Instrumental and Perceptual Evaluation of Hearing Devices — Methods and Applications

Von der Fakultät für Mathematik und Naturwissenschaften
der Carl von Ossietzky-Universität Oldenburg
zur Erlangung des Grades und Titels eines
Doktors der Naturwissenschaften (Dr. rer. nat.)
angenommene Dissertation

von
Christoph Völker (M. A.)
geboren am 22. März 1981
in Kiel

Gutachter: Prof. Dr. Dr. Birger Kollmeier

Zweitgutachter: Prof. dr. ir. Wouter A. Dreschler

Tag der Disputation: 26. August 2016

So habe auch ich's für gut gehalten, nachdem ich alles von Anfang an sorgfältig erkundet habe, es für dich, hochgeehrter Theophilus, in guter Ordnung aufzuschreiben.
(Evangelium nach Lukas 1,3; Luther-Übersetzung)

Abstract

The present thesis contributes towards the long-term goal of providing optimum evaluation and fitting methods for hearing devices that are based on the needs and preferences of the individual user. As an efficient and comprehensive fitting procedure should include both, instrumental and perceptual assessments of each setting of the hearing device, the current thesis considers the development and application of fitting methods utilizing both components. The first study describes the comparison of various pre-processing schemes. Among other things, the study serves as an initial point for further method developments within this thesis as it demonstrates that single auditory model-based measures are not able to predict the algorithm performance for the individual aided patients satisfactorily. Consequently, the second study focuses on perceptual test procedures for real persons. Two modifications for the standardized MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) were developed to increase the accessibility especially for elderly, technically non-experienced users with hearing impairments. It is shown that using the modification MUSHRA drag&drop leads to a slightly higher percentage of reliable data of the 50 tested persons than using the classical MUSHRA procedure. Emphasizing instrumental assessments again, the third study aims at applying instrumental auditory model-based measures for determining optimal parameter settings for complex, multidimensional algorithms within hearing devices. The introduced strategy combines and equally weights the model-based measures for the optimization. Realistic listening scenarios are specifically designed for the evaluation of the optimization approach with human test subjects. The study shows a mismatch between the predicted optimized parameter setting and the perceptual results and that the scenarios prove to be beneficial for a valid data acquisition. Finally combining instrumental and perceptual assessments, the fourth study merges the developed automated optimization approach and the modified subjective rating task and introduces the concept of an interactive trait-based fitting procedure: A manageable amount of reasonable alternative settings reflecting individual weightings of prototypical subjective traits is presented to the user, who subjectively decides about the preferred individual setting for the respective hearing device.

Next to the knowledge gained by the conducted studies, the main outcome of this thesis is the last-mentioned combination of perceptual assessment methods with instrumental prediction methods for the prescription of a 'path of probation' and the concurrent estimation of individual weightings of subjective traits. Further studies using this framework will be necessary to show the value of the introduced concept in terms of an optimized hearing device fitting procedure.

