximally 10 %, and even less at the output of the Aachen head. Likely due to the intensified statistical penalty mask processing in algorithm ELT (the GA optimized parameters, not shown here, are comparable to the canteen-optimized parameters in Figure 5.2), a small increase of objective speech intelligibility is observed.

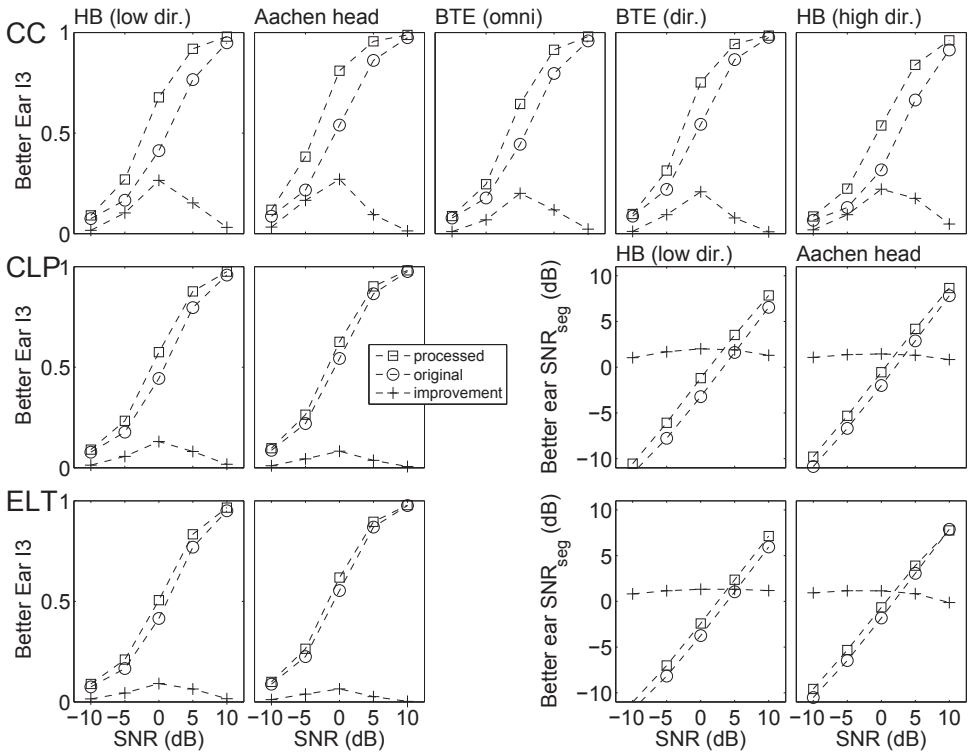

Figure 5.7: *Speech intelligibility enhancement of different front-end/post-processor combinations in the workshop environment at different mixing SNRs, assessed with the Better Ear I3 measure. Additionally, the intelligibility weighted Better Ear* SNR_{seg} *is given for algorithm* CLP *and* ELT*. The target speaker was fixed at 0 deg throughout the assessment.*

Robustness: All processors show a high degree of robustness through varying SNR conditions. In all cases, the maximum speech intelligibility improvement is reached under the 0 dB SNR condition.

Effect of front-end: No difference between directional and non-directional front-ends is observed at the output of algorithm CC. However, when using the BTE as front-end, the benefit of algorithm CC is slightly lower.

The above-mentioned temporal decorrelation between both channels of the HG recordings of the workshop scene shows to have at the most a small influence on the overall algorithmic performance (cf. the Aachen head at the output of algorithm CC). Implicitly, this observation also supports the assumption of high diffuseness of

the workshop scene recordings.
Among the binaural speech processors studied, it has been shown that algorithm CC is best suited for the suppression of the diffuse interference. Algorithm CLP and ELT, which are designed to suppress coherent interference from lateral directions, are rather ineffective in the suppression of diffuse interference.

5.4.3 One and two interferers in an anechoic environment

The next analysis examines the performance of the directional processors CLP and ELT in the presence of one, or two coherent interferers. The parameter sets of the previous optimization at an SNR of 0 dB and one interferer at 90 deg in Chapter 5.2.4 were used for the following assessment of algorithm ELT. For the execution of algorithm CLP, the two-interferer parameter set optimizations at an SNR of 0 dB of Chapter 5.2.3 were used. The reason for this difference with respect to the optimization setup has been given in Chapter 5.2.4.
We first analyze the results of algorithm CLP in the presence of one coherent interferer at different spatial positions. The outcomes are given in Figure 5.8.[5] Our observations on the above-introduced items, are given below.

Improvement of speech intelligibility: Speech intelligibility increases considerably across lower SNRs and noise azimuths. At many interference angles and mixing SNRs below 0 dB, an improvement of more than 40 % is observed. The optimal working range shows to be between a mixing SNR range of 0 and −10 dB for the directional receivers, and below −5 dB for the omni-directional receivers.

Robustness: If the unprocessed speech intelligibility is not too high, and if the interferer angle does not coincide with the target speaker, algorithm CLP generates a considerable improvement of speech intelligibility. Overall, the relative Better Ear I3 improvement (third row in Figure 5.8) shows no decline of speech intelligibility of the processed signal below a mixing SNR of 5 dB for all front-ends. An exception is seen for one angular interference position at −90 deg at the output of the BTE in the directivity mode.

Effect of front-end: The directionality of the front-end impacts several aspects found in the post-processing of algorithm CLP. Without regard to particular

[5] The figures in the remainder of this chapter are contour plots in order to observe the algorithmic performance of the binaural speech processors throughout a great many of spatial conditions and mixing SNRs. The ordinate of the subplots gives the mixing SNR, whereas the abscissa gives the location of the noise azimuths. The introduced Better Ear I3 measure comprises a range between 0 and 1 for no and full speech intelligibility, respectively. If the above-mentioned Mean Ear Q3 parameter is given, the scale from 0 to 10 renders the range from no speech quality to full speech quality. To evaluate the influence of reverberation or to facilitate the comparison with other studies, additionally the intelligibility weighted Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ is given.

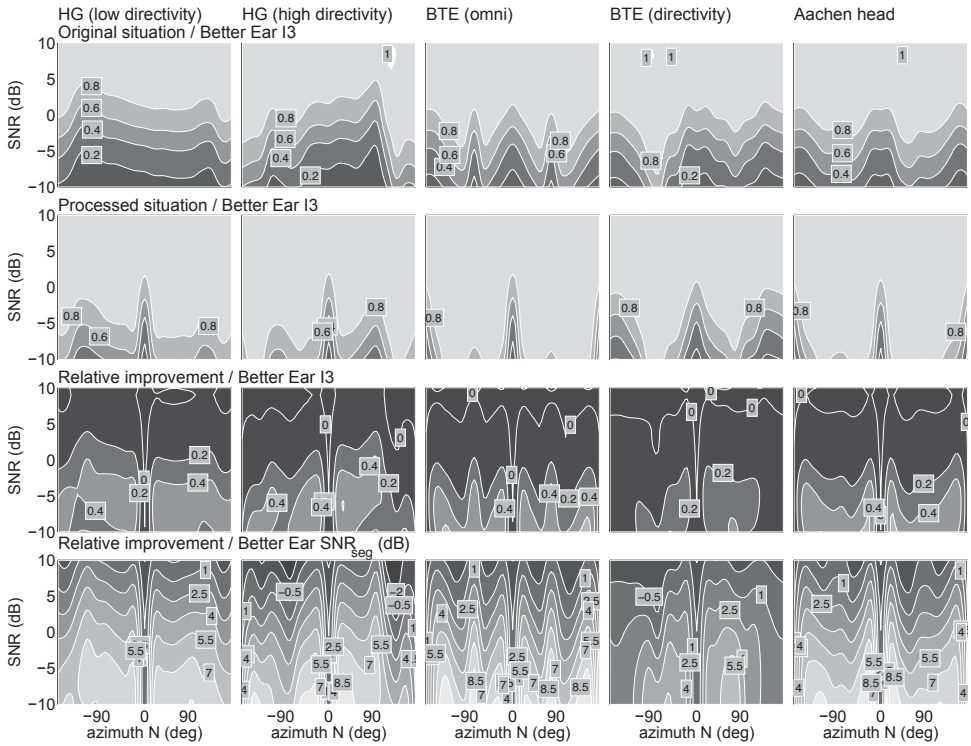

Figure 5.8: *Assessment of the speech intelligibility improvement of the speech processor CLP in the presence of one azimuth-variant interferer* N *and at the output of different front-ends (see the headings). The target speaker was fixed at 0 deg throughout the assessment. The improvement is assessed with the Better Ear I3 measure and the intelligibility weighted Better Ear* SNR_{seg}*. See the introduction of Chapter 5.4.3 for further explanation of the contour plots.*

exceptions, it can be seen that the performance of the binaural processor CLP is mainly determined by the ear-level SNR. A second dependency on the shape of the binaural parameters can be inferred.

Generally, it can be observed that omni-directional front-ends give a symmetrical intelligibility pattern with respect to the frontal plane (see the upper row in Figure 5.8). As already mentioned, the optimal working point of the CLP processor resides at higher mixing SNRs, when applying directional front-ends. Thereby a more azimuth-independent processing is observed.

Compared to the omni-directional BTE, the pinnae of the Aachen head introduce an advantage for lateral interference at ±90 deg in the unprocessed

case (first row in Figure 5.8). This observation might by explained by the diffraction effects of the head. At the contralateral BTE a pressure antinode exists at higher frequencies as a consequence of the frequency-dependent wave diffractions around the head. Following this rationale, the head shadow effect is reduced and speech intelligibility declines. This phenomenon is also observed with the Aachen head mannequin and the ear simulator, but smaller and rather flattened, possibly due to multiple diffractions within the pinna.

The directional hearing aids show an intelligibility pattern that is only roughly symmetrical with respect to the frontal plane. In fact, there is an asymmetry of speech intelligibility as a consequence of small directional differences among the bilateral beamformers, anatomical differences between the pinnae, fitting variation and interactions. The results of the original situation (see the top row in Figure 5.8) should be qualified, however, such that a source-level SNR definition would increase speech intelligibility in the frontal direction relative to lateral and retral directions, depending on the directionality of the beamformers. The application of the post-processor illustrates that the performance is mainly determined by the ear-level SNR, where the asymmetry of the input intelligibility is roughly maintained.

Due to the symmetry of binaural cues with respect to the frontal plane, i.e. the cone-of-confusion phenomenon in three dimensions, there is hardly any advantage in terms of binaural unmasking if the target speech is coming from the front and the noise is coming from the back, an observation that has also been made by Bronkhorst (2000). For bilaterally applied omni-directional receivers, this outcome, using the Better Ear I3 as a figure of merit, is generally reproduced, due to the absence of the head shadow effect.
The application of processor CLP to omnidirectional front-ends can hardly compensate for this effect, as the binaural cues of frontal and retral incidence resemble each other. Nevertheless, the results indicate that the confusion area has shrunk, i.e. it takes up a smaller angular volume after application of the post-processing to the BTE (omni-directional) and the Aachen head signals.
A different observation is made for the directional front-ends. Due to the fact that binaural level cues differ for fontal and retral incidence at the output of the HG front-ends applied here, there exists a spatial unmasking benefit (using the Better Ear I3 as a figure of merit) for coherent interference from the rear at 180 deg (top row in Figure 5.8). Moreover, as depicted in the second row of Figure 5.8, the CLP processor can exploit the difference of binaural cues between frontal and retral incidence, and consequently reduces the interference from behind the head.

Considering the average benefit of the algorithm CLP, a special case in this series of tests is formed by the results of the BTE (directional) front-end/CLP combination. Ambiguous binaural cues of the front-end, as shown in Figure 3.3, and a limited and binaurally unsymmetrical frequency transfer, as

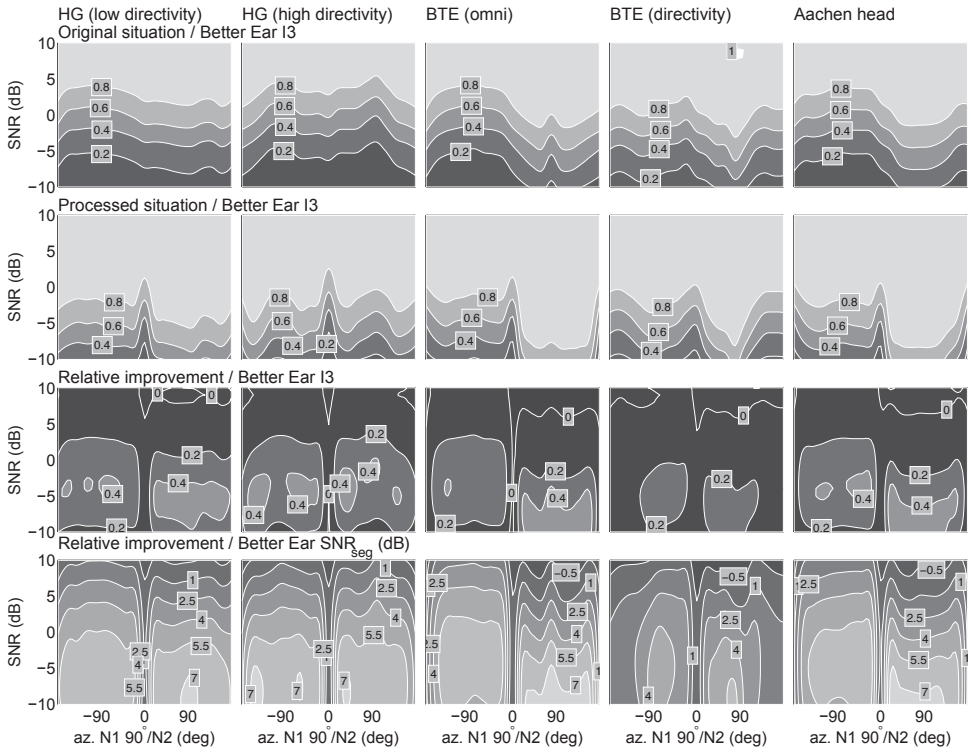

Figure 5.9: *Assessment of the speech intelligibility improvement of the speech processor CLP at the output of different front-ends (see the headings) in the presence of one static interferer (*N1*) at 90 deg and one azimuth-variant interferer (*N2*) in the range of −180 to 180 deg. The target speaker was fixed at 0 deg throughout the assessment. The improvement is assessed with the Better Ear I3 measure and the intelligibility weighted Better Ear* SNR_{seg}*. See the introduction of Chapter 5.4.3 for further explanation of the contour plots.*

shown in Figure 2.4, may account for the difficulty to enhance speech intelligibility equally well with the other front-ends. The probability-based weighting function is shown to adapt the CLP processing to the asymmetries of the front-end, however, optimally.

In the following analysis, the CLP processor is applied to the simultaneous suppression of a directionally invariant interferer N1 at 90 deg and an azimuth-variant interferer N2 that rotates across azimuths between −180 and 180 deg. The interferes are combined at an SNR of 0 dB, prior to the mixing with the target signal. Again, optimized algorithmic parameter sets of front-end/CLP combinations at a mixing

SNR of 0 dB are chosen. The results are given in Figure 5.9.

Generally, the findings with respect to speech intelligibility enhancement, robustness and the influence of the front-end correspond to the single interferer situations. Therefore, a detailed analysis is omitted. A difference is observed, however, with respect to the symmetry of the pre-processing and post-processing. If the interferers are located on opposite sides of the head the particular head-shadows cancel and speech intelligibility gets lower. Consequently, if both interferers are on the same side of the head, the head-shadow leads to an improvement of speech intelligibility. Furthermore, as a result of the increased acoustical complexity, the decrease of disjointness of sources in the time-frequency domain and the increased distortion in the estimated CLP weighting gain, the benefit of the enhancement is 5 to 10 % lower when compared to the single interferer test.

The next study assesses speech intelligibility at the output of algorithm ELT, in the presence of one and two simultaneous coherent interfering sources. The comparison between the front-ends is confined to a differentiation of the HG (low directivity) and the Aachen head. Figure 5.10 gives the results. The scene setups are identical to the previous experiments. Our observations are compared with the results of algorithm CLP, and are summarized under the categories of speech intelligibility enhancement, robustness and the influence of the front-end.

Improvement of speech intelligibility: On the whole, the performance of algorithm ELT is comparable to the performance of algorithm CLP. For most angular positions and lower mixing SNRs with a single interfering source, a speech intelligibility improvement of up to 40 % is observed. The optimal working range of algorithm ELT generally resides below an SNR of -5 dB and 0 dB at the output of the Aachen head and the HG (low directivity), respectively, in the single interferer case (left-hand columns in Figure 5.10).
Also in the presence of two interferers, algorithm ELT shows a robust improvement of speech intelligibility, a bit less powerful than algorithm CLP. The coarser directional lookup histogram resolution of algorithm ELT in comparison to algorithm CLP, as well as the elusive nature of envelope parameters in noise may account for this deterioration in performance.
The broader beamwidth of the ELT processor, constitutes another difference to algorithm CLP. This outcome is most likely a result of the lower directional resolution of envelope parameters. Furthermore, it has to be remembered that mainly the envelope ILD is used in the classification process, which has been shown to offer much less directional distinctness than the carrier IPD in psycho-acoustic tests (Stern et al., 2006).

No general advantage due to the modulation filterbank, ideally giving a co-modulation unmasking release gain, is observed. If the SNR_{seq} gain of algorithm CLT is subtracted from the gain of algorithm ELT (data not shown here), only at azimuths smaller than $-90°$ and higher than $90°$ a small but consistent improvement is found. However, the overall lack of a co-modulation

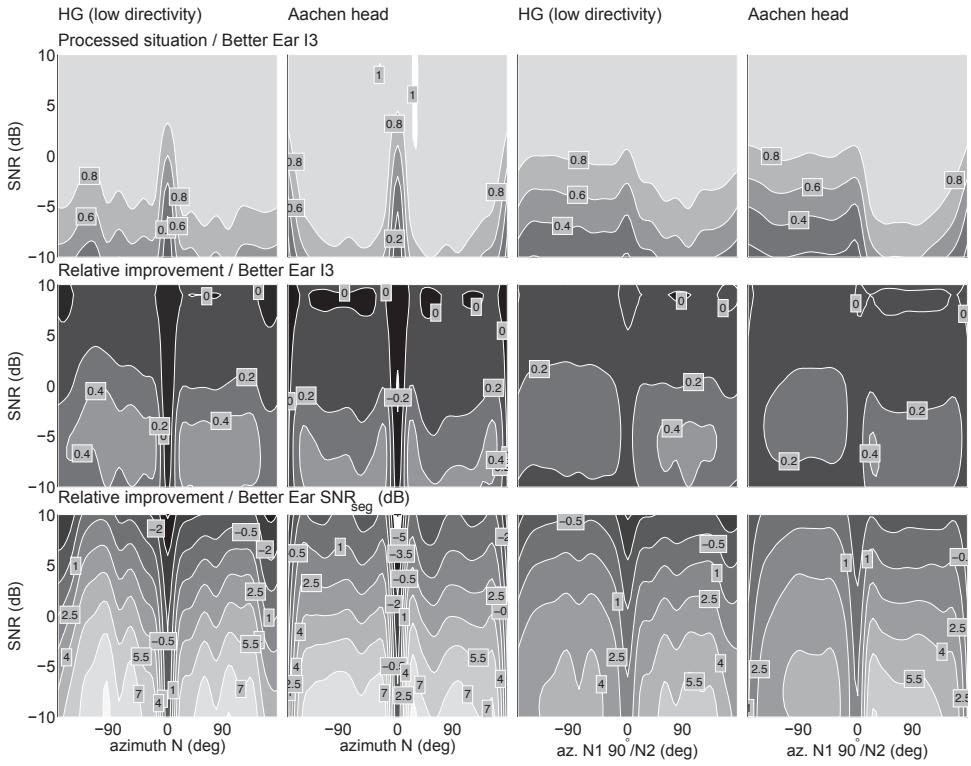

Figure 5.10: *Assessment of the speech intelligibility improvement of post-processor ELT in the presence of one static interferer at 90 deg, left columns, or one static interferer (*N1*) at 90 deg and one azimuth-variant interferer (*N2*) between −180 and 180 deg, right columns, at the output of different front-ends (see the headings). The target speaker was fixed at 0 deg throughout the assessment. The improvement is assessed with the Better Ear I3 measure and the intelligibility weighted Better Ear* SNR_{seg}*. See the introduction of Chapter 5.4.3 for further explanation of the contour plots.*

unmasking effect in comparison with algorithm CLP might be explained by the fact that the modulation spectra are not sufficiently resolved in order to exploit the combination of monaural and binaural cues optimally (see the discussion at the end of this chapter).

Robustness: The robustness of algorithm ELT is high throughout all conditions analyzed and comparable to algorithm CLP. The somewhat fluctuating Better Ear I3 dependency on the azimuth of the interferer (top row in Figure 5.10) might be caused by modulation transfer functions that to some degree overlap

with the modulation transfer-function of the target (see Figure 3.12 in Chapter 3.3.2).

Effect of front-end: Considering the directivity of the front-ends, an increased performance in the two-interferer situation is observed at the output of the HG (low directivity) front-end. Consequently, binaural parameters of the envelope at the output of the HG (low directivity) offer a better means to separate sources. To a lesser extent, a similar difference is seen in the single interference test, although the optimal working point is shifted towards lower SNRs, at the output of the mannequin.

5.4.4 Coherent interference in reverberation

In the final speech intelligibility enhancement test, the three binaural speech processors CC, CLP and ELT are assessed and compared in the MISM-simulated environment with one azimuth-variant interferer and a reverberation time of 0.2 s. The GA optimizations in the previous sections have already shown that it is only algorithm CLP that achieves a moderate enhancement in terms of $\mathrm{SNR}_{\mathrm{seg}}$ in the MISM-environment at an SNR of 0 dB. The other two binaural speech processors have demonstrated to gain only a small Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ increase in the range of 0.5 to 2 dB in the same environment.
The present analysis now questions whether the respective binaural processors work robustly for different interferer azimuths and different mixing SNRs, even though the Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ gain is low in the condition for which the algorithms are optimized. As mentioned above in Chapter 4.2.4, the results are assessed with the intelligibility weighted Better Ear $\mathrm{SNR}_{\mathrm{seg}}$, because the Better Ear I3 measure is not suitable when the target speech is distorted of reverberation. However, the $\mathrm{SNR}_{\mathrm{seg}}$ merit is also biased due to the influence on reverberation and cannot directly be translated to speech intelligibility. Therefore, the results of the current test have to be looked at carefully.
Similarly as in previous studies, GA-optimized parameter sets at a mixing SNR of 0 dB were chosen for the present assessment of each algorithm.