Zusammenfassung

Die vorliegende Arbeit liefert einen Beitrag zu dem langfristigen Ziel, Anpassungs- und Evaluationsmethoden für moderne Hörgeräte und deren Nutzer mit ihren individuellen Bedürfnissen und Wünschen bereitzustellen. Da eine effiziente, umfassende Anpassungsprozedur jede Einstellungsmöglichkeit des Hörsystems sowohl mithilfe technischer als auch perzeptiver Messungen untersuchen sollte, werden innerhalb dieser Arbeit Methoden für diese beiden Bereiche entwickelt und angewendet. Innerhalb der ersten Studie werden verschiedene binaurale Hörgerätealgorithmen zur Störgeräuschunterdrückung miteinander verglichen. Die Studie zeigt, dass einzelne computer-basierte auditorische Modelle nicht zufriedenstellend in der Lage sind, den individuellen Versorgungserfolg unterschiedlicher Hörsysteme vorherzusagen. Dementsprechend ist die zweite Studie zunächst den Vorgängen zur Datensammlung von realen Testpersonen gewidmet. Dabei wurden zwei Modifikationen für den vielfach eingesetzten MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) mit dem Ziel entworfen, die Bedienbarkeit insbesondere für ältere, schwerhörende Probanden mit wenig Technikerfahrung zu erhöhen. Dabei stellt sich heraus, dass die Verwendung der Modifikation MUSHRA drag&drop für die Mehrheit der 50 getesteten Personen zu den zuverlässigsten Daten führt. In der dritten Studie wird wiederum das Ziel verfolgt, computer-basierte auditorische Modelle zu nutzen, um optimale Einstellungen für komplexe, multidimensionale Hörgerätealgorithmen zu finden. Dabei werden allerdings nicht einzelne Modelle genutzt, sondern mehrere Modelle kombiniert eingesetzt. Zudem werden für die Überprüfung des computer-basierten Optimierungsansatzes mit realen Testpersonen speziell angefertigte Hörszenarien verwendet. Die Studie zeigt eine Diskrepanz zwischen der vorhergesagten optimalen Einstellung und den Ergebnissen der subjektiven Messung. Desweiteren erweist sich die Überprüfung innerhalb dieser Szenarien als vorteilhaft für eine valide Datenerhebung von den Testpersonen. Zuletzt vereint die vierte Studie einen weiterentwickelten computer-basierten Optimierungsansatz mit dem modifizierten perzeptiven Testverfahren zu dem Konzept einer interaktiven Präferenz-basierten Anpassungs- und Feineinstellungsprozedur: Der Nutzer wählt seine bevorzugte Einstellung für das jeweilige Hörsystem aus einer handhabbaren und sinnvollen Auswahl von Alternativen selbstständig aus. Die vorgegebenen Alternativen spiegeln dabei mögliche prototypische Nutzerpräferenzen und deren individuelle Abstufungen wider.

Das Hauptergebnis der vorliegenden Arbeit ist die Verknüpfung der perzeptiven Bewertungsmethode mit der technischen Optimierungsmethode, um einen ‘Path of Probation’ innerhalb von multidimensionalen Parameterräumen von Hörgerätealgorithmen einzustellen und eine gleichzeitige Erfassung der individuellen Abstufungen

von prototypischen Nutzerpräferenzen zu ermöglichen. Dazu sind noch weitere Untersuchungen nötig, welche zeigen werden, ob sich das vorgestellte Konzept als verbesserte Anpassungsprozedur eignet.

Contents

1 General Introduction

In a fact sheet from 2015, the World Health Organization states that 360 million people worldwide (328 million adults and 32 million children) have a disabling hearing loss, which refers to a hearing loss greater than 40 dB in the better hearing ear in adults and a hearing loss greater than 30 dB in the better hearing ear in children (World Health Organization, 2015). Estimating the world human population by 7.2 billion, 5% of the people have impaired hearing. A study from 2011 regarding the German population reveals that 13.1% have a self-stated hearing loss (Hougaard and Ruf, 2011). This higher amount with respect to the world's population can be explained by the high proportion of elderly people in Germany and the fact that hearing loss is associated with aging (presbycusis). Hearing impaired listeners suffer especially in complex acoustical situations as they encounter problems understanding speech in noisy and reverberant environments (e.g., Festen and Plomp, 1990; Humes, 1991). Recently, evidence has shown that there is a correlation between hearing loss and social isolation for elderly adults (Mick et al., 2014), i.e. some affected persons rather prefer being alone than exposing themselves to these complex acoustical situations.

One major goal of modern digital hearing aids therefore is to restore the ability to communicate in order to help hearing impaired listeners to fully participate in our society. To achieve this goal, the majority of algorithms in modern digital hearing aids are in general controlled by more than one parameter influencing the signal processing. These variable parameters lead to a multitude of possibilities for adjusting the respective algorithms, which offers the opportunity to fit the algorithms to each individual patient. However, the question remains how to effectively solve the multidimensional problem of finding the optimal fitting for each patient in all kinds of different listening situations. To tackle this question, both, algorithm developers and clinicians offer expert-based adjustments of the respective algorithm parameters. In hearing aid development, optimal algorithm parameters are commonly determined and evaluated by auditory-model based instrumental performance measures as e.g., the 'Hearing-Aid Speech Quality Index' (HASQI; Kates and Arehart, 2014), the standardized 'Perceptual Evaluation of Speech Quality' (PESQ; ITU-T —

Telecommunication Standardization Sector of ITU, 2001), and the quality measure 'PEMO-Q' (Huber and Kollmeier, 2006). Also purely technical distance measures as the signal-to-noise ratio (SNR) are used, or the optimization is heuristically based on informal listening by the algorithm developers (Rohdenburg et al., 2006). In clinical practice, the crucial hearing aid fitting task, in which the multidimensional hearing aid algorithms are matched to the individual hearing aid users, is performed by means of the commercial fitting software provided by the respective hearing aid manufacturer in combination with the expertise of the audiologist. Thus, prescriptive fittings are performed, which incorporate e.g., the audiometric hearing thresholds or suprathreshold measures like loudness discomfort levels of the patient. Examples for the commonly used threshold-based prescriptive fittings are NAL-NL1 (Byrne et al., 2001) and DSL [i/o] (Cornelisse et al., 1995). A detailed overview of various fitting strategies is given by Dillon (2012).

Evidence has shown that the prescriptive initial fitting is rarely the best fit for the individual patient (Leijon et al., 1984; Keidser and Dillon, 2006). In general, prescribed gains have been found to be slightly higher than the average preferred hearing aid responses (Wong, 2011). Jenstad et al. (2003) state that a prescriptive fitting will not guarantee patient satisfaction, particularly with the many available parameters on modern hearing aids for which values are not specified by fitting formulas. For example, the parameter regarding the sensitivity of an automatic noise reduction is not individually prescribed at all (Dillon et al., 2006). Most clinicians agree that the first prescribed setting serves as a reasonable starting point for a subsequent fine-tuning of the hearing aids, which can be regarded as the second stage of the complete fitting procedure. Due to mismatches between theoretical assumptions behind the fitting procedures and the actual real world outcomes of hearing aids, this crucial and difficult fitting process altogether can be very cost-intensive and time-consuming, e.g. incorporating many repeated visits at the audiologist (e.g., Boymans and Dreschler, 2012; Abrams et al., 2011).

The evaluation procedures for assessing hearing aid performance and benefit represent a vast and almost independent realm (Kiessling et al., 2006). One major evaluation objective of interest is to assess the real-world benefit of hearing aids, which, in research, can be captured within field studies (see e.g., Bentler et al., 2008). Discrepancies between the results of different evaluation objectives were found, namely

(a) between computer-based instrumental and perceptual measures (see e.g., Luts et al., 2010),

(b) between objective human measures and their subjective ratings (see e.g., Luts et al., 2010; Marzinzik and Kollmeier, 1999; Walden et al., 2000), and

(c) between laboratory and field trial results (see e.g., Bentler, 2005).

From this, it can be concluded that results gathered by a *narrow* evaluation using e.g. only instrumental measurements can be misleading.

The present thesis contributes towards the long-term goal of providing optimum fitting methods for hearing devices that are based on the needs and preferences of the individual user. The determination of the optimal algorithm out of a set of alternatives for each individual patient and accordingly the optimal individual setting for each algorithm thereby follows a comprehensive evaluation of hearing devices, which is presented in the following. A comprehensive evaluation includes the following items:

- physical measurements with e.g. a coupler simulating the human ear,
- instrumental measurements with e.g. auditory model-based performance measures, and
- perceptual measurements with different human listeners performing
  - objective measures like speech reception thresholds (SRTs) and
  - subjective ratings of attributes like quality or listening effort,

  both within the laboratory and a field test setup.

By performing physical measurements, the physical functionality of the test objects is determined, and it can be verified if the hearing devices act as they are supposed to. By conducting instrumental measurements, the maximum performance with respect to the aimed goals of the hearing devices is evaluated under maximally controlled conditions, i.e. the evaluation results remain stable as they are not dependent on a possible fluctuating performance of test persons. By performing perceptual measurements with human listeners (under most realistic conditions if necessary), the achievements of the hearing devices in the real world are determined.

A graphical representation of a comprehensive evaluation of hearing devices is given in Figure 1.1.