Figure 5.11 gives the results of speech enhancement in the simulated environment at the output of the Aachen head.

It can be seen that the relative Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ improvement is evenly distributed over noise angles and mixing SNRs. A laminar distribution of the Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ proves the robustness of all three binaural speech processors. Moreover, the processors CLP and ELT are able to enhance the Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ at lower mixing SNRs. The CLP processor offers the largest gains with more than 4 dB at mixing SNRs smaller than -5 dB. A small improvement of the Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ is observed at the output of processor CC at lateral positions, thereby showing a coarse directional differentiation by the binaural waveform coherence at zero lag.
Although the Better Ear $\mathrm{SNR}_{\mathrm{seg}}$ reflects the functioning of the processors at low

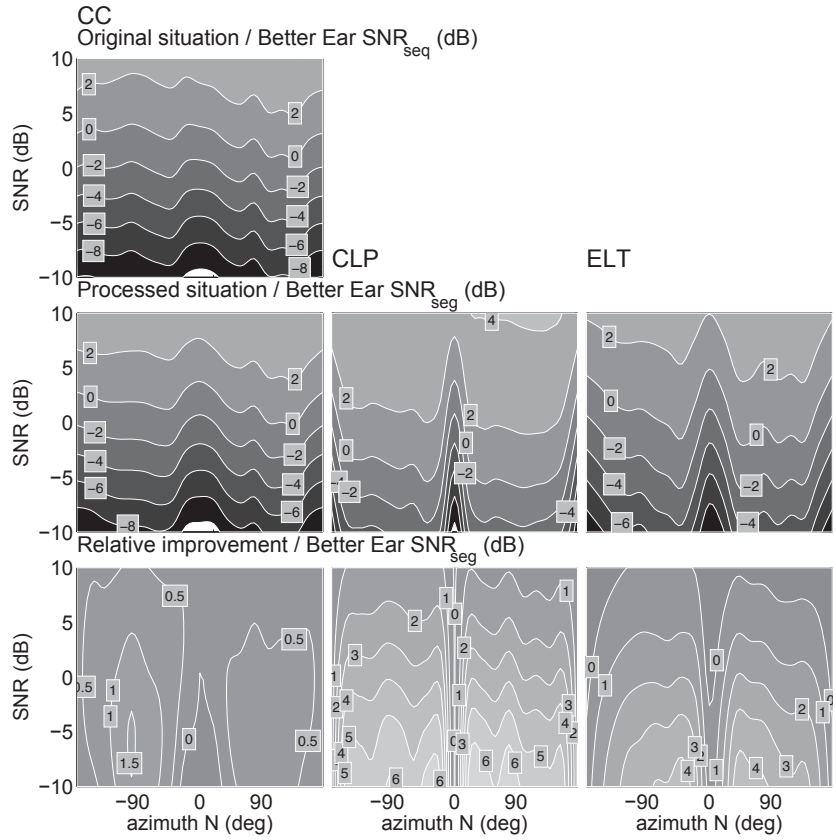

Figure 5.11: *Assessment of the speech enhancement of the binaural speech processors* CC, CLP *and* ELT *in the MISM-simulated environment with an RT of 0.2 s and an azimuth-variant interferer. The target speaker was fixed at 0 deg throughout the assessment. Both the target and the interfering speaker are in one meter distance from the Aachen head, which served as a front-end, while the reverberation radius is at 1.2 m. The algorithmic improvement is assessed with intelligibility weighted Better Ear* SNR_{seg}*. See the introduction of Chapter 5.4.3 for further explanation of the contour plots.*

mixing SNRs, it is unclear whether these enhance speech intelligibility or whether the trade-off between speech enhancement and distortion is inclined towards the latter.

5.4.5 The application of cepstral smoothing for quality enhancement

In the final study in this section, the musical noise suppression through the above-introduced cepstral smoothing technique is applied to the output of the binaural speech processor CLP and assessed.
A problem with assessing the impact of musical noise on intelligibility and quality is the fact that the objective measures of the present work were neither developed nor tested with the cepstral smoothing processing technique. Consequently, the outcomes of the following analysis have to be considered with caution.

We first briefly recapitulate the GA optimization of cepstral smoothing constants of Chapter 5.3. The results of the optimization have been given in Figure 5.5 for different algorithmic settings. For the purpose of the optimization, IBM processed speech has been subjected to cepstral smoothing. Instead of the I3 measure, which has been the standard intelligibility measure in this section, the STOI measure has been applied as an objective function in the GA optimization. The choice has been made to prevent an I3-related overestimation of a phase-loss to speech intelligibility, which has been observed with IBM processed speech at low SNR mixing SNRs (Taal et al., 2011b).
As a general optimization result, the STOI measure indicates no deterioration of speech intelligibility for the smoothed IBM masks, using different mixing SNR and mask-based local criteria combinations and a series of algorithmic variations. In an attempt to extract general cepstral parameter results from all IBM mask optimizations, the mean of the respective cepstral smoothing constants and the maximum mask attenuation has been obtained by averaging over all parameter results at particular mixing SNRs and mask-related local criteria combinations. For mixing SNRs and mask-related local criteria around 0 dB, the cepstral smoothing constants are essentially reduced to zero.

In the present analysis, the cepstral smoothing technique is applied to the soft-masks of algorithm CLP. Following the reasoning for the applicability of the Better Ear I3 and the Mean Ear Q3 measure in soft-mask speech enhancement tests, given in Chapter 4.2, the following analysis is based on these two metrics. It remains questionable, however, whether these measures reflect significant perceptual relevance of the cepstral smoothing technique.

Since it would render the effect of the cepstral processing useless if one applies the small-valued optimized cepstral smoothing constants of Chapter 5.3 around SNRs and mask-based local criteria of 0 dB, a set of heuristically found smoothing parameters are applied, which generate a considerable improvement of speech quality. The parameters are identical to the ones given in the lower right-hand plot of Figure 2.8 and have been found in preliminary listening. Hence, the algorithmic parameters of the cepstral smoothing technique are $\alpha_{\mathrm{loE}} = 0$, $\alpha_{\mathrm{hiE}} = 0.2$, $\alpha_{\mathrm{p}} = 0.3$, $\alpha_{\mathrm{n}} = 0.8$. Additionally, $A_{\mathrm{min}} = 0.1$. Consequently, by means of the application of an intense cepstral smoothing at the output of the CLP processor, the proof-of-concept shall

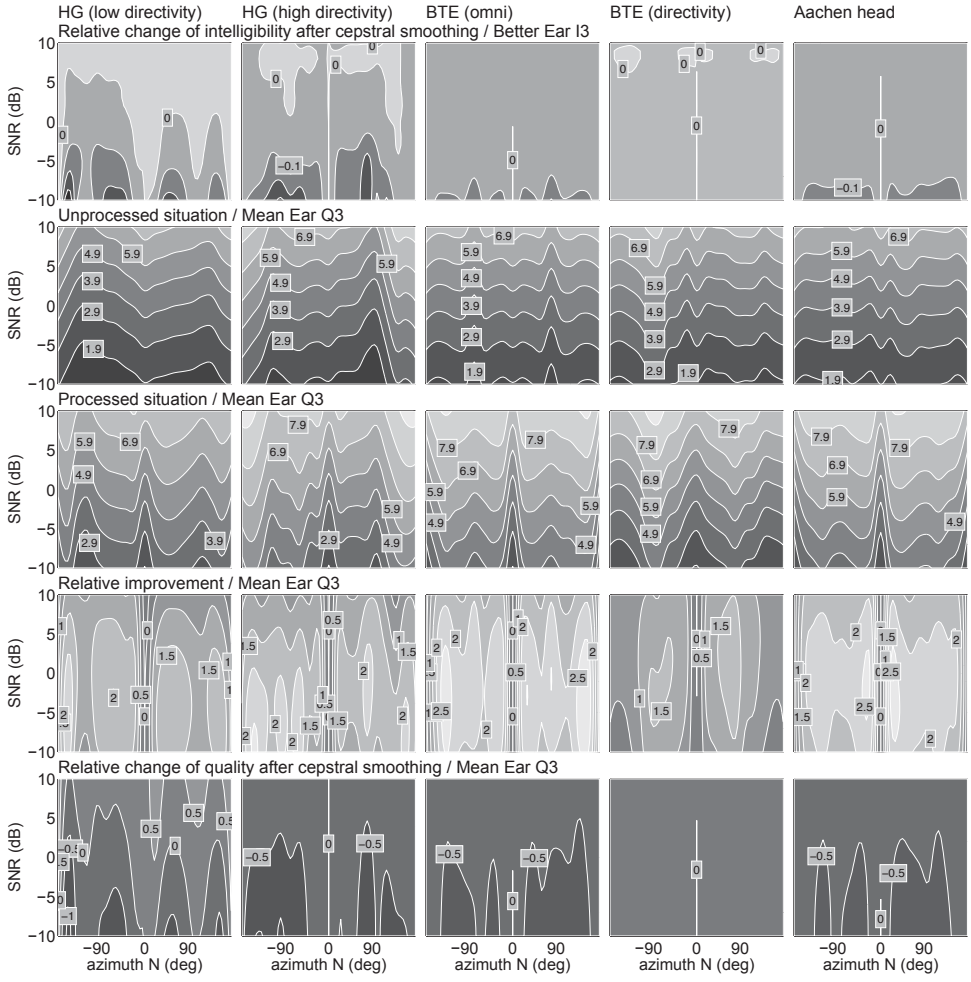

Figure 5.12: *Speech intelligibility and quality assessment of different front-end/CLP processor arrangements (see headings) with and without the cepstral smoothing technique, and as a function of mixing SNRs and noise-azimuths* N. *The target speaker was fixed at 0 deg throughout the assessment. The instrumental evaluation is based on the Better Ear I3 and the Mean Ear Q3 measure. See the introduction of Chapter 5.4.3 for further explanation of the contour plots.*

be given by studying the algorithmic robustness in terms of speech intelligibility and quality. To that purpose, the combined processing scheme was applied in an anechoic environment with a single azimuth-variant interferer. The CLP processor

used the optimized parameter sets of the two-interferer setup at ±90° and an SNR of 0 dB of Chapter 5.2.3.
The results of the musical noise suppression technique are given in Figure 5.12 and are compared with the outcomes of algorithm CLP without cepstral smoothing.

The first row of Figure 5.12 depicts the Better Ear I3 difference for the CLP enhancement with the downstream application of cepstral smoothing, relative to the CLP output without cepstral smoothing. A moderate decrease of speech intelligibility is seen at lower mixing SNRs. Despite that, at some spatial locations of the interference the decrease of predicted speech intelligibility can be as high as 20 %, especially at the output of the highly directional HGs at low mixing SNRs. However, a small increase of intelligibility is also observed at higher mixing SNRs. Overall, at a mixing SNR of 0 dB, the decrease of speech intelligibility is negligible, also for the HGs. Consequently, the algorithmic robustness of the cepstral smoothing technique—using the heuristically tuned parameters for a forceful application of cepstral smoothing—is high, and in most of the cases not detrimental to speech intelligibility.

Furthermore, Figure 5.12 presents the Mean Ear Q3 results at the output of algorithm CLP with and without cepstral smoothing. There the mean ear quality metric shows a moderate decrease of quality at lateral interference positions for most of the front-ends. The optimal working range of algorithm CLP in terms of quality is higher than found for speech intelligibility. As it was generally observed with the intelligibility improvement due to binaural processing under coherent interference conditions, the directional speech quality reveals conformal contours of the Mean Ear Q3 measure at the input and the output of the CLP processor. The difference of input and output quality at the CLP processor indicates an improvement of 15 to 25 % without cepstral smoothing applied.
Finally, the bottom row in Figure 5.12 gives the Mean Ear Q3 difference after application of the cepstral smoothing downstream to algorithm CLP, relative to the quality output without cepstral smoothing. At low mixing SNRs, a decline in quality is observed that lies in the range of 5 % and is, therefore, comparable to the loss of speech intelligibility if cepstral smoothing is applied. With respect to the influence of the front-end, however, the quality seems hardly affected.[6]

5.4.6 Discussion and conclusions

This section assessed and compared the capabilities of fifteen binaural front-end and binaural CASA speech processor combinations, under a variety of noise conditions. The improvement in terms of speech intelligibility due to the binaural speech pro-

[6] An exception from this observation represents the BTE (directivity) front-end, that has been shown to allow only for a moderate improvement of speech intelligibility in coherent interference conditions. As can be seen, the quality gain is equally affected. No decline of quality is observed when applying cepstral smoothing—possibly due to the low gain speech enhancement of algorithm CLP at the output of this front-end.

cessors at the output of different bilaterally applied front-ends with and without beamforming has been objectively predicted.

First, the hypothesis of Chapter 2 has been confirmed: binaural CASA-based speech processors are appropriate for the suppression of coherent interference. If applied as a post-filter to diffuse field optimized beamforming front-ends, a complementary working principle for a high improvement of speech intelligibility can be realized. Contrary to the suppression capabilities of up to 40 % in terms of speech intelligibility, when suppressing coherent interference, in diffuse noise the benefit is not given when applying processor ELT and small to moderate when applying the algorithms CLP and CC. Hence, none of the three binaural algorithms presented generate a high improvement of speech intelligibility under diffuse conditions. This observation was made in the tested canteen and workshop environment.
As far as a ranking of the post-filters in these conditions is concerned, it was shown that algorithm CLP performs best in the canteen environment and reaches a speech intelligibility enhancement of approx. 20 % at a mixing SNR of 0 dB. In the highly diffuse workshop environment algorithm CC is in the lead, in particular as it has shown to be only slightly affected by the directionality of the front-end at a given ear-level SNR. Algorithm ELT could not reach a benefit in terms of speech intelligibility in diffuse noise conditions. Nevertheless, the small to moderate algorithmic performance in diffuse noise conditions can become very beneficial around the 50 % intelligibility level, i.e. the SRT, because every dB of noise reduction at this level can improve the absolute speech intelligibility by 15 % (Plomp, 1978).
In anechoic and coherent interferer environments the qualifications of the respective algorithms are reversed. That is, algorithm CLP and ELT yield speech intelligibility improvements of up to 40 %, while algorithm CC is unable to give improvements under such noise conditions.

None of the post-processors presented introduce a considerable loss of speech intelligibility in any of the noise conditions analyzed. This is a result of the GA optimization of the parameter sets, in which the binaural processors have shown to adapt well to particular acoustic environments.
In addition, the processing of the optimized binaural algorithms has shown to stay robust in situations where the mixing SNR, or the directions and amount of interferers changed. However, it remains an unanswered question whether the post-processors with certain parameter sets keep their robustness across completely different environments.

With respect to typical (diffuse-field like) cocktail-party problems, an SNR of 0 dB can be assumed (see e.g., Bronkhorst, 2000). Therefore we have opted to centre the optimization and the analysis of the binaural processors around this SNR. However, compared to an estimated SNR of 0 dB in a diffuse sound field, the coherent interference setups at an SNR of 0 dB used in this chapter pose a much easier speech-in-noise task, due to the disjointness of the sources across the transform domains. As it was previously described, using an invariant SSN interference and monaural presenta-

tion, the SRT of the normal hearing is usually found at −4 to −5 dB (George, 2007). If speech and noise are spatially separated and binaurally presented, this SRT can be further reduced by −12 dB (Bronkhorst, 2000).
Consequently, the optimization results in coherent interferer conditions at an SNR of 0 dB rather reflect an SRT bottom line adjustment for the hearing impaired, than for the normal hearing. The results of the assessment reveal that the favourable working points of the algorithms CLP and ELT lie much below 0 dB, although these were optimized at an SNR of 0 dB. Future research should therefore optimize and analyze the processors at lower SNRs in coherent noise environments. By those means, it is likely that the speech intelligibility gain at low SNRs and in coherent interference environments, can further be improved.

A general dependency of the binaural post-filters' performance on the binaurally averaged ear-level SNR is a further important finding. Consequently, a directional front-end may linearly enhance speech intelligibility and a binaural post-processor will, irrespective of the actual SNR at the source-level, improve the overall intelligibility gain further, for the greatest part based on the ear-level SNR and approximately independent of the front-end.
Future tests should assess the overall benefit of a particular front-end and back-end combination. As far as the HG front-end is concerned, we may refer to the subjective and objective evaluation of Merks (2000) and Boone (2006). For instance, Boone (2006) measured an intelligibility weighted DI of 7.2 dB at the output of the HG (high directivity)[7]. As the HG are optimized in an ideal diffuse noise field, the front-end will be the main contributor to the overall benefit. Under coherent noise conditions this benefit will deteriorate and the contribution of a binaural post-processor, e.g. algorithm CLP, will increase. However, as above-mentioned, the derivation of an SNR value at the output of a nonlinear speech enhancement algorithm can not fully explain the actually intelligibility gain. In a first approximation we may conclude from the findings of the present chapter that the SNR loss of the front-end in the presence of coherent interference can be roughly compensated for by a binaural post-filter, as demonstrated by the intelligibility weighted Better Ear $\mathrm{SNR}_{\mathrm{seg}}$. As a consequence, the combined processing scheme makes few assumptions on the noise condition and is, therefore, universally applicable, provided an appropriate parameter set is used in the CASA-based post-filter. Choosing the right parameter set may only be a minor problem in future applications, considering recent progress in scene classification (Bach et al., 2011).

In conclusion, the potential of the analyzed CASA speech processors in diffuse conditions falls short of expectation, if we consider the cocktail-party performance of the model ASA process. The reason for the unsatisfactory separation was revealed in the statistical analysis of binaural parameters in noise, given in Chapter 3. There it was demonstrated that interaural differences, be it the binaural waveform or binaural envelope parameter, are severely degraded if the sound field has a diffuse

[7]The hearing aid programme is termed "high1" in Boone (2006); it provides a benefit of 7.2 dB, as assessed with the directivity index method of the ANSI S3.35- 2005 standard.

character and the mixing SNR is below or equal 0 dB. Also the binaural waveform coherence, which is applied in algorithm CC, could not efficiently be exploited for several physical reasons (see below). On the other hand, the separation power of concurrent speakers in anechoic situations is higher than provided by the model ASA process. This finding has already been made by Peissig (1992) in subjective testing, by comparing word scores of spatial configurations with and without the processing of an algorithm similar to the here applied CLP processor. Consequently, for the suppression of coherent interferers, the application of the CLP processor can be recommended.

Summary and outlook algorithm CC

The binaural waveform coherence was not the subject of a thorough theoretical analysis in this thesis. Such a study had been undertaken by, e.g. Martin (2001). In the present work, the delimiting factors of the MSC application in the source separation process have been mentioned and can be summarized as being the distance between the receivers, i.e. the distance between the ears, the reverberation radius, the radial distance of the target signal as well as the noise sources, and the time-frequency trade-off in the implementation of algorithm CC.

With the exception of a moderate intelligibility improvement in diffuse environments, algorithm CC could not deliver an improvement in coherent interferer situations. This result is in agreement with subjective tests performed by Peissig (1992), who found that algorithm CC is merely beneficial in highly reverberating conditions in terms of speech quality. In a similar fashion, Jeub et al. (2009) demonstrated a moderate gain of 1.39 dB (using a segmental SNR measure) in the suppression of reverberation by employing a binaural implementation of algorithm CC.
For the suppression of reverberation, Peissig (1992) applied a time constant $\breve{\tau}$ for the averaging of the MSC in the range of 48 to 333 ms, which is in the range of the syllable to the phoneme rate of speech (i.e. approximately from 3 to 20 Hz). The genetic optimization of the present work led to values of $\breve{\tau}$ between 10 ms and 40 ms. The Q3 measure, which was originally developed to reflect the perceptual quality perception of additive noise, centre-clipping and peak-clipping on speech quality (Kates and Arehart, 2005a), has shown to be not 'aware', i.e. not developed, for an improved quality perception at $\breve{\tau} > 35$ ms. Further research should analyze different time constants for transitional and steady-state parts in speech.

There are several possibilities to refine the MSC-based weighting approach by the calculation of the coherence at a range of (possibly auditory plausible) increments. Such an approach would establish a more complete image of the acoustical scene. Hence, it may enable the algorithmic suppression of an interferer at lateral positions. With an MSC implementation at zero lag the algorithm is generally not capable to suppress lateral sources. Nevertheless, by dint of the genetic adaption of the algorithmic parameters to a scene with coherent interference and slight reverb, see

Figure 5.11, a robust improvement of up 1.5 dB at the output of the Aachen head was achieved.
Furthermore, a multi-lag approach for calculating the binaural coherence was shown to be of use in modelling localization phenomena (Faller and Merimaa, 2004). The precedence effect (p. 253 ff., Moore, 2003), which can give an important source grouping cue, would be contained in these feature spaces. For example, Opdam (2010) extended the model of Albani et al. (1996) by a first-wavefront extraction approach and demonstrated how sources can be located correctly even in highly reverberating circumstances. Based on these possibilities, (possibly parametric) decision rules may be extracted from a multi-feature localizer and relayed to a mask-based weighting approach.

Summary and outlook algorithm ELT

The poor performance of algorithm ELT in diffuse conditions is a consequence of the application of elusive binaural envelope parameters in the source separation process. The noise susceptible nature of these parameters has been demonstrated in Chapter 3.3.2 and is corroborated by a series of related works (see e.g. Blauert, 1997). Especially time differences of the binaural envelopes have to have a strong binaural envelope coherence in order to contribute to a consistent directional image. The binaural envelope coherence shows, however, even in moderately noise-corrupted stimuli, to be strongly degraded (Rakerd and Hartmann, 2010).

Conceptually, the four-dimensional feature space (a time-varying sound power representation in a binaural carrier and modulation frequency domain) allows separation and target source enhancement in accordance with the conjectured physiological organization of the basal model ASA process (Kollmeier and Koch, 1994). However, the current means to implement this multi-dimensional feature space are likely to be too simple. Remember that algorithm ELT uses a subset of binaural parameters and, therefore, a subset of all cues that are available for the model ASA process. Moreover, the model ASA process is a mutual bottom-up and top-down process that binds available cues based on hypotheses. Given the flexibility of the ASA process, the here observed shortcomings of algorithm ELT under diffuse noise conditions constitute no surprise.

One important achievement of the present work, is the redesign of algorithm ELT such that it has been made able to suppress laterally coherent noise sources. We think this feature was not in the initial implementation of algorithm ELT of Kollmeier and Koch (1994). The main reason for this improvement lies in the analysis/synthesis window length of the STFT bandpass-filtered signal.
This finding is a result of the study of binaural parameters in modulation spectra. It turns out that the analysis window of STFT bandpass-filtered signals must be short, in order to maintain the disjointness property of concurrent sources in the modulation spectra. However, this requirement is in conflict with the fact that the

decomposition of modulation frequencies of the envelope needs longer analysis windows, in order to yield stable and well-resolved modulation power spectral density estimates. By using an analysis window length of 16 ms (Table 2.3) we have found a compromise between these two requirements.
Kollmeier and Koch (1994) opted for a frame length of 40 ms for the modulation frequency analysis. The implications of this choice are revealed by a looking at the weighting function A7 in Kollmeier and Koch (1994), which demonstrates that mainly the statistical penalty measures of the ILD and IPD are applied in the noise suppression process, instead of the binaural parameters (see Chapter 2.4.3). Our assumption is further supported if one considers that Kollmeier and Koch (1994) have assessed their algorithm at SNRs between -10 to -2 dB, in partly babble noise, with and without a reverberation time of 1.33 s, and by using anechoic ILD and IPD reference recordings (of a different artificial head than used in the enhancement setup) for the segmentation process. In the light of the findings in Chapter 3.3.2 on the statistics of binaural envelope parameters in noise, it seems justified to conclude that algorithm ELT was originally not designed to take decisions based on correct localization judgments, but rather on variations of spatial cues around the midline.
Although our implementation strictly followed the intended working principle of algorithm ELT, as it was proposed by Kollmeier and Koch (1994), it differed in a variety of functional features. As it has been mentioned, the envelope ITD was applied instead of the envelope IPD, and the envelope ILD parameter was magnified. Both actions have been brought in to increase the separation power of algorithm ELT. For the same reason, a pattern-based a posteriori probability weighting method has been incorporated. This approach, which functionally mimics top-down processes of the conjectured binaural pattern-driven model hearing, allows for an efficient adaptation to the binaural parameters at the output of a particular front-end.
The performance of algorithm ELT has shown to be on a par with algorithm CLP in single coherent interferer conditions, which, however, offers a doubled bandwidth and a higher quality output, due to directional energy-based masks rather than directional F0-based masks.

However, what has been found to be a requirement in the suppression of lateral coherent noise sources, turns out to be a disadvantage in diffuse noise conditions. Mainly because the modulation spectra are not well resolved in analysis/synthesis windows of 16 ms length, the algorithm shows an increased susceptibility to diffuse interference and renders any improvement of speech intelligibility impossible under such conditions. Therefore, the GA optimization reduces in the filtering power of algorithm ELT such that no deterioration of speech intelligibility is generated. Although, the ELT has been optimized at an SNR of 0 dB in these diffuse conditions, it has been shown that the GA optimization yields results that generalize well to the observed range of an SNR of -10 to 10 dB.
Using a higher resolution of the modulation filterbank, Kollmeier and Koch (1994) reported benefit in the range of an SNR of 2 dB in diffuse noise conditions. In a similar study, Wittkop et al. (1997) reported a moderate success under adverse

conditions with the implementation of Kollmeier and Koch (1994). Beyond that, it is not known to the author of the present work whether supplementary developments of algorithm ELT have been undertaken by the group of Kollmeier or others. However, a publication of Woods et al. (1996) reflects an effort to construct a binaural and modulation-based separation algorithm in a different way. In this work, speech enhancement was based on the combination of a binaural carrier algorithm and a monaural envelope algorithm. To this end, a binaural algorithm, similar to algorithm CLP of this work, was combined with monaural cepstral pitch detection and a modulation-based algorithm. As a result, a multi-feature representation of the binaural signal was used to obtain estimates of the target energy at a certain time-frequency bin. Consequently, two enhancement algorithms were efficiently combined and they generated a complementary benefit, as assessed with a correlation-based SNR measure[8].
Recently, Woodruff et al. (2010) advanced in the same direction, by combining binaural and several monaural cues (including a pitch estimation) for an enhanced separation performance in adverse conditions. The authors have demonstrated an SNR improvement of up to 3.6 dB (measured as the SNR of the estimated signals relative to signals produced by IBMs). As their method aims to estimate masks as a whole, meaning across a great many of time-frequency bins, it yet remains to be clarified, whether a real-time application of such an approach is realizable.

Despite the poor performance of algorithm ELT in diffuse noise, the general approach is neat in its appearance. Therefore, we think that algorithm ELT will not fall into disuse, since it provides a link between the spatial location and the pitch of concurrent speakers. Therefore, conceivably, one could use algorithm ELT in a supervised fashion for estimating the pitch when two or more speakers are overlapping on the tonotopic axis, i.e. excite the same carrier frequencies, or when speech is unvoiced and general pitch estimation methods are weak. As an example, a version of algorithm ELT was combined with versions of algorithm CLP and CC (Schlesinger and Boone, 2008). Based on the STI as an objective measure of speech intelligibility, which turned out in later works not to be well correlated with the speech intelligibility of nonlinearly processed speech (Schlesinger et al., 2010), the overlay of multiple algorithms was shown to offer the possibility for an enhanced noise suppression. In future work, a revision of such a combination appears worthwhile.

Another approach to an enhanced noise suppression in the modulation domain was recently presented by So and Paliwal (2010), in which a Kalman filter approach was monaurally applied. The Kalman approach offers the advantage to filter the amplitude and the phase spectrum based on an adaptive MMSE criterion. As it was recently shown by Paliwal et al. (2011), including the phase in the range of the low frequencies in the filtering process, can contribute to speech intelligibility. In this study Paliwal and colleagues examined the influence of the modulation magnitude and phase spectrum to speech intelligibility. The results suggest that the modulation

[8]See e.g. Goldsworthy and Greenberg (2004) for a description of this correlation-based SNR measure.

phase spectrum is more important to speech intelligibility than the phase spectrum of the waveform carrier. However, combining the modulation magnitude spectrum with the phase of the carrier before reconstructing the waveform, as it is also implemented in the ELT algorithm, was shown to enhance intelligibility. Moreover, Paliwal and colleagues found that short frames of the modulation analysis-synthesis filterbank, as short as 32 ms in their study, contribute to intelligibility, if the modulation magnitude spectrum is retained and the modulation phase spectrum is discarded. In contrast, if the modulation phase spectrum is retained and the modulation magnitude spectrum is discarded, longer frames, typically in the range of 250 ms, are needed to maintain a certain amount of speech intelligibility. In conclusion, these results give valuable guidelines for successful application of the modulation-based noise suppression approach.

In order to avoid a metallic sounding output signal, which has been observed in the present work and which was reported by others, Schimmel et al. (2007) transformed the modulation mask into a time-varying filter, with which the degraded input signal is subsequently enhanced in the time domain.

Summary and outlook algorithm CLP

Algorithm CLP is demonstrably the most efficient binaural CASA processor under all test conditions. The binaural temporal difference parameter of the carrier waveform, which is applied in the noise suppression process of algorithm CLP, has shown to be the most decisive and robust binaural criterion in noise. Psycho-acoustic tests widely support this result (Stern et al., 2006). Consequently, algorithm CLP is well suited for the directional noise suppression approach. As a result, the algorithm has demonstrated to be very beneficial in coherent noise conditions, and moderately effective in diffuse noise conditions.
We expect that the present implementation of algorithm CLP with the a posteriori probability weighting approach, as well as the GA-based parameter optimization process, outperforms the classical implementations of algorithm CLP of Peissig (1992), Kollmeier et al. (1993) and Wittkop and Hohmann (2003), which lack a pattern-based separation and the here proposed model-based optimization approach.
A subjective evaluation of the CLP processor was conducted by Gaik and Lindemann (1986) and Peissig (1992).[9] In both studies the algorithm's quality in anechoic and coherent interference conditions was demonstrated. Moreover, by a combination with algorithm CC, Peissig (1992) measured a subjective speech intelligibility improvement of up to 25 % under reverberant conditions. Therefore, and as previously mentioned, the combination of algorithm CLP and algorithm CC is a good option for future investigations.

[9]Note that Kollmeier et al. (1993), Wittkop et al. (1997) and Wittkop and Hohmann (2003) evaluated a combination of the CC and CLP algorithm. Therefore, a direct comparison with the results of algorithm CLP (or CC) is not possible.

A remark on the reverberation test setup

Although reverberation was considered throughout the assessment of the binaural speech processors in this chapter, it has to be emphasized that the results have to be considered with care. For instance, the reverberation time in the study of Peissig (1992) was in the range of 2 to 3 s (the recordings were performed in a damped reverberation chamber) and, therefore, it was considerably higher than in the present simulated reverberation environment with an RT of 0.2 s. For this reason and with respect to localization experiments that have been performed in MISM-simulated and real-world reverberation conditions at our institute (Opdam, 2010), we believe that the speech intelligibility enhancement results of the MISM-simulated reverberation conditions are underestimated.

As the MISM used here calculates the first two reflections and subsequently adds a random exponentially decaying energy tail, any spatial correlation with the target signal is lost, except the first early reflections. Hence, the spatial correlation in higher order reflections, which might be useful in a directional separation approach, cannot be exploited.

In fact, the influence of reverberation on speech is not yet fully understood. Therefore, it may be a flawed to base solutions in the field of speech enhancement solely on (possibly inadequately) simulated environments.

Depending on the interference effects with reflections, for example, certain speech modulations may increase, while others are damped. An MISM-approach that does not account for standing waves, i.e. a non-diffuse sound field, will not provide these effects. George (2007) analyzed the influence of reverberation and noise, separately and jointly, on speech perception for normal hearing and hearing impaired people. Reverberation, as the author found, degrades speech quality and leads to a smearing of the modulation of the target and the masking signal. Consequently, the co-modulation masking release as well as the chance of listening in the gaps, are reduced. As hearing impaired people severely suffer from a reduced temporal acuity and an elevated auditory threshold (together with other auditory and non-auditory effects), future research should extend the statistical analysis of binaural parameters of Chapter 3 as well as their application in noise suppression tasks to real-world reverberant environments.

6

Conclusions and Outlook

6.1 Conclusions

A revision of binaural computational auditory scene analysis (CASA) processors, and their serial combination with non-adaptive beamforming front-ends of different degrees of directivity has been given. The approach offers an approximation of the minimum mean square error solution to the speech-in-noise problem.
The main intention of this study has been to establish a comprehensive understanding of the possibilities that the presented combined processing scheme provides for the enhancement of noise-corrupted speech. Three binaural speech processors have been studied. Their basic designs refer to the algorithm of Gaik and Lindemann (1986), here referred to as algorithm CLP, to the algorithm of Kollmeier and Koch (1994), here referred to as algorithm ELT, and to the algorithm of Allen et al. (1977), here referred to as algorithm CC. By means of a model-based improvement and a model-based assessment of speech intelligibility, the study pursued the optimal application of binaural speech processors in varying sound scenes.

The main question of this research can be affirmed, in that the proposed combination of bilaterally applied diffuse-field optimized beamforming filters and an adaptive binaural speech processor allows complementarily for the attenuation of diffuse noise and coherent noise, respectively.

Subsequent to the introduction of the algorithmic concepts in Chapter 2, the work has been subdivided into three parts. In Chapter 3 the statistics of binaural parameters have been analyzed and a soft-decision classification approach has been presented. In Chapter 4 a study on objective measures of speech intelligibility for binaurally and nonlinearly processed speech has been given. In Chapter 5 the techniques of a model-based improvement and assessment of speech intelligibility have been combined to yield an optimal solution for the speech-in-noise problem. The main findings of each chapter are summarized below.

Statistics of binaural parameters in noise

The source separation power of binaural speech processors depends on the distinctness of binaural parameters. This distinctness is strongly related to the strength and spatial distribution of the interference. In order to study this dependence, the binaural parameters of the fine-structure of the waveform as well as the corresponding parameters of the envelope at the output of different directional and non-directional hearing aids have been analyzed under varying noise conditions. Reference experiments have been conducted with an artificial head.

The study provides an understanding of natural as well as artificial binaural parameters in noise. A first salient insight of the statistical results is that directional hearing aids alter the front-back ambiguity around the interaural axis of natural binaural cues. This natural ambiguity is, in three dimensions, well-known as the cone of confusion artifact for narrow-band sounds. As it has been shown, besides the attenuation of sounds from the side and the rear due to the directional processing of a beamformer, the cone of confusion is warped.

In diffuse noise fields and low SNR conditions, it has been demonstrated that binaural parameters of both the fine-structure of the waveform and the corresponding envelope are weak indicators of directivity. No clear difference has been found in terms of the directivity characteristics of directional front-ends under the same conditions, using the ear-level and binaurally averaged SNR in the mixing process. The study thereby explains why the presented binaural speech processors offer only a modest performance in diffuse interference conditions. Nevertheless, when concurrent sources are coherent in space, i.e. a condition that allows for high degree of separability through the signal transformations, binaural parameters have shown to be an accurate source classifier.

As it has been shown, the binaural parameters are not equally affected by signal degradation. The fine-structure IPD and the envelope ILD can be considered the most reliable parameters in noise. The envelope IPD parameter and the envelope ITD analogy were revealed to be the most sensitive parameters to noise. With respect to the effect of the front-end, the directional hearing aids demonstrated a positive effect on the fine-structure IPD, especially at low frequencies. The ILD of the carrier of the waveform as well as of the corresponding envelope, on the other hand, evinced less variation, and hence less separation power, as a consequence of the directional processing of the front-ends.

The challenge of binaural speech processors remains in the optimal activation of binaural parameters in a schema-based fashion. Using Bayes' statistics, a soft decision approach was introduced by Harding et al. (2005) and has been adopted for the binaural algorithms of Gaik and Lindemann (1986) and Kollmeier and Koch (1994) in the present work.

In view of the unfavourable characteristics of binaural parameters for a source separation in diffuse noise fields, we have to consider that the model hearing process

combines many more cues in demanding circumstances. Consequently, it may discount equivocal cues of binaural disparity in favour of timbre and modulation. Future CASA-based systems might advance in the source separation problem through a combination of multiple cue-based separation strategies as well as a weighted activation of these in a supervised fashion. Universal concepts underlying such a development have been proposed (Blauert, 2011; Kolossa, 2011).

Assessment of speech intelligibility of binaural and nonlinearly processed speech

The instrumental speech intelligibility assessment of binaurally and nonlinearly processed speech is a relatively young field. Therefore a universal measure is still missing. The instrumental evaluation is driven by the necessity that subjective evaluation offers little active insight during the development of speech processors due to the high complexity of the task. In the present work, for instance, each binaural speech processor comprises a set of algorithmic parameters that need to be tuned to a specific acoustical scene as well as to the peculiarities of a certain front-end. Therefore, a set of instrumental measures has been analyzed and tested against subjective intelligibility scores of binaurally and nonlinearily processed speech.

A binaural and speech-based STI has been developed. For the purpose of copying the binaural processing, the coincidence model of Jeffress (1948) has been applied to calculate the binaural interaction effect. In frequency regions where the head-shadow effect dominates, a better ear approach has been chosen. Both effects contribute to the binaural advantage that is observed when listening with two ears instead of one. In comprehensive listening tests of the speech intelligibility of binaurally processed speech, the proficiency of the developed STI-version in mimicking the binaural processing has been widely confirmed. Despite this success, the metric has shown to fail in the assessment of nonlinear distortions, in particular in the assessment of envelope thresholding distortions which are a simulation of varying filter-gain functions, i.e. CASA-based post-filters.

In a follow-up study, several speech intelligibility measures, which label the relative information content, have been proposed and compared to existing measures, e.g. the I3 measure of Kates and Arehart (2005b) or the STOI measure of Taal et al. (2010). In a listening test of the intelligibility of linearly and nonlinearly processed speech, the potential of the proposed measures has been ruled out by the I3 measure, among others. Furthermore, the I3 measure could be improved in the prediction of nonlinearly processed speech, on the basis of the applied speech material. However, in order to operate with a well-established measure, the original I3 metric has been chosen for the optimization and assessment of binaural speech processors in this work. To account for the dominating binaural effect, the I3 metric has been extended to incorporate the head-shadow effect, which is calculated in a better ear fashion per combination of short-time frame and critical band.

Optimization and assessment of binaural post-processors

The present work proposes the application of a genetic algorithm (GA) in the holistic framework of model-based improvement and model-based assessment of speech intelligibility. To that end, the Genetic Algorithms for Optimization Toolbox of Houck et al. (1995) has been applied in its default settings and the Better Ear I3 measure served as a cost-function. To obtain an indication of the optimization complexity, the reproducibility of the GA solutions has been examined by running several GA optimizations for each test setup. Additionally, the GA-optimized parameter sets have been applied in changing acoustic environments, which yields an indication of the robustness and the generalizability of a certain solution.
Overall, the GA optimization of algorithmic parameter sets produces practical, optimal and psycho-acoustically relevant solutions. While the efficiency of the GA procedure is a consequence of the survival of the fittest strategy, the regularity of the solutions is a product of the interplay of CASA algorithms in the improvement and assessment of speech intelligibility. Therefore, the GA approach has proven to be a very efficient means for the parameter adjustment. Moreover, the holistic optimization method provides solution strategies that may underly the ranking of low-level cues in the model hearing process, as shown here for the ranking of binaural cues. By way of example, for front-end and scene combinations in which the fine-structure ILD is significantly included in the directional weighting process, the ILD never gained more than half of the algorithmic weight in the total directional filtering process. By trend, these results correspond to psycho-acoustic tests about the trading of binaural cues in noise (Rakerd and Hartmann, 2010).

The subsequent assessment of the combined processing schemes provided the understanding that binaural speech processors are approximately independent of the directivity of the front-end. In accordance with the nature of binaural statistics, a main dependence of the performance on the ear-level and binaurally averaged SNR has been isolated. As a consequence, when employed as a post-processor, a speech intelligibility improvement of a binaural speech processor adds to the improvement gained by the directional processing of the front-end. A limited frequency transfer and equivocal binaural cues of the front-end, however, have been shown to detract from the separation power.

The comparison of different binaural speech processors demonstrated the superiority of algorithm CLP under diffuse and in coherent noise conditions. For instance, in the babble noise of a lively canteen the CLP algorithm reaches an absolute and objectively estimated improvement of approximately 20 % of intelligibility, as assessed with the Better Ear I3. In coherent noise conditions, algorithm CLP can achieve an absolute improvement of more than 40 % of speech intelligibility. The gain has shown to be independent for a wide range of target/masker angles, SNR conditions and front-ends.

Due to the susceptibility of the interaural temporal differences of the envelope to

diffuse noise fields, as it has been demonstrated in Chapter 3, algorithm ELT showed to be merely beneficial under highly coherent noise conditions. Nevertheless, in these sound scenes it offers a separation power that is comparable to algorithm CLP in the attenuation of one competing sound source.
The ability of algorithm ELT to suppress lateral sources indicates the proper application of binaural parameters of the envelope in the source separation task. This result contrasts with the findings of Kollmeier and Koch (1994), in which no strong attenuation of coherent interference from lateral directions was reported. The improved operation in coherent interference followed from our revision of the algorithm ELT, a statistical study of binaural parameters and an optimal pattern-based application.
However, following the strict objective in this work of employing binaural parameters in the noise suppression process, our implementation of algorithm ELT lost the benefit of an estimated SNR of 2 dB that was gained by Kollmeier and Koch (1994) in negative SNR as well as diffuse noise conditions. The present study led to the conclusion that this benefit in diffuse sound fields appears not to be a consequence of a direct spatial filtering, but a result of the second weighting function of algorithm ELT. This weighting function is based on the standard deviation of binaural envelope parameters in a joint centre and modulation-frequency domain. In application, its working principle comes down to the deviation of binaural envelope parameters from the median plane. Therefore, a diffuse-field gain can be observed, basically indicating the audiological benefit when filtering in the modulation domain.
Although the present implementation employs a weighting function based on a statistical penalty measure too, it has not been possible to verify the diffuse-field benefit with the present implementation. The reason for this shortcoming can be attributed to coarser modulation spectra given by a decreased analysis frame-length, which is necessary for the suppression of lateral interference.

With respect to algorithm CC, almost no improvement of speech intelligibility under coherent noise conditions has been gained. This finding has been expected, since the grouping scheme of this algorithm is inappropriate to differentiate the angular positions of coherent sound sources. Nevertheless, under highly diffuse noise conditions, as e.g. a workshop environment, an objective absolute improvement of speech intelligibility of up to 25 % has been achieved.

Finally, the results of the stochastic optimization reflect the need for an adjustment of the model-based algorithms in changing acoustics. As it has been shown, in the worst case, the GA optimization prevents a deterioration of speech intelligibility through the binaural filtering process. In all observed cases, the optimization process adapts the binaural speech processors in the best possible way to the boundary conditions and allows for robust speech intelligibility gains, even when the SNR or the angular positions of maskers change.

6.2 Outlook

The motivation behind this work has been to give a revision of binaural speech processors, which are connected in series to binaural front-ends. For the purpose of a structured analysis, the binaural speech processors have been analyzed in its basic algorithmic design and no combination has been considered.

As far as a refinement of this study is concerned, the following factors could be varied: the influence of the global mixing SNR used in the optimization has not yet been defined on the overall performance of the post-processor as well as the resulting robustness in changing SNR conditions. Furthermore, the SNR criterion of the IBMs used in the classification task of algorithm CLP and ELT could be studied, and compared to the choices made here. In addition, the overall gain of the combined processing scheme needs to be studied by defining diffuse and non-diffuse sound scenes with a source-dependent SNR.
As long as instrumental measures constitute a compromise for a small set of particular degradations and a certain dimension of speech perceptions, e.g. speech intelligibility, subjective evaluations are mandatory to verify the benefit of certain solutions to the speech-in-noise problem. Consequently, the results given here should be evaluated in a listening test. In addition, subjective tests should incorporate higher cognitive factors of speech perception, such that the overall ease of listening will be evaluated. As it has been discussed, the ease of listening combines several perceptual dimensions and complements the prediction of the aid's benefit.