By a comprehensive evaluation, a broad knowledge base is generated which might help to identify and to even explain discrepancies between the results within this evaluation framework and in the long term to determine optimal individual fittings of

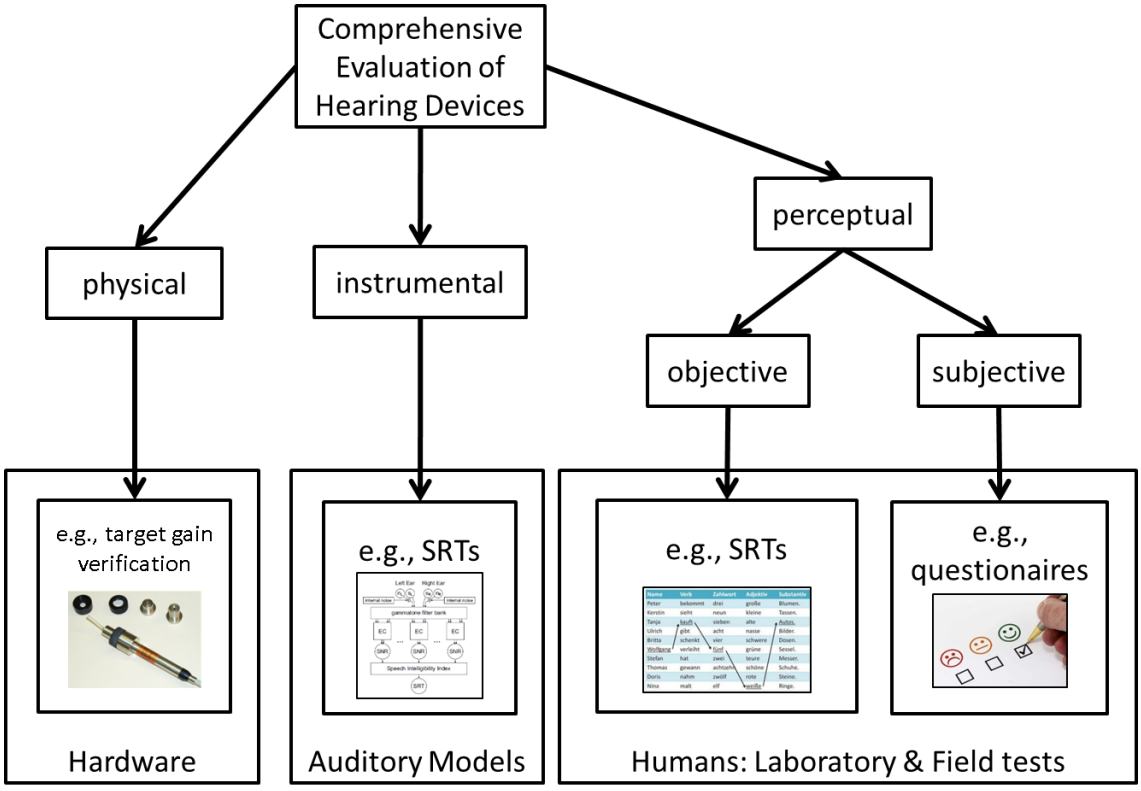

Figure 1.1: *Graphical representation of a comprehensive evaluation of hearing devices.*
Note: SRTs = speech reception thresholds.

hearing aids. Regarding the evaluation, it is furthermore essential that appropriate methods for calculating and collecting relevant data are used at the end of each branch (cf. Figure 1.1). This may refer to e.g. an applicable use of several auditory model-based performance measures or to the development of novel test methods for perceptual evaluation, which are presented within the following chapters of this thesis.

Thesis structure

Chapter 2 serves as an initial point of the thesis. Eight different hearing aid noise reduction algorithms were perceptually tested with ten normal-hearing and twelve hearing-impaired listeners. The empirical data also was compared with predictions by three common instrumental measures and individual predictions by the binaural speech intelligibility model (BSIM; Beutelmann et al., 2010). One major outcome of this study was that neither the common instrumental measures were able to predict the averaged perceptual data satisfactorily in all tested noise conditions, nor the

personalized binaural speech intelligibility model to predict the benefits obtained from the algorithms at an individual level. The discrepancies within the results from the study emphasize once more that a narrow evaluation using instrumental measures alone can be misleading and justify the claim for a comprehensive evaluation of hearing devices.