In a next step, binaural processors could be combined, as e.g. done with algorithm CLP and CC in the works of Peissig (1992) and Wittkop and Hohmann (2003), and analyzed, updated and compared to the fundamental algorithms of this work.
Analogously, combinations of binaural with monaural modulation-based approaches have shown to be beneficial in terms of an SNR improvement (Woods et al., 1996; Woodruff et al., 2010). It appears worthwhile to combine these multi-cue algorithms with the advanced methods in pattern recognition, as recently proposed by Weiss et al. (2011), for prospectively bridging the gap between CASA and ASA.

There are several potential studies in terms of signal-dependency that can be carried out. To begin with, a redesign of the presented binaural speech processors with varying signal-dependent filterbanks can be suggested as an approach to increase the time-resolution of the filterbank in the presence of transients or to increase the spectral resolution in the presence of harmonic speech. Thereby, the constraint imposed by the uncertainty principle that results in a compromise in a fixed analysis-synthesis approach can be alleviated.
The DFT has been applied in this work, leading to a constant frequency resolution. In future work, this approach should be compared to highly resolved auditory filterbanks, as e.g. a gammatone filterbank applied in the processor of Roman et al. (2003).
In addition, the incorporation of signal-dependency in the estimation of the power

spectral densities proved to be beneficial (see e.g. Thiergart et al., 2010), and should be applied to the algorithms presented here.
Furthermore, a signal-dependent enhancement of transients demonstrated to improve speech intelligibility considerably (Yoo et al., 2007). For single channel speech enhancement, Mauler (2010) recently proposed the amplification of transients, and verified an improvement of speech intelligibility for cochlear implant users.

In the present work, pattern-driven decision rules and scene-optimized parameter sets are employed for the noise suppression task. In order to employ this lookup data in an optimal way across different backgrounds, classification and scene analysis algorithms have to be used. Such a combination of a binaural localizer that steers a binaural post-processor to yield an optimal speech quality output was proposed in Boone et al. (2010). Future classifiers should advance to combine background classification as well as scene analysis. As shown in Boone et al. (2010), a scene analysis offers the parametric description of varying nature of binaural parameters with e.g. a Gaussian-mixture model. The comparison of such parametric lookup tables with histogram-based methods, as presented here, is deemed an important future task.

The instrumental measure of speech intelligibility that was applied throughout the assessment of binaural processors constitutes a compromise that has been achieved in the scope and context of the presented work. As such, the measure is not satisfying. Future work should undertake efforts to derive a measure that accounts more precisely for the binaural advantage as well as the effect of nonlinear noise-reduction on speech intelligibility. We believe that simple physical measures are incapable of attaining that goal. Higher processes of speech perception need to be included in a comprehensive model of speech intelligibility. Prospectively, we hope to refine the speech-based and binaural STI, in order to obtain a comprehensive model for predicting the intelligibility of binaurally as well as nonlinearly processed speech. An attempt is shown in Schlesinger (2012), in which a transient-based STI is developed and optimized against perception. On untrained data, this transient-based STI method shows to be well suited to predict the intelligibility of linearly and nonlinearily processed speech.

A

Appendix: Algorithmic definitions

A.1 Formulation of the Wiener filter

In the following a simplified derivation of the Wiener filter in the frequency domain and in the notation of this work is given. A more thorough derivation of the Wiener filter can be found in Hänsler and Schmidt (2004).
The filter assumes short-time stationarity of the signals and that the power spectral densities are known. For the reconstruction of the time series $y(\iota)$ of a target signal from a noise-corrupted signal $x(\iota)$ through a filtering process with $w(\iota)$:

$$y(\iota) = w(\iota) * x(\iota), \tag{A.1.1}$$

the Wiener filter is the MMSE solution of:

$$w_{\mathrm{opt}}(\iota) = \underset{w}{\operatorname{argmin}} \, \mathrm{E}\left[(s(\iota) - y(\iota))^2\right], \tag{A.1.2}$$

where $s(\iota)$ is the original target signal. After calculating the Fourier Transform of the time series, this expression becomes:

$$w_{\mathrm{opt}}(d) = \underset{w}{\operatorname{argmin}} \, \mathrm{E}\left[(s(d) - w(d)x(d))^2\right], \tag{A.1.3}$$

and by separating the target signal and noise $v(d)$:

$$w_{\mathrm{opt}}(d) = \underset{w}{\operatorname{argmin}} \, \mathrm{E}\left[(s(d) - w(d)(s(d) + v(d)))^2\right]. \tag{A.1.4}$$

After minimization with $w(d)$ as a variable and the replacement of the expectation operator with linear operations, one gains the Wiener filter in the frequency domain:

$$w_{\mathrm{opt}}(d) = \frac{\phi_{\mathrm{ss}}(d)}{\phi_{\mathrm{ss}}(d) + \phi_{\mathrm{vv}}(d)}, \tag{A.1.5}$$

where $\phi_{\mathrm{ss}}(d)$ and $\phi_{\mathrm{vv}}(d)$ denote the power spectral density of the signal and noise, respectively.

A.2 The direct DFT-IDFT filter approach in speech enhancement

The DFT is generally applicable for stationary and deterministic signals. In speech processing the DFT is usually calculated over segments of short durations (15 to 35 ms), in which speech is approximately stationary. The approach is widely known as short-time Fourier transform (STFT) and constitutes a very efficient means in the filtering process. In here, we give the windowed analysis and the overlap-add synthesis method, in brief the DFT-IDFT approach.

The discrete time sequence of a speech signal $x(\iota)$ of length N_ι, sampled at f_s, is partitioned into overlapping frames of length N_χ using a window function $\chi(\iota)$ to reduce spectral leakage. Throughout this work the Hanning window $\chi(\iota) = (1 - \cos(2\pi\iota/N_\chi))$ with $\iota = 0, 1, \ldots, N_\chi$ has been applied. To avoid a circular convolution in the filtering process, the window function $\chi(\iota)$ is (for example) appended with an array of zeros of length $N_\mathrm{d} - N_\chi$, with N_d being the length of the DFT and the requirement that $N_\mathrm{d} \geq 2N_\chi - 1$. Throughout this work, usually a frame shift of 50 %, i.e. $\Delta T = N_\chi/2$ is applied. Using these definitions, perfect reconstruction can be approximated with the DFT-IDFT approach.

Analogously to Equation (2.1.1), the DFT of the time sampled speech sequence $x(\iota)$ can be computed with:

$$x(d, n) = \sum_{\iota=0}^{N_\mathrm{d}-1} \chi(\iota)\tilde{x}(n\Delta T + \iota)e^{-j2\pi\iota\frac{d}{N_\mathrm{d}}}, \tag{A.2.6}$$

where $d = 0, 1, \ldots, N_\mathrm{d} - 1$ and n are the frequency bin and the frame index, respectively. Subsequent to a signal modification in the STFT domain, the stimulus is reconstructed by the IDFT and overlap-add reconstruction (Loizou, 2007):

$$\tilde{x}(\iota) = \sum_{n=-\infty}^{\infty} \frac{1}{N_\mathrm{d}} \sum_{d=0}^{N_\mathrm{d}-1} x(d, n\Delta T)e^{j2\pi d\frac{\iota}{N_\mathrm{d}}}, \tag{A.2.7}$$

therein the frame shift ΔT is multiplied with n to align the short-time frames in the reconstructed signal. For real-world implementations, the DFT length can be shortened by a factor of two, if instead of zero padding a synthesis window is applied prior to the overlap-add synthesis procedure (Mauler, 2010).

A.3 Range and mean standard deviation

This appendix presents the calculation methods of the range parameter m' and the mean standard deviation σ', as used in Chapter 3.3.

The calculation starts with a sample value $\check{o}_{\check{i}}$ of a stochastic variable $\check{O}$, which is given through:

$$\check{o}_{\check{i}} = \mathrm{E}\{\check{O}\} + \check{\epsilon}, \tag{A.3.8}$$

where the expectation of the error $\check{\epsilon}$ is:

$$\mathrm{E}\{\check{\epsilon}\} = 0. \tag{A.3.9}$$

The arithmetic mean is calculated with:

$$\bar{o} = \frac{1}{\check{n}} \sum_{\check{i}=1}^{\check{n}} \check{o}_{\check{i}}. \tag{A.3.10}$$

Given $\bar{o}$ represents the mean of a binaural cue at a certain angle of incidence θ, the range (i.e. the width of the cue variation) m' of this binaural cue is calculated as:

$$m' = \max[\bar{o}(\theta)] - \min[\bar{o}(\theta)]. \tag{A.3.11}$$

The standard deviation σ of the sample value is calculated as:

$$\sigma = \sqrt{\frac{1}{\check{n}-1} \sum_{\check{i}=1}^{\check{n}} (\check{o}_{\check{i}} - \bar{o})^2}. \tag{A.3.12}$$

Consequently, the mean standard deviation σ' of a binaural cue is computed as:

$$\sigma' = \frac{1}{N_\theta} \sum_{u=1}^{N_\theta} \sigma_\theta, \tag{A.3.13}$$

in which N_θ is the number of discrete target source positions and σ_θ is the standard deviation of a certain binaural parameter at one target direction.

A.4 Statistics of the model-assessment

The r^2 measure calculates the quotient of the variance in the model to the original variance in the data. Thereby, r^2 gives the amount of variance of the original data that is explained by the model, also known as the 'Bestimmtheitsmaß' (Bortz, 2005). It can be calculated as:

$$r^2 = \frac{\sigma_s^2 - \sigma_{s-\hat{s}}^2}{\sigma_s^2}, \tag{A.4.14}$$

where σ_s^2 is the original variance of, for instance, the subjective scores $s_{\check{i}}$ and $\sigma_{s-\hat{s}}^2$ is the variance of the residuals, i.e. the differences between the measurements $s_{\check{i}}$ and the model-predictions $\hat{s}_{\check{i}}$.

Kendall's τ is a rank statistic and calculates the monotonic relation between intelligibility scores and objective scores. The statistic is calculated as:

$$\tau = \frac{\check{p}_c - \check{p}_d}{\frac{1}{2}\check{n}_\mathrm{p}(\check{n}_\mathrm{p} - 1)}, \tag{A.4.15}$$

where $\check{p}_c$, $\check{p}_d$ and $\check{n}_\mathrm{p}$ determine the concordant pairs, discordant pairs of conditions and the amount of the tested conditions, respectively (Bortz, 2005).

B

Appendix: Interaural parameters of the binaural envelope signal

This appendix extends the statistical analysis on interaural parameters of the envelope in Chapter 3.3. For a description of the figures, the reader is referred to Section 3.3.2.

In addition, the binaural magnification method of Section 3.3.2 is analyzed in greater detail. A comparison between the original envelope ILD calculation method with the magnification method at the output of the HG (low directivity) is given in the Figures B.5 and B.6, respectively. The results indicate an approximate doubling of all m' values. Finally, Figure B.7 shows the magnified envelope ILD in the canteen situation at an SNR of 0 dB.

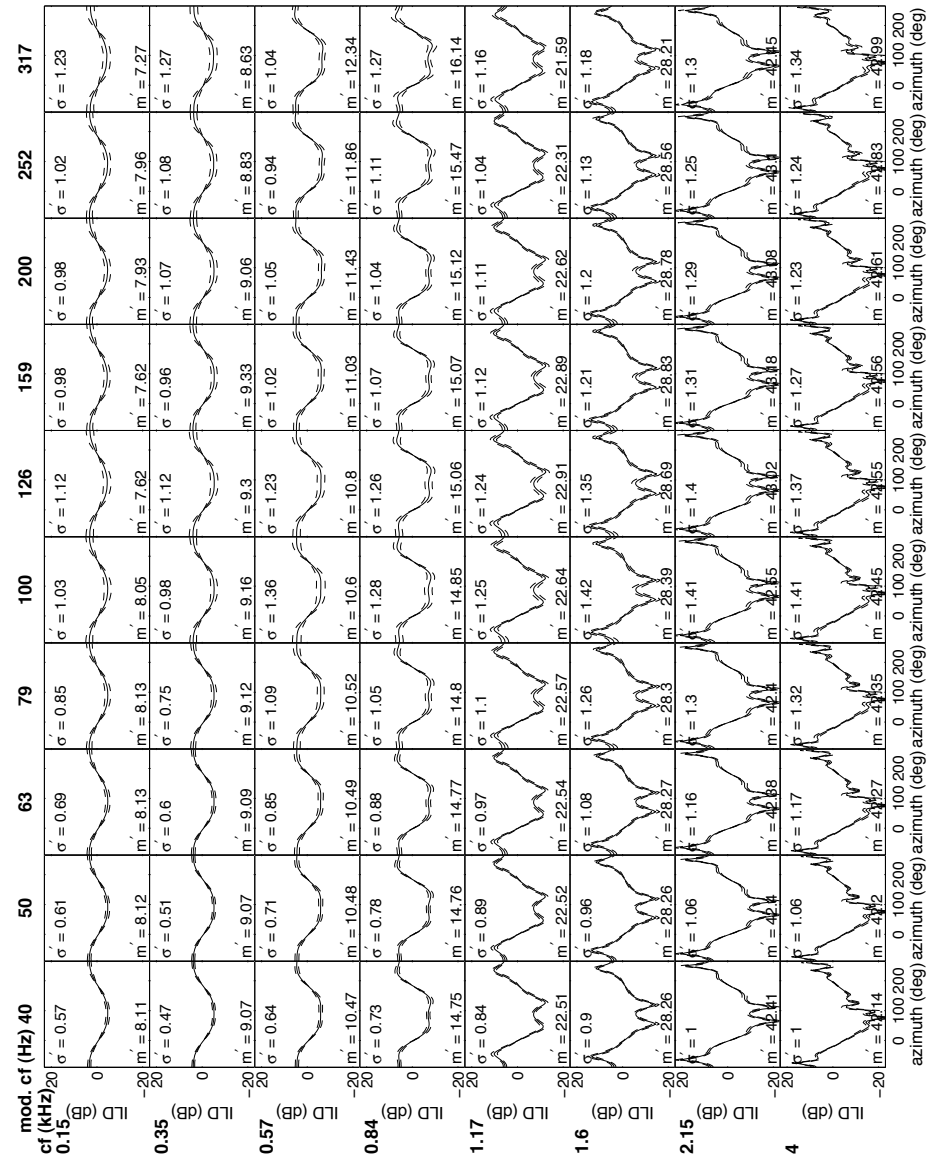

Figure B.1: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based ILD parameter at the Aachen head as a function of source direction, analyzed in carrier and modulation band combinations that are centred at the specified frequencies (*cf *and* mod. cf*, resp.). The SNR was set to 60 dB.*

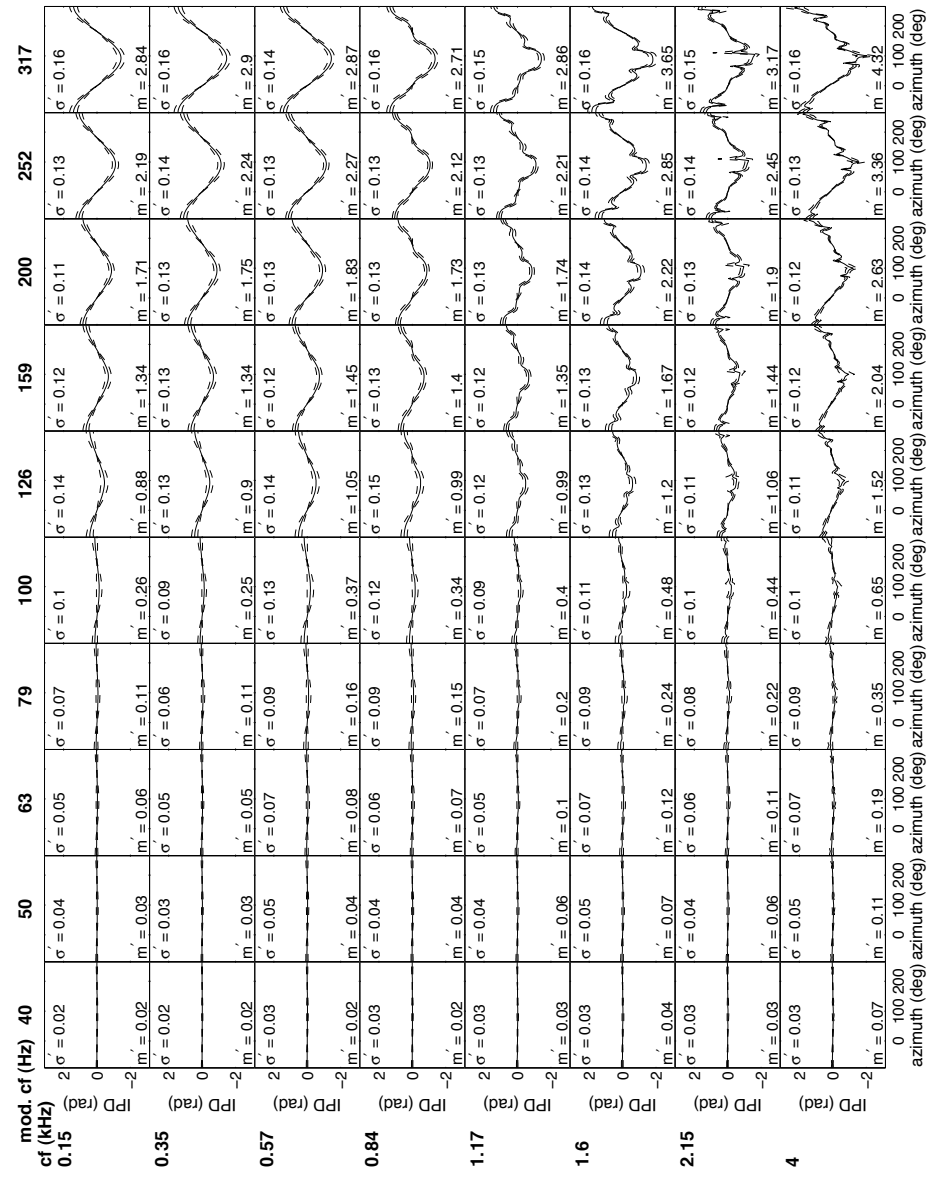

Figure B.2: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based IPD parameter at the Aachen head as a function of source direction, analyzed in carrier and modulation band combinations that are centred at the specified frequencies (*cf *and* mod. cf*, resp.). The SNR was set to 60 dB.*

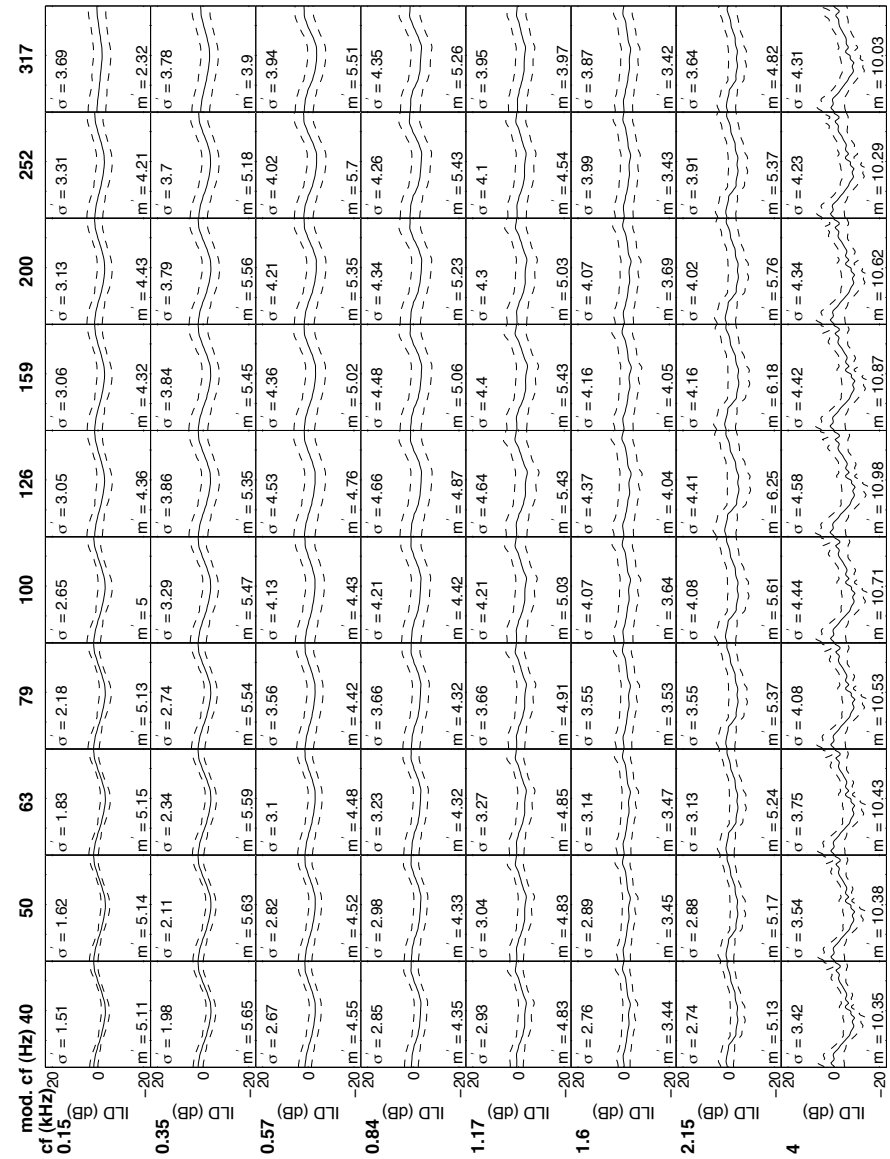

Figure B.3: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based ILD parameter at the Aachen head as a function of source direction, analyzed in carrier and modulation band combinations that are centred at the specified frequencies (*cf *and* mod. cf*, resp.). The SNR was set to 0 dB and the interference is a lively canteen.*