Contributing to the perceptual-subjective branch (cf. Figure 1.1), chapter 3 describes two modifications of the standardized MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA; ITU-R — Radiocommunication Sector of ITU, 2014a), which were designed to maximize the accessibility of MUSHRA for elderly and technically non-experienced listeners, who constitute the typical target group in hearing aid evaluation. The three MUSHRA variants were assessed based on subjective and objective measures, e.g. test-retest reliability, discrimination ability, time exposure, and overall preference. The results showed that both modifications can be used to obtain compatible final rating results, and both are preferred over the classical MUSHRA procedure. The comparison of the MUSHRA variants demonstrated that the intuitive modification MUSHRA drag&drop can be generally recommended.

While it had been shown in chapter 2 that the prediction power of single auditory model-based performance measures is limited, a novel technique for combining several models was examined in chapter 4. A model-based approach to generate optimized settings for multidimensional hearing aid algorithms without a direct supervision of a professional, e.g. an audiologist, was introduced and evaluated. The automated parameter optimization strategy combines different instrumental measures and extracts the optimal setting from the minimal intersection between the calculated iso-performance contours within the multidimensional parameter space of the hearing device. The approach was used to exemplary optimize a binaural coherence-based noise reduction scheme within this chapter. Together with two further algorithm adjustments by human experts, the outcomes of the different algorithms settings were compared. The comparison included instrumental and perceptual measures, i.e. SRTs and subjective attributes (listening effort, localization, naturalness, preference), which were tested in attribute-specific noise scenarios. The model-based optimization approach did not reach the perceptual performance of expert-driven settings for the chosen algorithm. Possible explanations and suggested solutions for further developments are given in the respective chapter. Testing in attribute-specific scenarios, as described in this chapter, proved to be useful to determine a possible

real-world benefit within controlled laboratory measurements.

Chapter 5 describes the fusion of further developments of the separate components from chapters 3 and 4 to introduce a concept for an interactive patient-centered fitting and fine-tuning of hearing aids. An objective method to systematically extract feasible parameter settings for complex multidimensional hearing aid algorithms is introduced. The method uses differently weighted measures to prescribe a reasonable 'path of probation' within the algorithm's parameter space to gather candidates for a subsequent subjective rating. It is assumed that these objective-measures-generated parameter sets estimate the individual weights on certain personal traits like the attitudes towards noise suppression and introduced distortions (i.e.'noise haters' and 'distortion haters'). Serving the purpose for a subjective rating, a fast and intuitive multi-stimulus test is presented combining a discrimination and classification task to capture the user preferences for the optimal algorithm setting. The use and interplay of the complementary methods was exemplary shown on a multidimensional coherence-based noise reduction algorithm. It is argued that the combined use of the two methods will help to reconcile modern multidimensional hearing aid algorithms with the individual patient and promises a higher user satisfaction by the direct patient involvement and decision-making in the fitting process.

2 Comparing binaural pre-processing strategies: Speech intelligibility of normal-hearing and hearing-impaired listeners

ABSTRACT

A comprehensive evaluation of eight signal pre-processing strategies, including directional microphones, coherence filters, single-channel noise reduction, binaural beamformers, and their combinations, was undertaken with normal-hearing (NH) and hearing-impaired (HI) listeners. Speech reception thresholds (SRTs) were measured in three noise scenarios (multitalker babble, cafeteria noise, and single competing talker). Predictions of three common instrumental measures were compared with the general perceptual benefit caused by the algorithms. The individual SRTs measured without pre-processing and individual benefits were objectively estimated using the binaural speech intelligibility model. Ten listeners with NH and 12 HI listeners participated. The participants varied in age and pure-tone threshold levels. Although HI listeners required a better signal-to-noise ratio to obtain 50% intelligibility than listeners with NH, no differences in SRT benefit from the different algorithms were found between the two groups. With the exception of single-channel noise reduction, all algorithms showed an improvement in SRT of between 2.1 dB (in cafeteria noise) and 4.8 dB (in single competing talker condition). Model predictions with binaural speech intelligibility model explained 83% of the measured variance of the individual SRTs in the no pre-processing condition. Regarding the benefit from the algorithms, the instrumental measures were not able to predict the perceptual data in all tested noise conditions. The comparable benefit observed for both groups suggests a possible