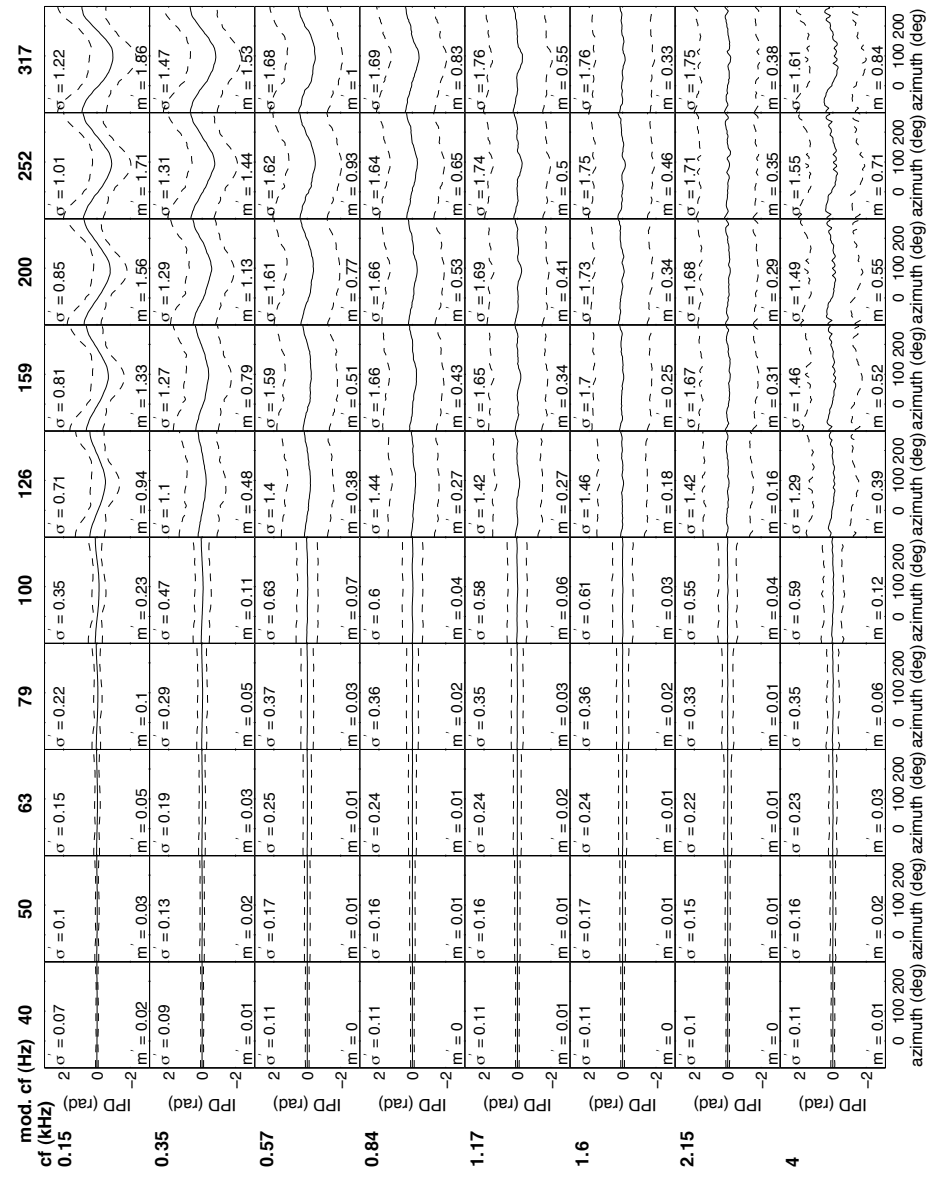

Figure B.4: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based IPD parameter at the Aachen head as a function of source direction, analyzed in carrier and modulation band combinations that are centred at the specified frequencies (*cf *and* mod. cf*, resp.). The SNR was set to 0 dB and the interference is a lively canteen.*

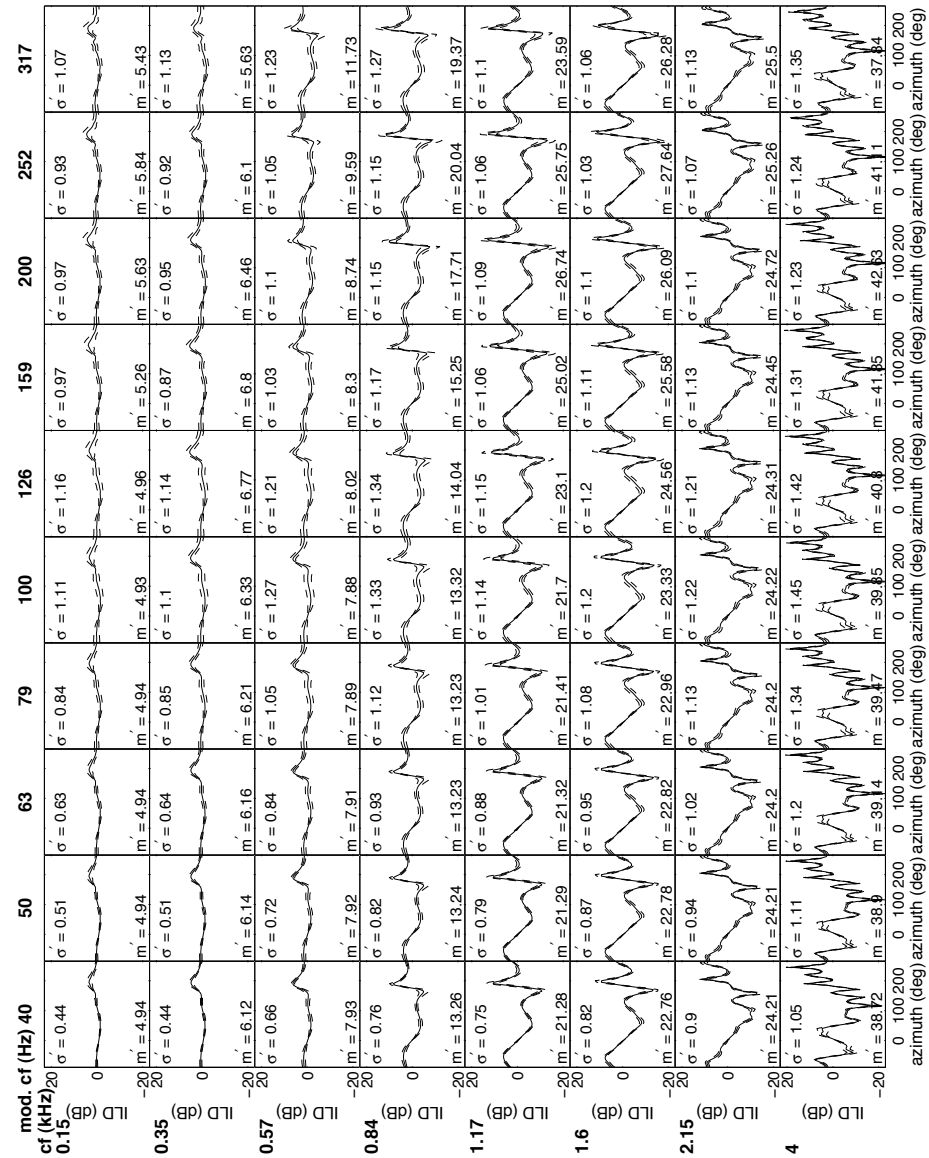

Figure B.5: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based ILD parameter at the HG (low directivity) as a function of source direction, analyzed in carrier and modulation band combinations that are centred at the specified frequencies (*cf *and* mod. cf*, resp.). The SNR was set to 60 dB.*

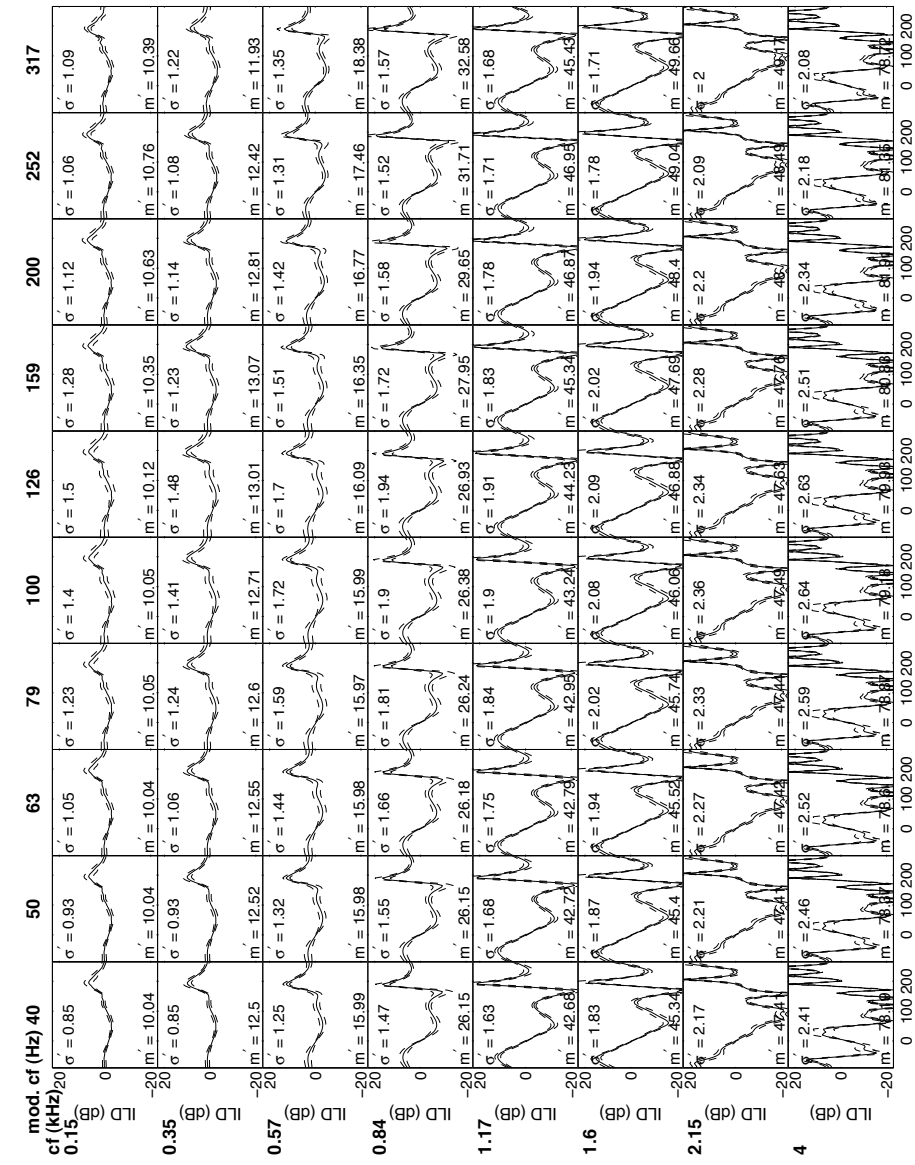

Figure B.6: *The PDFs with mean (solid line) and standard deviation (dashed line) of the magnified envelope-based ILD parameter at the HG (low directivity) as a function of source direction, analyzed in carrier and modulation band combinations that are centred at the specified frequencies (*cf *and* mod. cf*, resp.). The SNR was set to 60 dB. The ILD magnification procedure of Equation (3.3.7) was applied.*

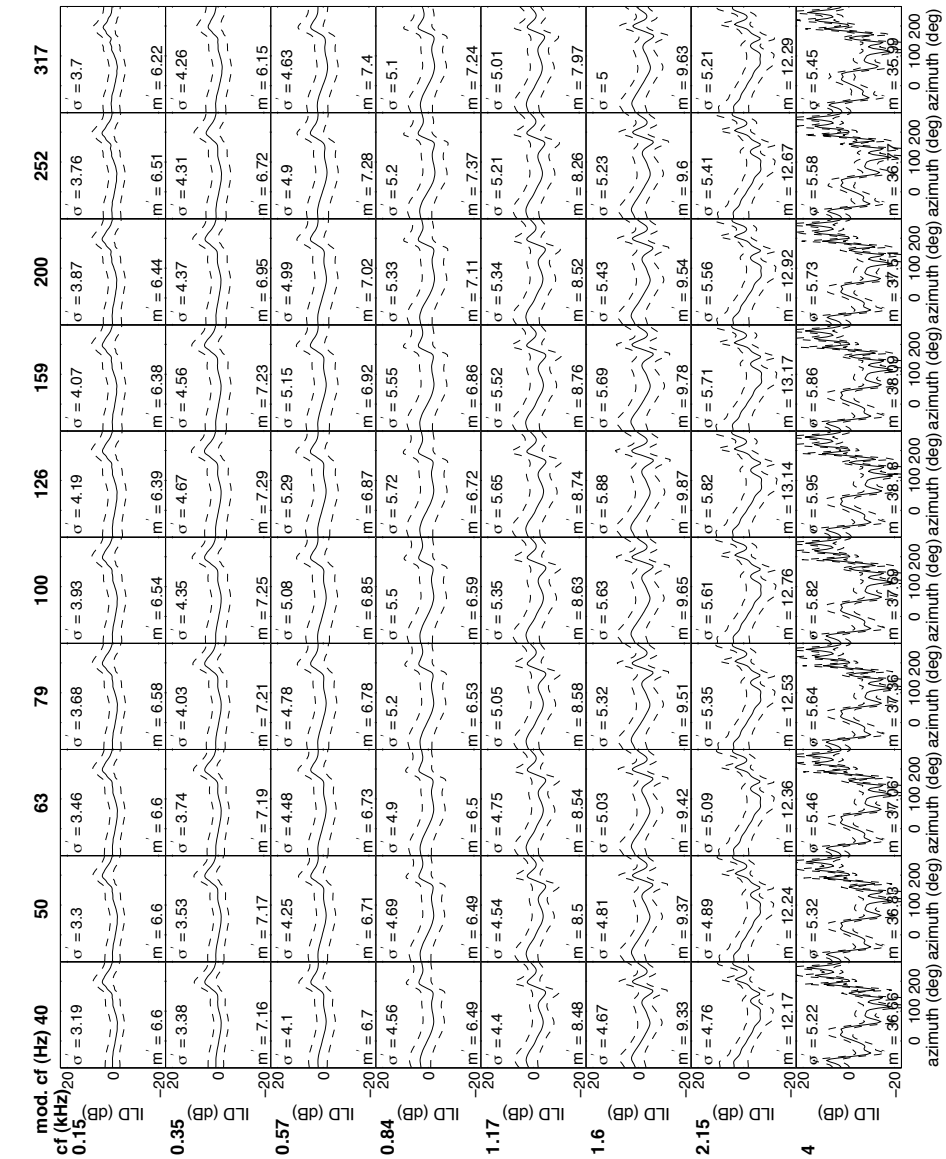

Figure B.7: *The PDFs with mean (solid line) and standard deviation (dashed line) of the magnified envelope-based ILD parameter at the HG (low directivity) as a function of source direction, analyzed in carrier and modulation band combinations that are centred at the specified frequencies (*cf *and* mod. cf*, resp.). The SNR was set to 0 dB and the interference is a lively canteen. The ILD magnification procedure of Equation (3.3.7) was applied.*

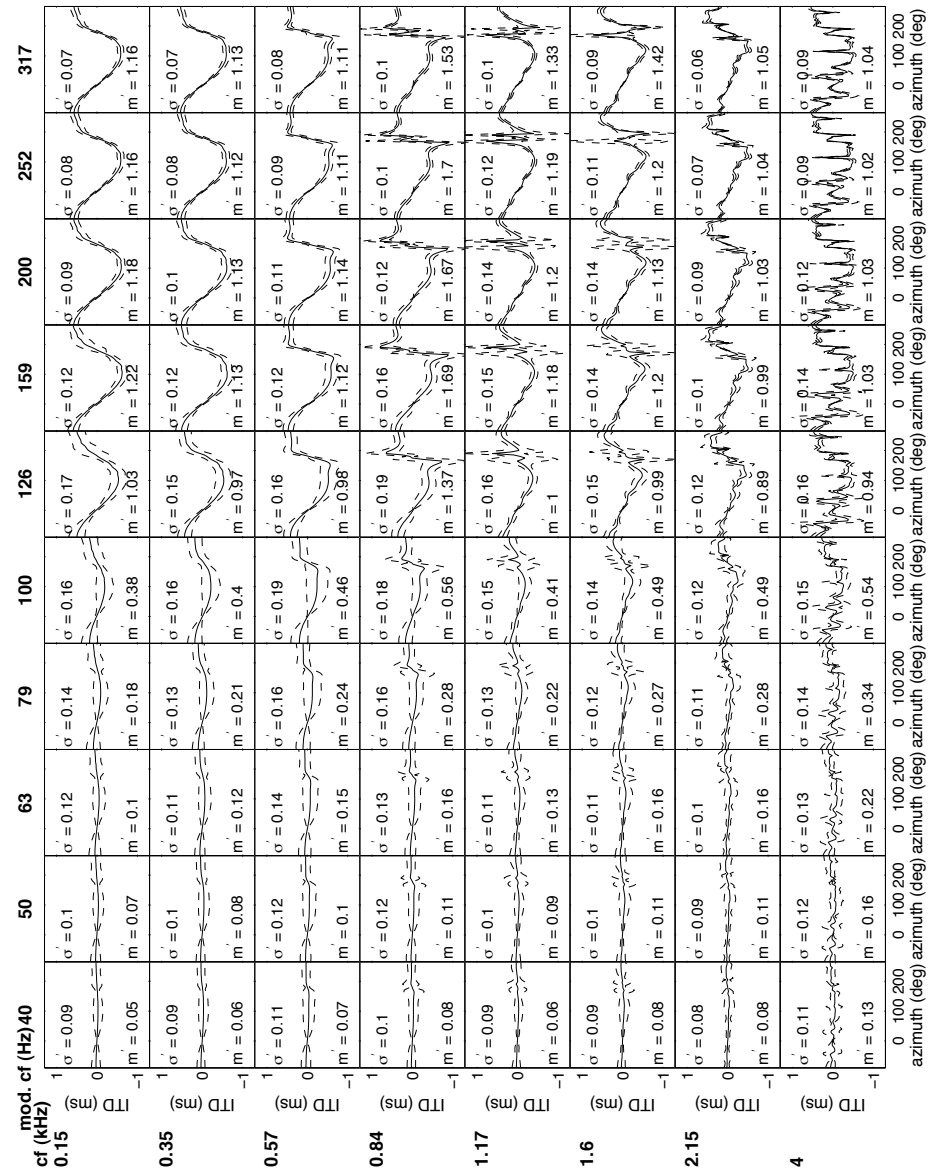

Figure B.8: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based ITD parameter at the HG (low directivity) as a function of source direction, analyzed in carrier and modulation band combinations that are centred at the specified frequencies (*cf *and* mod. cf*, resp.). The SNR was set to 60 dB.*

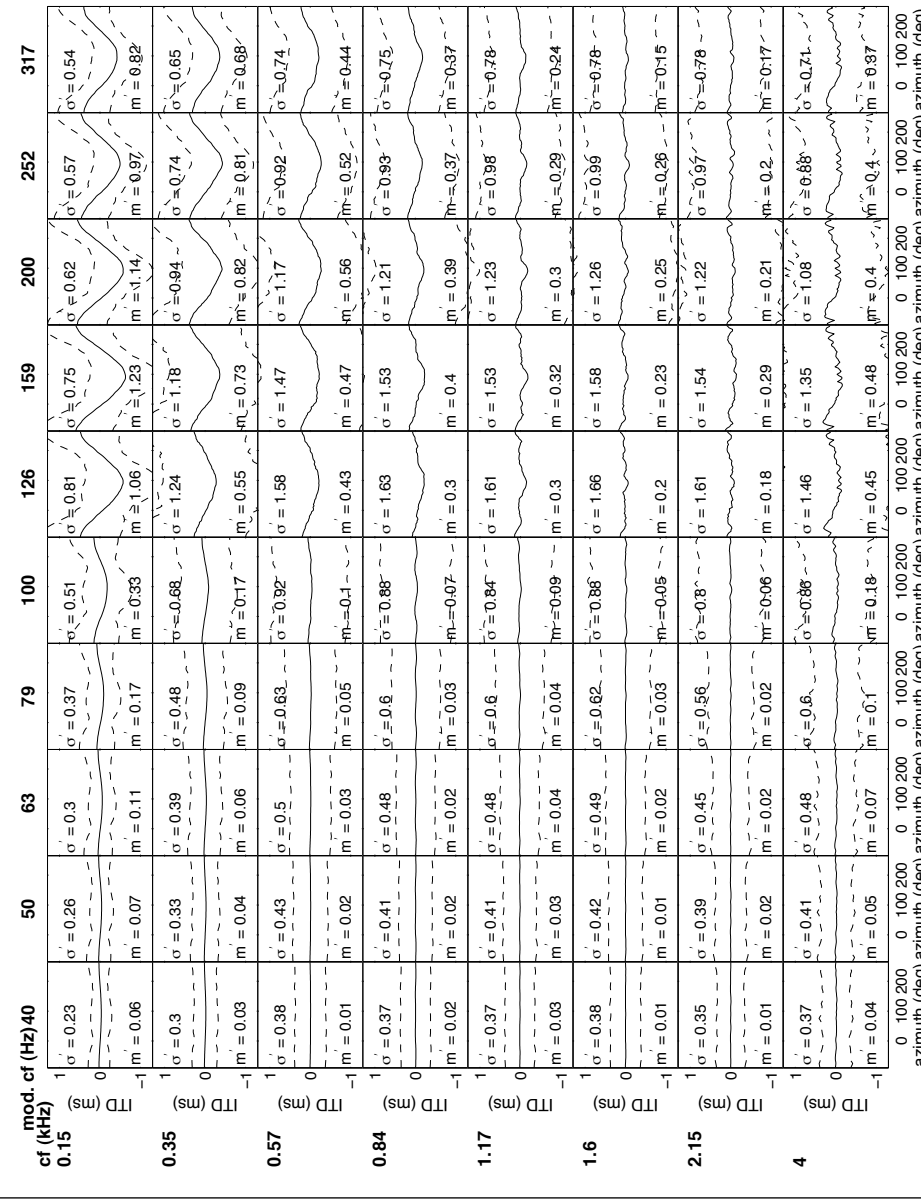

Figure B.9: *The PDFs with mean (solid line) and standard deviation (dashed line) of the envelope-based ITD parameter at the HG (low directivity) as a function of source direction, analyzed in carrier and modulation band combinations that are centred at the specified frequencies (*cf *and* mod. cf*, resp.). The SNR was set to 0 dB and the interference is a lively canteen.*

C

Appendix: A comparative study on speech intelligibility measures for nonlinearly processed speech