This chapter is a reformatted reprint of "Comparing Binaural Pre-processing Strategies III: Speech Intelligibility of Normal-Hearing and Hearing-Impaired Listeners", Christoph Völker, Anna Warzybok, and Stephan M. A. Ernst, Trends in Hearing (19), 1–18. The original article can be found at https://doi.org/10.1177/2331216515618609.

application of noise reduction schemes for listeners with different hearing status. Although the model can predict the individual SRTs without pre-processing, further development is necessary to predict the benefits obtained from the algorithms at an individual level.

2.1 Introduction

The World Health Organization states that 360 million people worldwide have to deal with disabling hearing loss (World Health Organization, 2015). For adults, a disabling hearing loss refers to a hearing loss greater than 40 dB HL in the better ear. People with hearing loss suffer especially in complex acoustical situations. It is well known that they encounter great difficulties understanding speech in noisy and reverberant environments (e.g., Plomp, 1986; Humes, 1991). Modern digital hearing aids offer a number of approaches designed to solve this problem. The aim of hearing aids is to improve speech intelligibility while not degrading the signal quality by applying speech enhancement techniques, noise and feedback reduction schemes, or directional microphones.

In clinical practice, the benefit from hearing aids is usually measured by comparing the intelligibility scores of single words or sentences presented in quiet with and without hearing aids. However, an increasing number of studies have measured speech intelligibility in noise, which is closer to the acoustical environment that hearing- impaired (HI) listeners are faced with in their daily life. As there is no single unified way of assessing the benefit from hearing aids in noise, some studies are restricted to the relatively simple condition involving speech recognition in stationary noise (e.g., Peeters et al., 2009). Other studies have investigated the improvements caused by hearing aids within different more complex noise types, including babble or cafeteria noise (e.g., Cornelis et al., 2012; Healy et al., 2013; Luts et al., 2010). Still, several studies have indicated that the benefit from hearing aids measured in such controlled acoustical conditions did not match the benefit reported by the users in the everyday listening conditions (Bentler et al., 1993a,b; Cord et al., 2004).

In our view, a comprehensive evaluation incorporates (a) instrumental model-based measures determining the effectiveness of signal processing schemes and (b) subjective experiments accessing the efficacy in realistic test environments. Accordingly, the current article, being part of a collaborative research project to comprehensively evaluate state-of-the-art binaural signal pre-processing schemes, is aimed at further closing the gap between real-life performance and laboratory measures.

Eight advanced signal pre-processing strategies, including directional microphones, a coherence filter, single-channel noise reduction (SCNR), and binaural beamformers, as well as a no pre-processing (NoPre) condition serving as a reference, were implemented and subsequently evaluated. Three noise conditions, based on typical everyday listening situations, were designed to measure their potential benefit. These conditions involved speech recognition in multitalker babble noise, cafeteria ambient noise (CAN), and noise from an intelligible competing talker spatially separated from the target speaker.

This study tested the effects of the different strategies in normal-hearing (NH) and HI listeners. The results were compared with the outcomes from instrumental measures (Baumgärtel et al., 2015b) and results from bilateral cochlear implant (CI) users (Baumgärtel et al., 2015a).

2.2 Method

2.2.1 Listeners

Ten NH listeners and 12 HI listeners participated in this study. The group of NH listeners consisted mostly of students and employees of our department with self-reported NH. Four female and six male listeners participated, ranging in age from 21 to 33 years, with an average age of 27.3 years.

The tested HI group included 8 male and 4 female listeners, ranging in age from 21 to 66 years (with a mean of 44.3 years). Table 2.1 summarizes the information about the HI participants. The listeners are rank-ordered according to age, followed by their averaged thresholds over the frequencies 500, 1000, 2000, and 4000 Hz (4PTA) in their better ear. The first column ID is used to index the nth listener. The 4PTA for the better ear across HI listeners ranged from 20 dB HL to 55 dB HL, with a mean of 43.9 dB HL.