This appendix gives a comparison of three intrusive speech-based speech intelligibility measures.
For their assessment a subjective speech intelligibility test has been performed. The speech material has been taken from the semantically unpredictable sentence (SUS) corpus, which was developed by Ramirez et al. (2009). The speech files were recorded and digitized at 44.1 kHz. The clean and distorted wave forms were convolved with the HRTFs of an artificial head in a particular acoustical scene and corrected for the headphones that were used in the listening test. The masking signal was presented at a fixed level of 70 dB(A) SPL and the target level was changed to the respective SNRs used in the different test conditions. The recordings were stored for further analysis with the objective intelligibility measures. Four subjects of normal hearing ($<$ 15 dB (HL) for both ears) participated in the diotic test with three trials per condition. The conditions ranged from no deterioration to several forms of linear deteriorations using echo, reverberation and a single masker of speech shaped stationary noise. Situations 24 to 27 indicate nonlinear envelope threshold distortions from soft to severe, respectively, which increase the modulation depth abnormally and can therefore lead to an overestimation by a modulation-based intelligibility measure (Goldsworthy and Greenberg, 2004). Table C.1 lists the conditions assessed in this study.
The noise-corrupted and distorted sentences of the listening test were stored for further analysis with objective intelligibility measures. Moreover, silent passages in the speech files, in here defined as the RMS level lower than -50 dB in frames of 32 ms length relative to the overall RMS level, lead to an erroneous increase of the STI with the envelope regression method (Payton and Shrestha, 2008) and were, therefore, excluded with a VAD algorithm.
Three speech-based intelligibility measures have been applied in this study: the STMI of Elhilali et al. (2003) in an optimized envelope regression version of Goldsworthy and Greenberg (2004), the STI in an optimized envelope regression version of

Table C.1: *List of distortion conditions. The abbreviation bw. indicates the bandwidth employed in the stimuli. Wide-band (wb.) has a frequency range from 0.05 to 7 kHz and full-band (fb.) from 0 to 22.05 kHz.*

cond.	bw.	noise type	SNR (dB)
1	fb.		inf
2—5	fb.	white additive noise	[−8, −4, 0, 0]
6—9	fb.	pink additive noise	[−8, −4, 0, 0]
10—13	fb.	SSN male voice	[−8, −4, 0, 0]
14—18	wb.	SSN male voice	[−8, −5, −2, 1, 4]
19	fb.	Echo 50 ms	inf
20	fb.	Echo 50 ms plus SSN male voice	6
21	fb.	Echo 150 ms	inf
22	fb.	Echo 150 ms plus SSN male voice	6
23	fb.	Echo 150 ms plus SSN male voice	12
24—27	fb.	Envelope thresholding [50, 60, 75, 90 %]	inf
28	fb.	Reverberation 0.4 s	inf
29—31	fb.	SSN male voice plus rev. 0.4 s	[0, 3, 6]
32	fb.	Reverberation 1.5 s	inf
33—35	wb.	SSN male voice plus rev. 1.5 s	[6, 10, 15]

Goldsworthy and Greenberg (2004) and the coherence-based SII (CSII) of Kates and Arehart (2005b). As it regards the implementation of the STMI, the Neural Systems Laboratoy toolbox (URL *http://www.isr.umd.edu/Labs/NSL/*) was used for this purpose. The parameters of the NSL toolbox were set to 8 ms for the frame length, 8 ms for the time constant, the critical level ratio followed a linear function, NSL option −2, and the NSL octave shift option was set to 0. The STMI resolution was set to 4, 6, 8, 12, 16, 24 and 32 Hz for analyzing the temporal modulation and with 0.25, 0.375, 0.5, 0.75, 1, 1.5, 2, 3, 4, 6, 8 in cycles per octave for analyzing the modulation along the frequency dimension (scale). Subsequent to the STRF representations with the NSL toolbox, the optimized envelope regression method of Goldsworthy and Greenberg (2004) has been applied to the clean and the deteriorated signal. Finally, the transmission indices of the frequency channels have been weighted with the band-importance function of Pavlovic (1987) for average speech and summed to our version of the STMI.

The STI has been analyzed with a frame-length of 0.5 s and with the envelope regression method of Goldsworthy and Greenberg (2004). However, in our implementation, the modulation transfer in 30 auditory bands, as given in Chapter 4.1.2,

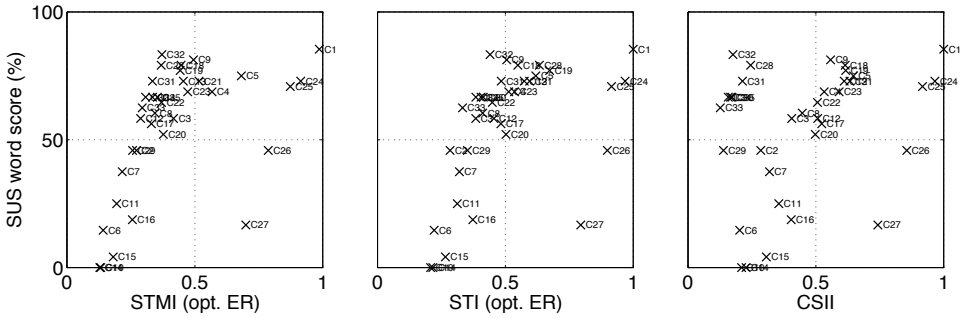

Figure C.1: *Analysis of three speech-based measures of speech intelligibility, the STMI, the STI and the CSII. The specification* opt. ER *denotes the optimized regression method of Goldsworthy and Greenberg (2004).*

has been analyzed. The CSII was calculated according to the definition of Kates and Arehart (2005b), see also Chapter 4.2.

Figure C.1 gives the results. As can be seen, all three measures correctly indicate the distortion of additive noise. Moreover, reverberation distortions are correctly predicted by the STMI and the STI. The CSII is a purely spectral measure and, therefore, cannot account for the reverberation, which is a temporal distortion. In addition, the STMI shows a smaller spread of linear distortions and some nonlinear distortions along a common curve (phase jitter and peak clipping) than the STI. Most important in the context of this thesis is the finding that none of these objective measures is capable to predict the impact of nonlinear envelope thresholding conditions. Chapter 4.2 departs from this observation with the development of a measure that offers a functional relationship between subjective scores and objective prediction for additive noise and envelope thresholding distortions.

D

Appendix: Artificial nonlinear signal distortions

This appendix gives the calculation methods of several nonlinear signal distortions.

D.1 Peak clipping

The peak clipping (symmetric) distortion used in the listening tests of this thesis has been calculated according to Kates and Arehart (2005b). Peak clipping is associated with amplifier and receiver saturation effects in hearing aids. The peak clipping distortion thresholds are calculated from the histograms of the magnitude of the signal samples. Here, subsequent to exclusion of silent parts of the sentences at the beginning and the end, the cumulative distribution of the magnitudes of the sentences was calculated. The threshold $\hat{\epsilon}$ was adjusted by a certain percent proportion of the cumulative distribution (see test conditions) according to:

$$s_{\mathrm{peak-clipping}}(\iota) = \begin{cases} \hat{\epsilon} & \text{if} \quad s(\iota) > \hat{\epsilon}, \\ s(\iota) & \text{if} \quad -\hat{\epsilon} \leq s(\iota) \leq \hat{\epsilon}, \\ -\hat{\epsilon} & \text{if} \quad s(\iota) < -\hat{\epsilon}, \end{cases} \tag{D.1.1}$$

where $s(\iota)$ is the clean input and $s_{\mathrm{peak-clipping}}(\iota)$ is the distorted output.

D.2 Envelope thresholding

The envelope thresholding (symmetric) distortion used in the listening tests of this thesis have been calculated according to Kates and Arehart (2005b). Envelope thresholding is associated with nonlinear noise suppression methods, which reduce the signal amplitude in low-level regions. The envelope thresholding distortion thresholds are calculated from the histograms of the magnitude of the signal samples. Here, subsequent to exclusion of silent parts of the sentences at the beginning

and the end, the cumulative distribution of the magnitudes of the sentences was calculated. The threshold $\hat{\epsilon}$ was adjusted by a certain percent proportion of the cumulative distribution (see test conditions) according to:

$$s_{\mathrm{envelope-thresholding}}(\iota) = \begin{cases} s(\iota) & \text{if} \quad s(\iota) > \hat{\epsilon}, \\ 0 & \text{if} \quad -\hat{\epsilon} \leq s(\iota) \leq \hat{\epsilon}, \\ s(\iota) & \text{if} \quad s(\iota) < -\hat{\epsilon}, \end{cases} \tag{D.2.2}$$

where $s(\iota)$ is the clean input and $s_{\mathrm{envelope-thresholding}}(\iota)$ is the distorted output.

D.3 Phase jitter

The phase jitter distortion used in the listening tests in this thesis have been calculated according to Elhilali et al. (2003). Such a distortion can occur on telephone channels if the power supply is fluctuating. Phase jitter destroys the carrier, while leaving the envelope intact. For this reason, the distortion is interesting in relation to the STI method. The STI analyzes the modulation depth of the envelope and is, therefore, generally unaffected by phase jitter distortions. Elhilali et al. (2003) showed that the classical STI method is absolutely unaffected by such distortion. However, in own experiments (not shown here, but partly rendered in Figure 4.4) using a speech-based STI, we have found a responsiveness of the method towards this distortion, although not as strong as found with the STMI.
The phase jitter distortion can be calculated with:

$$s_{\mathrm{phase-jitter}}(\iota) = \Re\left\{ s(\iota) e^{j\Omega(\iota)} \right\} = s(\iota)\cos(\hat{\Omega}(\iota)), \tag{D.3.3}$$

where $s(\iota)$ is the clean input, $s_{\mathrm{phase-jitter}}(\iota)$ is the distorted output and $\Omega(\iota)$ is the phase jitter function modeled as a uniform random process over $(0 : 2\hat{\Xi}\pi)$ with $0 < \hat{\Xi} < 1$. For $\hat{\Xi} = 1$ the signal becomes a modulated white noise carrier and the intelligibility declines to zero.

Bibliography

Adamy, J., Voutsas, K. and Willert, V. (2003). "Ein binaurales Richtungshörsystem für mobile Roboter in echoarmer Umgebung (A Binaural Sound Localization System for Mobile Robots in Low-reflecting Environments)". *at–Automatisierungstechnik/Methoden und Anwendungen der Steuerungs-, Regelungs-und Informationstechnik*, 51(9/2003), pp. 387–395.

Albani, S., Peissig, J. and Kollmeier, B. (1996). *Psychoacoustics, Speech and Hearing Aids*, chapter Model of binaural localization resolving multiple sources and spatial ambiguities, pages 227–232. World Scientific, Singapore.

Allen, J. B., Berkley, D. A. and Blauert, J. (1977). "Multimicrophone signal-processing technique to remove room reverberation from speech signals". *The Journal of the Acoustical Society of America*, 62(4), pp. 912–915.

ANSI/ASA (S3.5-1997 (R2007)). "American National Standard Methods for Calculation of the Speech Intelligibility Index". Technical report, American National Standards of the Acoustical Society of America.

Bach, J.-H., Anemüller, J. and Kollmeier, B. (2011). "Robust speech detection in real acoustic backgrounds with perceptually motivated features". *Speech Communication*, 53(5), pp. 690–706.

Barker, J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, chapter Robust automatic speech recognition. IEEE Press/Wiley-Interscience.

Barker, J., Josifovski, L., Cooke, M. and Green, P. (2000). "Soft decisions in missing data techniques for robust automatic speech recognition". In *Sixth International*

Conference on Spoken Language Processing (ISCA) 2000, pages 373–376, Beijng, China.

Baskent, D., Eiler, C. L. and Edwards, B. (2007). "Using genetic algorithms with subjective input from human subjects: Implications for fitting hearing aids and cochlear implants". *Ear and hearing*, 28(3), pp. 370380.

Beerends, J. G., van Buuren, R., van Vugt, J. and Verhave, J. (2009). "Objective speech intelligibility measurement on the basis of natural speech in combination with perceptual modeling". *Journal of the Audio Engineering Society*, 57(5), pp. 299–308.

Beutelmann, R. and Brand, T. (2006). "Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners". *The Journal of the Acoustical Society of America*, 120(1), pp. 331–342.

Beutelmann, R., Brand, T. and Kollmeier, B. (2010). "Revision, extension and evaluation of a binaural speech intelligibility model". *The Journal of the Acoustical Society of America*, 127(4), pp. 2479–2497.

Bitzer, J. and Simmer, K. U. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*, chapter Superdirective Microphone Arrays. Springer-Verlag.

Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press.

Blauert, J. (2011). "Epistemological bases of binaural perception - a constructivists' approach". In *Forum Acusticum 2011*, Aalborg, Denmark.

Bodden, M. (1993). "Modeling human sound source localization and the cocktail-party-effect". *Acta Acustica*, 1(1), pp. 43–55.

Boldt, J. B. and Ellis, D. P. W. (2009). "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation". In *17th European Signal Processing Conference (EUSIPCO)*, pages 1849–1853, Glasgow, Scotland.

Boone, M. M. (2006). "Directivity measurements on a highly directive hearing aid: the hearing glasses". In *AES 120th Convention, Paris, France*.

Boone, M. M., Opdam, R. C. G. and Schlesinger, A. (2010). "Downstream speech enhancement in a low directivity binaural hearing aid". In *Proceedings of 20th International Congress on Acoustics, ICA*, Sydney, Australia.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. Springer, Medizinverlag.

Brainard, M. S., Knudsen, E. I. and Esterly, S. D. (1992). "Neural derivation of sound source location: resolution of spatial ambiguities in binaural cues". *The Journal of the Acoustical Society of America*, 91(2), pp. 1015–1027.

Brand, T. and Kollmeier, B. (2002). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests". *The Journal of the Acoustical Society of America*, 111(6), pp. 2801–2810.

Bregman, A. S. (1990). *Auditory scene analysis: the perceptual organization of sound*. The MIT Press.

Breithaupt, C. and Martin, R. (2008). *Advances in Digital Speech Transmission*, chapter Noise Reduction-Statistical Analysis and Control of Musical Noise. John Wiley & Sons Ltd.

Bronkhorst, A. W. (2000). "The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions". *Acta Acustica united with Acustica*, 86, pp. 117–128.

Brown, G. and Wang, D. L. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, chapter Neural and perceptual modeling. IEEE Press/Wiley-Interscience.

Brown, G. J. and Palomaki, K. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, chapter Reverberation. IEEE Press/Wiley-Interscience.

Cherry, E. C. (1953). "Some experiments on the recognition of speech, with one and two ears". *The Journal of the Acoustical Society of America*, 25(5), pp. 975–979.

Christiansen, C., Pedersen, M. and Dau, T. (2010). "Prediction of speech intelligibility based on an auditory preprocessing model". *Speech Communication*, 52(7-8), pp. 678–692.

Dau, T., Kollmeier, B. and Kohlrausch, A. (1997). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers". *The Journal of the Acoustical Society of America*, 102(5), pp. 2892–2905.

De Vries, D., Hulsebos, E. M. and Baan, J. (2001). "Spatial fluctuations in measures of spaciousness". *The Journal of the Acoustical Society of America*, 110(2), pp. 947–954.

Desloge, J., Rabinowitz, W. and Zurek, P. (1997). "Microphone-array hearing aids with binaural output-Part I: Fixed-processing systems". *IEEE Transactions on Speech and Audio Processing*, 5(6), pp. 529–542.

Dietz, M., Ewert, S. D. and Hohmann, V. (2009). "Lateralization of stimuli with independent fine-structure and envelope-based temporal disparities". *The Journal of the Acoustical Society of America*, 125(3), pp. 1622–1635.

Dillon, H. (2001). *Hearing aids*. Thieme Medical Pub., Stuttgart.

Dörbecker, M. and Ernst, S. (1996). "Combination of Two–Channel Spectral Subtraction and Adaptive Wiener Post–Filtering for Noise Reduction and Dereverberation". In *Proceedings of the European Signal Processing Conference (EUSIPCO) 1996*, pages 995–998, Trieste, Italy.

Duquesnoy, A. J. (1983). "Effect of a single interfering noise or speech source upon the binaural sentence intelligibility of aged persons". *The Journal of the Acoustical Society of America*, 74(3), pp. 739–743.

Duquesnoy, A. J. and Plomp, R. (1983). "The effect of Hearing Impairment on the Speech-Reception Threshold of Hearing-Impaired Listeners in Quiet and Noise". *The Journal of the Acoustical Society of America*, 73(6).

Durant, E., Wakefield, G., Van Tasell, D. and Rickert, M. (2004). "Efficient perceptual tuning of hearing aids with genetic algorithms". *IEEE Transactions on Speech and Audio Processing*, 12(2), pp. 144–155.

Durlach, N. I. (1960). "Note on the equalization and cancellation theory of binaural masking level differences". *The Journal of the Acoustical Society of America*, 32(8), pp. 1075–1076.

Durlach, N. I. and Colburn, H. S. (1978). *Handbook of perception*, volume 4, chapter Binaural phenomena. New York: Academic Press.

Durlach, N. I. and Pang, X. D. (1986). "Interaural magnification". *The Journal of the Acoustical Society of America*, 80(6), pp. 1849–1850.

Elhilali, M., Chi, T. and Shamma, S. A. (2003). "A Spectro-Temporal Modulation Index (STMI) for assessment of speech intelligibility". *Speech Communication*, 41, pp. 331–348.

Ellis, D. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, chapter Model-based scene analysis. IEEE Press/Wiley-Interscience.

Elzinga, H. (2010). "Speech source localization with binaural CASA approaches". Master's thesis, Technical University of Delft, The Netherlands.

Eneman, K., Leijon, A., Doclo, S., Spriet, A., Moonen, M. and Wouters, J. (2008). *Advances in digital speech transmission*, chapter Auditory-profile-based Physical Evaluation of Multi-microphone Noise Reduction Techniques in Hearing Instruments. John Wiley & Sons Ltd.

Faller, C. and Merimaa, J. (2004). "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence". *The Journal of the Acoustical Society of America*, 116(5), pp. 3075–3089.

FCA (2007). "Federal Cartel Authority Germany: Resolution Administrative Procedure B3-578/06".

Fels, J. (2008). *From Children to Adults: How Binaural Cues and Ear Canal Impedances Grow*. PhD thesis, RWTH Aachen, Germany.

Gaik, W. and Lindemann, W. (1986). "Ein digitales Richtungsfilter, basierend auf der Auswertung interauraler Parameter von Kunstkopfsignalen". In *Fortschritte der Akustik–DAGA*, volume 86, pages 721–724, Oldenburg, Germany.

George, E., Zekveld, A., Kramer, S., Goverts, S., Festen, J. and Houtgast, T. (2007). "Auditory and nonauditory factors affecting speech reception in noise by older listeners". *The Journal of the Acoustical Society of America*, 121(4), pp. 2362–2375.

George, E. L. J. (2007). *Factors affecting speech reception in fluctuating noise and reverberation*. PhD thesis, Vrije Universiteit, The Netherlands.

Gnewikow, D., Ricketts, T., Bratt, G. and Mutchler, L. (2009). "Real-world benefit from directional microphone hearing aids". *Journal of rehabilitation research and development*, 17(23), pp. 29–33.

Goldsworthy, R. L. and Greenberg, J. E. (2004). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations". *The Journal of the Acoustical Society of America*, 116(6), pp. 3679–3689.

Goupell, M. J. and Hartmann, W. M. (2007). *Hearing–From Sensory Processing to Perception*, chapter Interaural Phase and Level Fluctuations as the Basis of Interaural Incoherence Detection. Springer.

Greenberg, J. and Zurek, P. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*, chapter Microphone-Array Hearing Aids. Springer-Verlag.

Hamacher, V., Kornagel, U., Lotter, T. and Puder, H. (2008). *Advances in Digital Speech Transmission*, chapter Binaural Signal Processing in Hearing Aids. John Wiley & Sons Ltd.

Handelsblatt (2010). *Mafia-Methoden bei Hörgeräten: Ohr um Ohr, Zahn um Zahn*. http://bit.ly/8XEtB7.

Hänsler, E. and Schmidt, G. (2004). *Acoustic echo and noise control a practical approach*. John Wiley & Sons Ltd.

Harding, S., Barker, J. and Brown, G. (2005). "Mask estimation for missing data speech recognition based on statistics of binaural interaction". *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1), pp. 58–67.

Hartmann, W. M. (1997). *Signals, sound, and sensation*. Modern Acoustics and Signal Processing. American Institute of Physics.

Hohmann, V. (2008). "Modellbasierte Signalverarbeitung in Hörgeräten". In *Fortschritte der Akustik - DAGA*, pages 21–24, Dresden, Germany.

Holube, I. and Kollmeier, B. (1996). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated prediction model". *The Journal of the Acoustical Society of America*, 100(3), pp. 1703–1716.

Houck, C., Joines, J. and Kay, M. (1995). "A genetic algorithm for function optimization: a Matlab implementation". *North Carolina State University, Raleigh, NC, Technical Report*.

Housseinzadeh, D. and Krishnan, S. (2007). "Combining vocal source and MFCC features for enhanced speaker recognition performance using GMMs". In *Proceedings of the 9th IEEE International Workshop on Multimedia Signal Processing*, Crete, Greece. MMSP.

Hu, G. and Wang, D. L. (2004). "Monaural speech segregation based on pitch tracking and amplitude modulation". *IEEE Transactions on Neural Networks*, 15(5), pp. 1135–1150.

Hu, Y. and Loizou, P. (2007). "A comparative intelligibility study of single-microphone noise reduction algorithms". *The Journal of the Acoustical Society of America*, 122(3), pp. 1777–1786.

Hulsebos, E. (2004). *Auralization using wave field synthesis*. PhD thesis, TU Delft, The Netherlands.

Humes, L. (2002). "Factors underlying the speech-recognition performance of elderly hearing-aid wearers". *The Journal of the Acoustical Society of America*, 112(3), pp. 1112–1132.

Jeffress, L. (1948). "A place theory of sound localization". *Journal of comparative and physiological psychology*, 41(1), pp. 35–39.

Jeub, M., Schafer, M. and Vary, P. (2009). "A binaural room impulse response database for the evaluation of dereverberation algorithms". In *IEEE 16th International Conference on Digital Signal Processing, 2009*, pages 550–554, Santorini, Greece.

Kates, J. and Arehart, K. (2010). "The Hearing-Aid Speech Quality Index (HASQI)". *Journal of the Audio Engineering Society*, 58(5), pp. 363–381.

Kates, J. M. and Arehart, K. H. (2005a). "A model of speech intelligibility and quality in hearing aids". In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 53–56, New Paltz, United States of America.

Kates, J. M. and Arehart, K. H. (2005b). "Coherence and the speech intelligibility index". *The Journal of the Acoustical Society of America*, 117(4), pp. 2224–2237.