Figure 2.1 shows the audiometric data of the 12 HI listeners (left panel: right ear, right panel: left ear). The mean thresholds with corresponding standard deviation are displayed as thick lines; individual thresholds are shown as thinner solid lines. The HI listeners showed slight-to-moderate bilateral sensorineural hearing impairments. For all but one listener, the across-ear asymmetry in hearing thresholds was within 25 dB for any audiometric frequency between 125 Hz and 8 kHz. The participant with asymmetric hearing loss (ID 1, see Table 2.1) showed a mean asymmetry across ears of 17 dB, averaged across all audiometric frequencies. For the other listeners, the averaged asymmetry did not exceed 11 dB.

Table 2.1: *ID, Gender (M/F), Age and Hearing Thresholds of the HI Listeners.*

			Audiometric thresholds (dB HL)					
ID	M/F	Age	250	500	1k	2k	4k	8k
1	M	21	5	0	25	45	60	65
			10	30	50	65	60	70
2	M	28	15	25	45	50	50	50
			20	30	40	50	45	50
3	M	28	5	25	75	65	55	25
			5	20	75	70	60	75
4	M	34	15	25	50	55	65	60
			20	30	45	50	60	65
5	F	40	35	45	60	60	45	70
			45	50	65	60	50	80
6	F	40	15	45	50	55	60	75
			25	55	60	60	55	80
7	M	43	15	25	35	65	70	70
			15	20	35	55	60	70
8	F	43	30	40	50	55	55	60
			30	40	50	55	60	65
9	M	61	5	5	10	10	55	60
			5	10	15	30	55	65
10	F	61	30	45	45	45	65	85
			40	50	50	50	75	100
11	M	66	15	30	35	35	55	60
			15	20	20	35	50	80
12	M	66	40	45	45	50	75	105
			40	45	45	50	70	80

Note. The values in the upper lines for each listener specify thresholds of the left ear, and the values in the lower lines specify the right ear.
HI = hearing-impaired; HL = hearing loss.

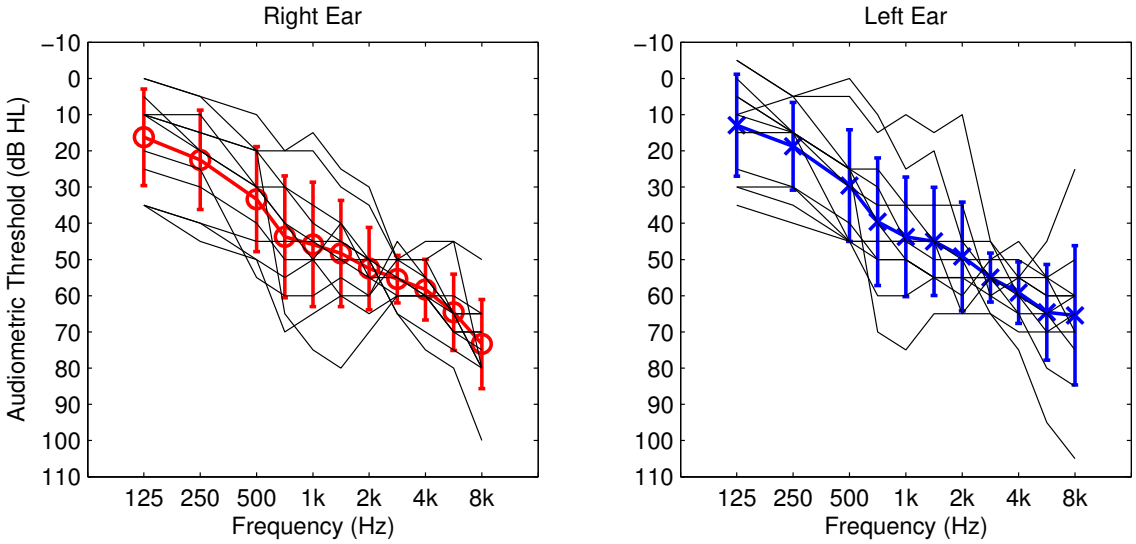

Figure 2.1: *Audiometric data of the 12 hearing-impaired listeners (left panel: right ear; right panel: left ear). The mean thresholds (marked for the right ear with circles and with crosses for the left ear) ± standard deviation are displayed thick, individual thresholds are printed in thinner solid lines.*