Ketwaru, S. (2010). "Localization with hearing aids in noisy conditions using a Bayesian algorithm". Bachelor thesis (not published), Technical University of Delft, The Netherlands.

Kim, G. and Loizou, P. C. (2010). "A new binary mask based on noise constraints for improved speech intelligibility". In *Proceedings of the Interspeech Conference*, Makuhari, Japan.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T. and Wang, D. L. (2009). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech". *The Journal of the Acoustical Society of America*, 126(3), pp. 1415–1426.

Kocinski, J., Libiszewski, P. and Sek, A. (2011). "Spatial efficiency of blind source separation based on decorrelation - subjective and objective assessment". *Speech Communication*, 53(3), pp. 390–402.

Kollmeier, B. and Koch, R. (1994). "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction". *The Journal of the Acoustical Society of America*, 95(3), pp. 1593–1602.

Kollmeier, B., Peissig, J. and Hohmann, V. (1993). "Real-time multiband dynamic compression and noise reduction for binaural hearing aids". *Journal of rehabilitation research and development*, 30, pp. 82–82.

Kolossa, D. (2011). "High-Level Processing of Binaural Features". In *Forum Acusticum 2011*, Aalborg, Denmark.

Leijon, A. (2007). *Hearing–From Sensory Processing to Perception*, chapter Articulation Index and Shannon mutual information. Springer.

Li, J., Sakamoto, S., Hongo, S., Akagi, M. and Suzuki, Y. (2011). "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication". *Speech Communication*, 53(5), pp. 677–689.

Liu, C., Wheeler, B. C., O'Brien Jr, W. D., Bilger, R. C., Lansing, C. and Feng, A. S. (2000). "Localization of multiple sound sources with two microphones". *The Journal of the Acoustical Society of America*, 108(4), pp. 1888–1905.

Lockwood, M., Jones, D., Bilger, R., Lansing, C., O'Brien Jr, W., Wheeler, B. and Feng, A. (2004). "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms". *The Journal of the Acoustical Society of America*, 115(1), pp. 379–391.

Loizou, P. and Kim, G. (2011). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions". *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1), pp. 47–56.

Loizou, P. C. (2007). *Speech enhancement: Theory and practice*. Signal processing and communications. Taylor & Francis Group, LLC.

Lotter, T. and Vary, P. (2006). "Dual-channel speech enhancement by superdirective beamforming". *EURASIP Journal on Applied Signal Processing*, 2006(63297), pp. 1–14.

Ludvigsen, C., Elberling, C., Keidser, G. and Poulsen, T. (1990). "Prediction of Intelligibility of Non-linearly Processed Speech". *Acta Otolaryngol. (Stockh.)*, 469, pp. 190–195.

Madhu, N. (2009a). *Acoustic source localization: Algorithms, applications and extensions to source separation*. PhD thesis, Ruhr-Universität Bochum, Germany.

Madhu, N. (2009b). "Data-Driven Mask Generation for Source Separation". In *International Symposium on Auditory and Audiological Research (ISAAR)*, Marienlyst, Denmark.

Madhu, N. (2009c). "Note on measures for spectral flatness". *Electronics Letters (IET)*, 45(23), pp. 1195–1196.

Madhu, N., Breithaupt, C. and Martin, R. (2008). "Temporal smoothing of spectral masks in the cepstral domain for speech separation". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008*, pages 45–48, Las Vegas, United States of America.

Madhu, N., Spriet, A., Jansen, S. and Wouters, J. (2010). "The myth of the Ideal Binary Mask in speech enhancement". Technical report, ExpORL, Dept. Neurosciences, K.U.Leuven.

Martin, K. (1997). "Echo suppression in a computational model of the precedence effect". In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (ASSP), 1997*, New Paltz, United States of America.

Martin, R. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*, chapter Small Microphone Arrays with Postfilters for Noise and Acoustic Echo Reduction. Springer-Verlag.

Mauler, D. (2010). *Advances in Single-Channel Noise Reduction for Hearing Instruments*. PhD thesis, Ruhr-Universität Bochum, Germany.

Merks, I. (2000). *Binaural Application of Microphone Arrays for Improved Speech Intelligibility in a Noisy Environment*. PhD thesis, Delft University of Technology, The Netherlands.

Mesgarani, N., Shamma, S. and Slaney, M. (2004). "Speech discrimination based on multiscale spectro-temporal modulations". In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2004*, volume 1, pages 601–604, Montreal, Canada.

Miller, G. (1947). "The masking of speech". *Psychological Bulletin*, 44(2), pp. 105–129.

Minnaar, P., Olesen, S. K., Christensen, F. and Møller, H. (2001). "Localization with binaural recordings from artificial and human heads". *Journal of the Audio Engineering Society*, 49(5), pp. 323–336.

Moore, B. C. J. (2003). *An Introduction to the Psychology of Hearing*. Academic Press - An imprint of Elsevier Science.

Nix, J. (2005). *Localization and Separation of Concurrent Talkers Based on Principles of Auditory Scene Analysis and Multi-Dimensional Statistical Methods*. PhD thesis, Carl von Ossietzky Universität Oldenburg, Germany.

Nix, J. and Hohmann, V. (2006). "Sound source localization in real sound fields based on empirical statistics of interaural parameters". *The Journal of the Acoustical Society of America*, 119(1), pp. 463–479.

Opdam, R. (2010). "Binaural CASA algorithm for speech source localization". Master's thesis, Technical University of Delft, The Netherlands.

Oppenheim, A. V. and Schafer, R. W. (1975). *Digital Signal Processing*. Prentice-Hall.

Paliwal, K., Schwerin, B. and Wojcicki, K. (2011). "Role of modulation magnitude and phase spectrum towards speech intelligibility". *Speech Communication*, 53(3), pp. 327–339.

Paliwal, K. and Wojcicki, K. (2008). "Effect of Analysis Window Duration on Speech Intelligibility". *IEEE Signal Processing Letters*, 15, pp. 785–788.

Pavlovic, C. V. (1987). "Derivation of primary parameters and procedures for use in speech intelligibility predictions". *The Journal of the Acoustical Society of America*, 82(2), pp. 413–422.

Payton, K. L., Braida, L. D., Chen, S., Rosengard, P. and Goldsworthy, R. (2002). *Past, Present and Future of the Speech Transmission Index*, chapter Computing the STI using speech as a probe stimulus. TNO Human Factors: Soesterberg, The Netherlands.

Payton, K. L. and Shrestha, M. (2008). "Analysis of short-time speech transmission index algorithms". *The Journal of the Acoustical Society of America*, 123(5), pp. 3071–3071.

Peissig, J. (1992). *Binaurale Hörgerätestrategien in komplexen Störschallsituationen*. PhD thesis, Göttingen Georg-August Universität, Germany.

Pichora-Fuller, K. (2009). "How cognition might influence hearing aid-design, fitting, and outcomes". *The Hearing Journal*, 62(11), pp. 32–38.

Pichora-Fuller, M., Schneider, B. and Daneman, M. (1995). "How young and old adults listen to and remember speech in noise". *The Journal of the Acoustical Society of America*, 97(1), pp. 593–608.

Plomp, R. (1978). "Auditory Handicap of Hearing Impairment and the limited Benefit of Hearing Aids". *The Journal of the Acoustical Society of America*, 63(2), pp. 533–549.

Plomp, R. (1986). "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired". *Journal of Speech and Hearing Research*, 29(2), pp. 146–154.

Plomp, R. and Mimpen, A. (1979). "Improving the reliability of testing the speech reception threshold for sentences". *International Journal of Audiology*, 18(1), pp. 43–52.

Rakerd, B. and Hartmann, W. M. (2010). "Localization of sound in rooms. V. Binaural coherence and human sensitivity to interaural time differences in noise". *The Journal of the Acoustical Society of America*, 128(5), pp. 3052–3063.

Ramirez, J.-P., Raake, A. and Reusch, D. (2009). "Intelligibility assessment method for semantically unpredictable sentences in German". In *Fortschritte der Akustik–DAGA*, pages 1013–1015, Rotterdam, The Netherlands.

Rennies, J., Brand, T. and Kollmeier, B. (2010). "Modellierung binauraler Sprachverstaendlichkeit in verhallter Umgebung". In *Fortschritte der Akustik–DAGA*, pages 989–990, Berlin, Germany.

Rhebergen, K. S. and Versfeld, N. J. (2005). "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuation noise for normal-hearing listeners". *The Journal of the Acoustical Society of America*, 117(4), pp. 2181–2192.

Riesz, R. (1928). "Differential intensity sensitivity of the ear for pure tones". *Physical Review*, 31(5), pp. 867–875.

Rohdenburg, T. (2008). *Development and objective perceptual quality assessment of monaural and binaural noise reduction schemes for hearing aids*. PhD thesis, Carl von Ossietzky Universität Oldenburg, Germany.

Roman, N., Wang, D. L. and Brown, G. J. (2003). "A classification-based cocktail-party processor". *in Proc. of Neural Information Processing Systems*.

Sarampalis, A., Kalluri, S., Edwards, B. and Hafter, E. (2009). "Objective measures of listening effort: Effects of background noise and noise reduction". *Journal of Speech, Language, and Hearing Research*, 52(5), pp. 1230–1240.

Schimmel, S. M., Atlas, L. E. and Nie, K. (2007). "Feasibility of single channel speaker separation based on modulation frequency analysis". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2007*, volume 4, Honolulu, United States of America. IEEE.

Schlesinger, A. (2012). "Transient-based Speech Transmission Index for predicting intelligibility in nonlinear speech enhancement processors". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012 (submitted)*, Kyoto, Japan.

Schlesinger, A. and Boone, M. M. (2008). "Improving speech intelligibility based on a conjunction of multiple perceptual models". In *The Journal of the Acoustical Society of America*, volume 123, pages 3722–3722.

Schlesinger, A. and Boone, M. M. (2010). "Speech intelligibility assessment in binaural and nonlinear hearing aids". In *2nd Workshop on Speech in Noise: Intelligibility and Quality*, VU University Medical Center Amsterdam.

Schlesinger, A., Ramirez, J.-P. and Boone, M. M. (2010). "Evaluation of a speech-based and binaural Speech Transmission Index". In *Proceedings of the AES 40th Conference on Spatial Audio*, Tokyo, Japan.

Schlesinger, A., Ramirez, J.-P., Van Dorp Schuitman, J. and Boone, M. M. (2009). "Report on a binaural extension of the Speech Transmission Index method for nonlinear systems and narrowband interference". In *International Symposium on Auditory and Audiological Research*, Marienlyst, Denmark. ISAAR.

Schmitz, A. (1995). "Ein neues digitales Kunstkopfmesssystem". *Acta Acustica united with Acustica*, 81, pp. 416–420.

Scholz, K. (2008). *Instrumentelle Qualitätsbeurteilung von Telefonsprache beruhend auf Qualitätsattributen*. PhD thesis, Christian-Albrechts-Universität zu Kiel, Germany.

Seltzer, M., Tashev, I. and Acero, A. (2007). "Microphone Array Post-Filter using Incremental Bayes Learning to Track the Spatial Distributions of Speech and Noise". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2007*, Honolulu, United States of America.

Shield, B. (2006). *Evaluation of the Social and Economic Costs of Hearing Impairment*. Hear-it AISBL.

Simmer, K. U., Bitzer, J. and Marro, C. (2001). *Microphone Arrays: Signal Processing Techniques and Applications*, chapter Post-Filtering Techniques. Springer-Verlag.

So, S. and Paliwal, K. (2010). "Single-Channel Speech Enhancement Using Kalman Filtering in the Modulation Domain". In *Proceedings of the Interspeech Conference*, Makuhari, Japan.

Soede, W., Berkhout, A. J. and Bilsen, F. A. (1993). "Development of a directional hearing instrument based on array technology". *The Journal of the Acoustical Society of America*, 94(2), pp. 785–798.

Steeneken, H. and Houtgast, T. (2002). "Basics of the STI measuring method". *Past, present and future of the Speech Transmission Index. Soesterberg, the Netherlands: TNO Human Factors*, pages 13–43.

Steeneken, H. J. M. and Houtgast, T. (1980). "A physical method for measuring speech-transmission quality". *The Journal of the Acoustical Society of America*, 67(1), pp. 318–326.

Stern, R., Brown, G. and Wang, D. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, chapter Binaural sound localization. IEEE Press/Wiley-Interscience.

Taal, C., Hendriks, R., Heusdens, R. and Jensen, J. (2011a). "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech". *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), pp. 2125 –2136.

Taal, C. H., Hendriks, R. C. and Heusdens, R. (2010). "A short-time objective intelligibility measure for time-frequency weighted noisy speech". In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010*, pages 4214 –4217, Dallas, United States of America.

Taal, C. H., Hendriks, R. C., Heusdens, R. and Jensen, J. (2011b). "An evaluation of objective measures for intelligibility prediction of time-frequency weighted noisy speech". *The Journal of the Acoustical Society of America (in press)*.

Thiergart, O., Del Galdo, G., Prus, M. and Kuech, F. (2010). "Three-Dimensional Sound Field Analysis with Directional Audio Coding based on Signal Adaptive Parameter Estimators". In *Proceedings of the AES 40th Conference on Spatial Audio*, Tokyo, Japan.

TNO (2000). *Multilingual database*. TNO Human Factors Research Institute, Soesterberg, The Netherlands.

Van de Par, S., Trahiotis, C. and Bernstein, L. R. (2001). "A consideration of the normalization that is typically included in correlation-based models of binaural detection". *The Journal of the Acoustical Society of America*, 109(2), pp. 830–833.

Van Dorp Schuitman, J. (2009). *Shoebox v0.7.1 manual*. TU Delft, Delft, The Netherlands.

Van Wijngaarden, S. J. and Drullman, R. (2008). "Binaural intelligiblity prediction based on the Speech Transmission Index". *The Journal of the Acoustical Society of America*, 123(6), pp. 4514–4523.

Verhey, J. (2008). "Psychophysik, Physiologie und Modelle des binauralen Hörens". In *38. DGMP Tagung*, Oldenburg, Germany.

Vorländer, M. (2008). *Auralization: fundamentals of acoustics, modeling, simulation, algorithms and acoustic virtual reality*, volume 1. Springer.

Wang, D. and Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, chapter Fundamentals of computational auditory scene analysis. IEEE Press/Wiley-Interscience.

Weiss, R. J., Mandel, M. I. and Ellis, D. P. W. (2011). "Combining localization cues and source model constraints for binaural source separation". *Speech Communication*, 53(5), pp. 606–621.

Wittkop, T., Albani, S., Hohmann, V., Peissig, J., Woods, W. and Kollmeier, B. (1997). "Speech processing for hearing aids: Noise reduction motivated by models of binaural interaction". *Acta Acustica united with Acustica*, 83(4), pp. 684–699.

Wittkop, T. and Hohmann, V. (2003). "Strategy-selective noise reduction for binaural digital hearing aids". *Speech Communication*, 39, pp. 111–138.

Woodruff, J., Prabhavalkar, R., Fosler-Lussier, E. and Wang, D. (2010). "Combining Monaural and Binaural Evidence for Reverberant Speech Segregation". In *Proceedings of the Interspeech Conference*, Makuhari, Japan.

Woods, W. S., Hansen, M., Wittkop, T. and Kollmeier, B. (1996). "A simple architecture for using multiple cues in sound separation". In *Fourth International Conference on Spoken Language (ICSLP) 1996*, pages 909–912, Philadelpha, United States of America.

Yoo, S. D., Boston, R., El-Jaroudi, A., Li, C.-C., Durrant, J. D., Kovacyk, K. and Shaiman, S. (2007). "Speech signal modification to increase intelligibility in noisy environments". *The Journal of the Acoustical Society of America*, 122(2), pp. 1138–1149.

You, H., Zhu, Q. and Alwan, A. (2004). "Entropy-based variable frame rate analysis of speech signals and its application to ASR". In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 549–552.

Zhang, P. X. and Hartmann, W. M. (2006). "Lateralization of sine tones–interaural time vs phase". *The Journal of the Acoustical Society of America*, 120(6), pp. 3471–3474.

Zheng, Y., Reindl, K. and Kellermann, W. (2009). "BSS for improved interference estimation for Blind speech signal Extraction with two microphones". In *3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 253 –256, Aruba, Dutch Antilles.

Zwicker, E. and Terhardt, E. (1980). "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency". *The Journal of the Acoustical Society of America*, 68(5), pp. 1523–1525.

Zwicker, E. and Zollner, M. (1987). *Elektroakustik (Electroacoustics), 2nd Edition*. Springer, Berlin, Heidelberg.

List of Symbols and Abbreviations

Symbols (except definitions in the appendix)

α	Smoothing constant
α_{x}	Cepstral time smoothing constant
α_{loE}	Cepstral time smoothing constant low-frequent speech envelope
α_{hiE}	Cepstral time smoothing constant high-frequent speech envelope
α_{p}	Cepstral time smoothing constant F0
α_{n}	Cepstral time smoothing constant noise
α_{γ}	Smoothing constant for estimating the coherence function
α_{PSD}	Smoothing constant PSD signal
$\mathring{\alpha}$	Smoothing constant for estimating MPSD
$\ddot{\alpha}$	Smoothing constant feature vector
β	Order Renyi entropy
$\Delta\gamma$	Normalized absolute magnitude coherence function
δ	Band importance function (speech intelligibility)
ϵ	Lower bound envelope correction for preventing division by zero
ϵ_{v}	Threshold value feature vector
ϵ_{hist}	Threshold in histograms for ITD/IPD/ILD
$\hat{\epsilon}$	Threshold of speech distortion
ε	Mask criterion
ζ	Modulation at the rate of speech
$\hat{\zeta}$	Theoretical modulation at the rate of speech calculated from an SNR
η	Relative criterion mask
θ	Azimuth angle
θ_{t}	Direction of the target signal
ϑ	Elevation angle

ι	Time sample time domain
κ	Stabilization constant (MVDR beamformer)
λ	Wave length [m]
μ	Sample mean
μ_{sb}	Sample mean of intensity envelope of s
μ_{xb}	Sample mean of intensity envelope of x
μ_{zb}	Sample mean of intensity envelope of z
ν	Stabilization constant (Wiener filter)
ξ	Balancing constant between IPD and ILD
o	Crest factor
π	Circle constant
φ	Variable of logistic function
ρ_ι	Feature vector of relative information content (function of time)
ϱ	Time increments of interaural cross correlation (STI)
σ	Sample standard deviation
σ'	Averaged standard deviation over all incidence angles
σ_t	Standard deviation IPD (envelope)
σ_L	Standard deviation ILD (envelope)
ς	Filter (pharynx)
τ	Kendall's τ
$\breve{\tau}$	Time constant [s]
υ	Source (glottis)
ϕ	Power spectral density
ϕ_ss	Power spectral density (target signal)
ϕ_vv	Power spectral density (noise signal)
$\mathring{\phi}$	Modulation power spectral density
$\Delta\varphi$	IPD of the fine-structure
$\Delta\mathring{\varphi}$	IPD of the envelope
χ	Analysis window function
$\mathring{\chi}$	Analysis window function of analysis in STFT domain
ψ	Variable of logistic function
ω	Analysis window STI
$\mathbf{\Gamma}_\mathrm{vv}$	Complex coherence matrix
$\vec{\Delta}$	Feature vector interaural parameters
Λ	Logistic function
Ξ	FIR filter (correction factor algorithm ELT)
$\mathbf{\Pi}_b$	Matrix of rounded-exponential filters
Υ	Mean objective score
$\mathbf{\Phi}_\mathrm{vv}^{-1}$	Inverse cross power spectral density matrix (noise correlation matrix)
Ψ	Steepness of the psychometric intelligibility function
Ω	SRT
$\aleph$	Objective function
$\imath$	Frame index CSII calculation

$\jmath$	Speaker index
ℓ	Microphone index
$\wp$	Arbitrary parameter solution
$\wp_{\clubsuit}$	Final parameter solution of the genetic algorithm
$\mathbf{a}$	Propagation vector
$A_{\min}$	Noise flooring constant
$A_{\max}$	Upper limitation of $\mathcal{M}_{\text{elt}}^{\text{ft}}$
b	Critical band index
c	Speed of sound [m/s]
d	DFT coefficient
d_{x}	Algorithmic parameter of lower cutoff frequency of $\Delta\gamma$ in mask
$d_{\text{x}\varphi}$	Upper cutoff frequency of IPD fine-struture
d_{xL}	Lower cutoff frequency of ILD fine-structure
d_{xo}	Upper cutoff frequency FIR filter (correction factor algorithm ELT)
d_{cb}	Critical band index subsequent averaging in STFT domain
d_{l}	Critical band lower bound in STFT domain
d_{u}	Critical band upper bound in STFT domain
e	Algorithmic parameter for compressing or expanding a mask
$e_{\sigma\text{L}}$	Algorithmic parameter for compression / expansion of $\mathcal{M}_{\sigma\text{L}}$
$e_{\sigma\text{t}}$	Algorithmic parameter for compression / expansion of $\mathcal{M}_{\sigma\text{t}}$
$\mathcal{E}$	Envelope in the STFT domain
$\mathcal{E}'$	Lowpass filtered envelope
E	Expectation operator
f	Frequency [Hz]
f_{s}	Sampling frequency [Hz]
g	Cepstral DFT coefficient
g_{loE}	Cepstral DFT coefficient low-frequent speech envelope
g_{hiE}	Cepstral DFT coefficient high-frequent speech envelope
g_{p}	Cepstral DFT coefficient F0 frequency
$g_{\text{p}-\text{low}}$	Cepstral DFT coefficient lower F0 frequency bound
$g_{\text{p}-\text{high}}$	Cepstral DFT coefficient upper F0 frequency bound
h	Transfer-function
h_{post}	Transfer-function post-filter
i	One-third octave filter band index
$\mathbf{I}$	Identity matrix
j	Imaginary unit
k	Wave number [m^{-1}]
l	Distance between microphones [m]
L	Search space
ΔL	Interaural phase differences of the fine-structure waveform
$\Delta\mathring{L}$	ILD of envelope
$\Delta\tilde{L}$	Magnified ILD parameter (envelope)
m	Modulation frequency index

m_{xo}	Lower cutoff frequency for application of ITD (envelope)
m_{cb}	Critical modulation band
m_{l}	Critical modulation band lower bound
m_{u}	Critical modulation band upper bound
m'	Maximum interaural parameter range of the mean
M_{ι}	Amount of short-time frames (CSII)
$\mathcal{M}_{\mathrm{soft}}$	Soft-mask
$\mathcal{M}_{\mathrm{IBM}}$	Ideal binary mask
$\mathcal{M}$	Mask
$\mathcal{M}_{\mathrm{c}}$	Mask subsequent cepstral smoothing
$\mathcal{M}_{\mathrm{cc}}$	Mask of algorithm CC
$\mathcal{M}_{\mathrm{clp}}$	Mask of algorithm CLP
$\mathcal{M}_{\mathrm{elt}}^{\mathrm{ff}}$	Weighting function algorithm ELT in the modulation domain
$\mathcal{M}_{\mathrm{elt}}^{\mathrm{ft}}$	Weighting function algorithm ELT in the STFT domain
$\mathcal{M}_{\sigma\mathrm{L}}$	Mask based on σ_{L}
$\mathcal{M}_{\sigma\mathrm{t}}$	Mask based on σ_{t}
$\mathcal{M}_{\Delta\mathrm{L}}$	Mask of algorithm ELT based on ILD (envelope)
$\mathcal{M}_{\Delta\mathrm{t}}$	Mask of algorithm ELT based on ITD (envelope)
n	Short-time frame index STFT domain
n_{L}	Bin size of a two-dimensional matrix for smoothing ILD
n_{φ}	Bin size of a two-dimensional matrix for smoothing IPD
n_{t}	Bin size of a two-dimensional matrix for smoothing ITD
$n_{\sigma\mathrm{t}}$	Bin size of a two-dimensional matrix for estimating σ_{t}
$n_{\sigma\mathrm{L}}$	Bin size of a two-dimensional matrix for estimating σ_{L}
N	Amount of transducers
$N_{\hat{\chi}}$	Length of analysis window function of analysis in STFT domain
N_{dd}	Length of DFT for transformation to modulation domain
N_{d}	Length DFT
N_{ι}	Length of a speech token
N_{χ}	Length of the analysis window function (prior zero padding)
$N_{1/3}$	Amount of one-third octave band filters
N_{θ}	Number of azimuthal angles
N_{lp}	Order FIR filter (correction factor algorithm ELT)
$N_{\sum \mathrm{d}}$	Amount of DFT coefficients of a certain critical band
o	Modulation frame index
p	A posteriori probability
p_{s}	Likelihood of the psychometric function for understanding SUS
q	Tab index FIR filter (correction factor algorithm ELT)
r_{RT}	Reverberation radius [m]
r^2	Squared correlation measure
RT	Reverberation time [s]
s	Target speech signal
$\hat{s}$	Probability density function

$\Delta\mathring{t}$	ITD of envelope
ΔT	Frame shift analysis window
$\Delta\mathring{T}$	Modulation analysis window frame shift
u	Index of azimuthal positions
v	Noise signal
V	Room volume [m^3]
w	Single-channel filter
w_{post}	Single-channel post-filter
$\mathbf{w}$	Filter coefficients
x	Single channel input signal (STFT domain)
$\tilde{x}$	Single channel input signal (time domain)
$\mathring{x}$	Centre frequency and modulation spectrum
y	Speech enhanced signal
z	Difference measure between clean and degraded envelope

Abbreviations

ASA	Auditory Scene Analysis
aSNR	apparent Signal to Noise Ratio
BILD	Binaural Intelligibility Level Difference
BSS	Blind Source Separation
BTE	Behind The Ear
CASA	Computational Auditory Scene Analysis
CC	Carrier Coherence algorithm
CLP	Carrier Level Phase algorithm
CSII	Coherence based Speech Intelligibility Index
D50	Definition
DFT	Discrete Fourier Transform
DI	Directivity Index
EC	Equalization Cancellation
ELT	Envelope Level Time algorithm
ERB	Equivalent Rectangular Bandwidth
F0	Pitch frequency of speech
FR	Front Random index
GA	Genetic Algorithm
GSC	Generalized Sidelobe Canceler
HG	Hearing Glasses
HL	Hearing loss
HRTF	Head Related Transfer-Function
I3	Three-level weighted Intelligibility measure
IBM	Ideal Binary Mask
IDFT	Inverse Discrete Fourier Transform

ILD	Interaural Level Differences
IPD	Interaural Phase Differences
ITD	Interaural Time Differences
ITE	In The Ear
JND	Just Noticeable Differences
LSO	Lateral Superior Olive
MAP	Maximum A Posteriori
MPSD	Modulation Power Spectral Density
MF	Madhu Flatness measure
MISM	Mirror Image Source Model
MMSE	Minimum Mean Square Error
MSC	Magnitude Squared Coherence
MSO	Medial Superior Olive
MVDR	Minimum Variance Distortionless Response
$\mathrm{PDF_a}$	Probability Density Function of all sources
$\mathrm{PDF_t}$	Probability Density Function of the target source
PESQ	Perceptual Evaluation of Speech Qualitiy
Q3	Three-level weighted Quality measure
RE	Renyi Entropy
RMS	Root Mean Square
RT	Reverberation Time
SBE	Shannon Band Energy
SE	Shannon Entropy
SI	Speech Intelligibility
SII	Speech Intelligibility Index
SNR	Signal to Noise Ratio
$\mathrm{SNR_l}$	Local Signal to Noise Ratio (per time-frequency bin)
$\mathrm{SNR_{seg}}$	Segmental Signal to Noise Ratio
SSN	Speech Shaped Noise
SRT	Speech Reception Threshold
SPL	Sound Pressure Level
STI	Speech Transmission Index
$\mathrm{STI_{bin}}$	binaural Speech Transmission Index
$\mathrm{STI_{mon}}$	monaural Speech Transmission Index
STFT	Short-Time Fourier Transform
STMI	Spectro-Temporal Modulation Index
STOI	Short Time Objective Intelligibility measure
STRF	Spectro-Temporal Response Fields
SUS	Semantically Unpredictable Sentence
TMTF	Temporal Modulation Transfer-Function
VAD	Voice Activity Detection
WNG	White Noise Gain

Abstract

In Europe about one fifth of the population has difficulties with understanding speech in noisy and complex environments. Improving speech intelligibility in these conditions allows for the reintegration of the hearing impaired into a communication-oriented society and restores individual well-being to a high degree.
Commercially available hearing aid solutions are generally based on the amplification principle and successfully enhance speech understanding for severe grades of a hearing loss in silence. However, current hearing aid solutions do not restore speech intelligibility in noisy surroundings to an extent that is required by the majority of the hearing impaired.
Successful solutions that reconstruct the intelligibility of noise-corrupted speech are based on the principle of spatial sampling. By such means, a target speaker can be enhanced, whereas interference can be suppressed.

In this thesis, a set of standard binaural speech processors, that are based upon models of the auditory scene analysis, are revised, optimized and compared. The binaural speech processors are, furthermore, applied at the output of hearing aids with and without beamforming. As a result, two efficient spatial sampling schemes are combined to gain a high improvement of speech intelligibility in noisy environments.

The conjunction of statistical principles with perceptually motivated algorithms is one of the core focusses of this thesis. A broad statistical study on binaural parameters in different acoustic real-world scenes is given. Binaural parameters of the fine-structure of the waveform are compared to the binaural parameters of the envelope of the waveform. In addition, natural binaural parameters are compared to binaural parameters at the output of different hearing aids and directivity modes. As a result, the study provides a comprehensive insight into the behavior of binaural

parameters in noise, thereby sizing the possibilities of a binaural parameter-based noise suppression.
While earlier approaches of binaural speech processors generally employed primitive grouping schemes in the noise suppression task, a set of binaural speech processors presented here is equipped with a pattern-based a posteriori classification approach. By this approach, the auditory scene analysis is not only carried out in terms of different bottom-up approaches, but also with regard to a simplified top-down pattern-driven method. It is well-known that both processes—in its full extent—form a complement which underlies the unsurpassed capabilities of the auditory scene analysis.

Furthermore, a stochastic optimization of binaural speech processors at the output of different front-ends as well as in different acoustic environments is performed. To that end, a genetic algorithm is applied, which maximizes an objective function of binaural speech intelligibility. Subsequently, the robustness of the optimized binaural speech processors is assessed while changing acoustic scenes.
As will be shown, the holistic approach of a model-based improvement and a model-based assessment of speech intelligibility offers an efficient and task-oriented means for the improvement of speech intelligibility. However, an unsolved problem amounts to the development of an objective function of binaurally and nonlinearly processed speech. To date, there exists no comprehensive model of speech intelligibility.
In this thesis, a broad study on binaurally and nonlinearly processed speech is aimed at making an advance towards such a model. Derived from a series of listening tests, different models of speech intelligibility are presented, developed and compared.
In addition, the efficient cepstral smoothing technique is supplemented to the speech enhancement methods in this work. Cepstral smoothing allows for a suppression of musical noise, which is an unavoidable consequence of varying filter gain-functions. The method will be optimized for speech intelligibility and assessed as a second post-processor of the combined processing scheme.

Considering the results of this thesis, an important consequence for the application of binaural speech processors constitutes the fact that these are predominantly capable to suppress lateral coherent sound sources. As it has been shown in the present study, if binaural speech processors are applied at the output of beamformers, which are generally optimized to suppress diffuse interference, a complementary processing scheme can be designed. An estimate, which has been given in this study, of the benefit of binaural speech processors in terms of speech intelligibility shows an absolute improvement of more than 40 % in coherent interference conditions and an absolute improvement of up to 20 % in diffuse noise conditions.
Furthermore, the study has shown that binaural speech processors, which use the binaural differences of the fine-structure of the waveform as a classifier in the course of source separation, outperform existing binaural modulation-based and binaural coherence-based processors. No disadvantage to the performance of binaural speech processors has been found when binaural beamforming front-ends are applied. Moreover, the pattern-based noise suppression approach and the genetic speech intelli-

gibility optimization procedure have demonstrated to produce robust and efficient binaural speech processors, which—similarly to the model hearing process—show a high degree of plasticity with respect to a particular front-end and a particular sound scene.

With regards to the formulation of a comprehensive speech intelligibility measure, a binaural and speech-based Speech Transmission Index has been developed. Although the measure widely corresponds to subjective scores of binaurally presented speech by modeling the binaural interaction process as well as the head shadow effect, the measure has been shown to fail in the assessment of nonlinear binaural speech processors. In a follow-up study, a compromise for both of these objectives, i.e. binaural processing and handling of nonlinearity, has been incorporated in one measure, that was subsequently applied in the optimization tasks throughout this work.

The present study has been incorporated into a research project on continuity preserving signal processing (CPSP), that was supported by the Dutch Technology Foundation (STW). The project has its origins in the collaboration of two schools, the language, sound and cognition group at the KU Groningen and the sound control group at the TU Delft. The entire project comprises—besides the study at hand—the fields of keyword spotting in automatic speech recognition, the objective assessment of room acoustics and machine monitoring.

This thesis concludes a successful period of auditory research at the Section of Acoustical Imaging and Sound Control at TU Delft. Whereas former projects studied the audiological benefit of array technology and resulted in a market-launch of a beamforming solution, well-known as the hearing glasses of Varibel Innovations BV, the present study has broadened the scope to perceptually motivated principles. Due to the heterogeneity of the research field, we have opted to deliver a basic study on binaural speech intelligibility enhancement and assessment as opposed to a development of solitary solutions that lack generalizability. This way, we hope the present study lays the foundations for further advancements.

Samenvatting

Eén vijfde van de bevolking van Europa heeft moeite met het verstaan van spraak in lawaaiige en drukke omgevingen. Het verbeteren van de spraakverstaanbaarheid onder deze omstandigheden geeft slechthorenden de kans te re-integreren in een samenleving waar verbale communicatie zo belangrijk is, en herstelt het individueel welzijn in sterke mate.
De werking van commercieel verkrijgbare hoortoestellen is over het algemeen gebaseerd op het versterkingsprincipe en is daarmee succesvol in het verbeteren van de spraakverstaanbaarheid in rustige omgevingen. De huidige hoortoestellen zijn echter voor de meeste slechthorenden niet afdoende om de spraakverstaanbaarheid in lawaaiige omgevingen te verbeteren.
Effectieve oplossingen die de verstaanbaarheid van door omgevingsgeluid vervormde spraak verbeteren, werken met het principe van ruimtelijke bemonstering. Op deze manier kan de spraak van een bepaalde spreker worden versterkt en tegelijkertijd omgevingslawaai worden teruggedrongen.

In dit proefschrift wordt een aantal standaard binaurale spraakprocessoren die berusten op modellen voor auditieve omgevingsanalyse, bekend als 'computational auditory scene analysis' (CASA), gereviseerd, geoptimaliseerd en vergeleken. De binaurale processoren worden aansluitend toegepast op de uitgang van hoortoestellen met en zonder richtingswerkende bundelvorming. Uiteindelijk worden twee efficiënte ruimtelijke bemonsteringsmethoden gecombineerd om een grote winst in spraakverstaanbaarheid in ruis (omgevingslawaai) te behalen.

De combinatie van statistische principes met perceptieve algoritmes is een van de kernpunten van deze dissertatie. De binaurale parameters van verscheidene auditieve scenario's uit het dagelijkse leven worden uitgebreid statistisch onderzocht. Binaurale parameters van de fijnstructuur van het geluidsignaal worden vergeleken met de

binaurale parameters van de omhullende. Ook worden natuurlijke binaurale parameters vergeleken met die van de uitgangssignalen van hoortoestellen met verschillende richtinggevoeligheid. Het resultaat is dat de studie een uitgebreid inzicht geeft in de eigenschappen van binaurale parameters in ruis en daarmee een inschatting mogelijk maakt in hoeverre onderdrukking van omgevingsgeluid met deze methode mogelijk is.

Terwijl eerdere benaderingen van binaurale spraakprocessoren over het algemeen primitieve groeperingsmethoden gebruikten voor ruisreductie, wordt hier een aantal binaurale spraakprocessoren gepresenteerd die gebruik maken van a posteriori classificatie. Op deze wijze wordt CASA niet alleen vanuit verschillende bottom-up benaderingen, maar ook vanuit een vereenvoudigde top-down benadering verkregen. Het is algemeen bekend dat beide processen elkaar aanvullen, wat daarmee grote mogelijkheden biedt voor de toepassing van CASA.

Aansluitend wordt een stochastische optimalisatie uitgevoerd van binaurale spraakprocessoren op de uitgangssignalen van verschillende hoortoestellen die dienst doen als front-ends. Om dit te doen is er een genetisch algoritme toegepast, dat de binaurale spraakverstaanbaarheid maximaliseert. Vervolgens wordt de robuustheid van de geoptimaliseerde binaurale spraakprocessoren geëvalueerd onder verschillende akoestische omstandigheden.
Zoals zal worden aangetoond, is de holistische aanpak van verbetering en beoordeling van modelgebaseerde spraakverstaanbaarheid een efficiënte en taakgebaseerde methode. Toch blijft het objectief beoordelen van de binaurale en niet-lineair bewerkte spraak een onopgelost probleem. Tot de dag van vandaag bestaat er geen alomvattend beoordelingsmodel voor spraakverstaanbaarheid. In dit proefschrift wordt een stap gezet naar zo'n model door een brede studie naar binaurale en niet-lineair bewerkte spraak uit te voeren. Vanuit een aantal luisterexperimenten worden verscheidene modellen voor spraakverstaanbaarheid gepresenteerd, ontwikkeld en vergeleken.
Ook wordt de efficiënte cepstrale middelingsmethode toegevoegd aan de spraakverbetermethoden in dit werk. Cepstrale middeling biedt mogelijkheden voor het onderdrukken van tonale ruis; een niet te vermijden gevolg van de tijd- en frequentieafhankelijke signaalversterking. De methode wordt geoptimaliseerd voor spraakverstaanbaarheid en beoordeeld als een tweede postprocessor van het gecombineerde verwerkingsschema.

De resultaten van dit proefschrift maken duidelijk dat het toepassen van binaurale spraakverwerking voornamelijk geschikt is om laterale coherente geluidsbronnen te onderdrukken. Zoals is aangetoond in deze studie, kan daarmee een effectief verwerkingsschema worden ontworpen voor binaurale spraakprocessoren die aanvullend worden toegepast op het signaal van bundelvormende front-ends, aangezien bundelvormers over het algemeen geoptimaliseerd zijn in het onderdrukken van diffuse interferentie. Een schatting die in deze studie wordt gegeven voor de mate waarin binaurale spraakbewerking de spraakverstaanbaarheid ten goede komt, toont een

absolute verbetering aan van ruim 40 % bij coherente ruis, en een absolute verbetering tot 20 % onder diffuse ruiscondities.
De studie heeft ook aangetoond dat binaurale spraakprocessoren die de binaurale verschillen in de fijnstructuur van de golfvorm gebruiken bij het uit elkaar halen van de bronnen, beter presteren dan bestaande binaurale processoren die gebaseerd zijn op de omhullende van het signaal en op coherentie. Er is geen verslechtering van de spraakverstaanbaarheid geconstateerd wanneer binaurale spraakprocessoren worden gecombineerd met richtinggevoelige front-ends. Bovendien hebben de op patronen gebaseerde bronscheidingsaanpak en de genetische procedure voor het verbeteren van spraakverstaanbaarheid bewezen robuuste en efficiënte binaurale spraakprocessoren op te leveren, die—net als bij het modelgebaseerd hoorproces—een hoge mate van flexibiliteit demonstreren bij verscheidene front-ends en geluidsscenarios.

Om een complete oplossing voor spraakverstaanbaarheid te formuleren, zijn een binaurale en een spraakgebaseerde Speech Transmission Index ontwikkeld. Hoewel de oplossing in de meeste gevallen overeenkomt met subjectieve resultaten van binauraal gepresenteerde spraak door het binaurale interactieproces en de akoestische schaduwwerking van het hoofd te modelleren, is aangetoond dat de oplossing tekortschiet in het beoordelen van niet-lineaire binaurale spraakprocessoren. In een vervolgonderzoek is een compromis voor beide doelen—het binauraal verwerken en het omgaan met niet-lineariteit—samengevoegd in één maat, die vervolgens is toegepast in de optimalisatietaken in dit werk.

De huidige studie is ingepast in een onderzoeksproject over ‘continuity preserving signal processing’ (CPSP), dat werd ondersteund door de Technologiestichting STW. Het project heeft zijn oorsprong in de samenwerking tussen twee scholen, de language, sound and cognition-groep aan de KU Groningen en de sound control-groep aan de TU Delft. Waar vorige projecten de audiologische voordelen van array technologie bestudeerden en resulteerden in het op de markt brengen van een richtinggevoelige hooroplossing, welbekend als de hoorbril van Varibel Innovations BV, heeft de voorliggende studie het werkgebied verbreed naar principes die zijn geënt op perceptie. Door de breedte van het onderzoeksveld hebben we gekozen voor een fundamentele studie naar het verbeteren en het beoordelen van binaurale spraakverstaanbaarheid in plaats van het ontwikkelen van specialistische oplossingen die niet breed toepasbaar zijn. Op deze manier hopen we dat de voorliggende studie de basis legt voor toekomstige ontwikkelingen.

About the Author

Anton Schlesinger was born in Dresden, Germany, on October 3rd, 1977. He attended secondary school and received his 'Abitur' in 1998. In connection with his military service, which he spent as a clerk in a hospital, he entered a training course to become a professional purchase executive monitored by the German Chamber of Commerce and graduated in 2001. During that period he was enrolled as a remote student at the University of Hagen, Germany, and followed courses in politics.

In the academic winter terms 2000/01, he started studying Media Technology at Ilmenau University of Technology, Germany. In 2004/05 he took a one year break from his study for an internship as a consultant for classical room acoustics of recording studios and venues at the Walters-Storyk Design Group in Basel, Switzerland. Subsequently, he has been engaged as a founding member of a start-up for location-based services in Weimar, Germany. In August 2006, he received his Diplom in Media Technology with his thesis work on the three-dimensional measurement and holographic reconstruction of sound fields.

At the end of 2006, he joined the Laboratory of Acoustical Imaging and Sound Control at the Delft University of Technology, The Netherlands. In the succeeding four years, he conducted research in the field of audiology that resulted in this PhD thesis.

From the beginning of 2011, Anton Schlesinger is employed at the Institute of Communication Acoustics at the Ruhr-Universität Bochum, Germany, as a postdoctoral researcher. His current scientific interest is in binaural models of speech intelligibility and sound scene classification.

Acknowledgement

I like to remember a conversation with our dear colleague Aad van den Bos on the western orientation of science. He told me about his travels into the opposite direction and his contacts to scientists in Russia after the fall of the wall. He had learned the Russian language, and giving an example of the adorability of the Russian nature, he told what he had seen on TV: “After the return to ground, the cosmonaut had been dragged out of his space ship and was asked by a reporter for new insights. His answer: No God.”—Motivated by this highlighted laconism, I would like to present the acknowledgements for this work of the past five years with similar brevity, pondering that many more words could be written but not much more could be said.

I am very grateful to Marinus Boone for his guidance, help and critical judgement throughout the work on this thesis as a supervisor. My colleagues Lars Hörchens and Jasper van Dorp-Schuitman regarded this work with profound interest and support. Therefore I am deeply thankful. Juan-Pablo Ramirez conducted the listening tests of this thesis at Telekom Laboratories in Berlin. His excellent company has been a blessing. I am much indebted to Diemer de Vries and my promotor Dries Gisolf for their recognition of my work in room acoustics, their choice for me as a PhD student, as well as their generous support throughout the study.

Many thanks to the Dutch foundation of Technology and the helpful user committee, whose assemblies I appreciated. Furthermore, this work benefitted greatly from the help of Tjierd Andringa, Bea Valkenier, Hedde van de Vooren, Armin Kohlrausch, Joost Festen, Margaret van Fessem, Eric Verschuur, Mohammad Al-Bannagi, Alexey Kononov, Edo Bergsma, Henry den Bok, Gerrit van Dijk, Susan Pesman, Sarmad Malik, Christian Luther, Nilesh Madhu, Hendrik Elzinga, Cees Taal, Rob Opdam, Salim Ketwaru, Simone Romanow, Gert-Peter Gooiker, Sebastian Gergen, Rainer Martin, Dorothea Kolossa, Jens Blauert, Herbert Hudde, Brit Hopmann, Markus Haase, Cornelia Albring, Laurentiu Giogu, Wieland Sack, Ola Kopka, Sarah Maria Ruhrort and my parents, Beatrix and Helmut Schlesinger.

ISBN 987-94-6186-020-0