Except one listener (ID 9), all participants wore bilateral hearing aids on a daily basis. All but one participant with a longer distance to travel performed the tests on two different days. On the first test day, pure-tone thresholds and the speech reception thresholds (SRTs) for one noise condition were measured (2.5 hr). On the second day, the SRTs were measured for the other two noise conditions (2.5 hr). Sufficient breaks were provided between the measurements. The HI participants were remunerated for their participation and their travel expenses.

Approval for this experiment was obtained from the Carl von Ossietzky Universität Oldenburg ethics committee.

2.2.2 Signal Processing Strategies

Eight signal pre-processing strategies were evaluated in this study. The algorithm conditions and corresponding references are listed in Table 2.2. The processing schemes were described in detail and evaluated instrumentally by Baumgärtel et al. (2015b). The adaptive differential microphone (ADM) scheme and the SCNR do not provide a binaural link between the input channels and can be regarded as being monaural. The remaining six schemes are binaural pre-processing strategies, consisting of the ADM in combination with a coherence-based noise reduction (ADM + coh), a fixed and an adaptive version of a binaural beamformer using minimum variance distortionless response (MVDR) technique, and three variants of postfilters based on the binaural MVDR beamformer technique. Together with

Table 2.2: *List of Signal Pre-processing Strategies.*

#	Abbreviation	Algorithm
1	NoPre	no pre-processing
2	ADM	adaptive differential microphones (Elko and Nguyen Pong, 1995)
3	ADM + coh	adaptive differential microphones in combination with coherence-based noise reduction (e.g., Grimm et al., 2009; Luts et al., 2010)
4	SCNR	single channel noise reduction (e.g., Breithaupt et al., 2008; Gerkmann et al., 2008)
5	fixed MVDR	fixed binaural MVDR beamformer (e.g., Van Veen and Buckley, 1988; Doclo et al., 2010)
6	adapt MVDR	adaptive binaural MVDR beamformer (e.g., Griffiths and Jim, 1982)
7	com PF (fixed MVDR)	common postfilter based on fixed binaural MVDR beamformer (e.g. Simmer et al., 2001)
8	com PF (adapt MVDR)	common postfilter based on adaptive binaural MVDR beamformer (e.g. Simmer et al., 2001)
9	ind PF (adapt MVDR)	individual postfilter based on adaptive binaural MVDR beamformer (e.g. Simmer et al., 2001)

Note. MVDR = minimum variance distortionless response.

the NoPre condition serving as a reference, in total nine algorithm conditions were provided in the test protocol.

2.2.3 Compensation of Hearing Loss

To compensate for the hearing loss of HI participants, a multiband compressor scheme was used. The compressor divides the left and right input signals into nine overlapping filter bands and measures the sound pressure levels (SPLs) in each band. Dependent on the actual input band levels and the individual hearing threshold levels of the HI listener, insertion gains are calculated using the nonlinear fitting procedure (NL1) by the National Acoustic Laboratories (NAL; Byrne et al., 2001). The NAL-NL1 prescription rule aims at maximizing speech intelligibility while constraining loudness to be normal or less. The obtained first fit of the procedure was used. The insertion gains are applied to the input signals by the multiband compressor. The time constants for attack and release were 20 ms and 100 ms, respectively.

When testing NH participants, the multiband compressor did not apply any gain. The dynamic compressor was — together with the evaluated noise reduction schemes — realized as part of the real-time capable Master Hearing Aid (MHA; Grimm et al., 2006).

2.2.4 Speech Reception Thresholds

In this study, the comparison of the binaural signal preprocessing schemes was done using adaptive speech intelligibility measurements in three noise conditions. The adaptive SRT measurement converged on the 50% intelligibility level. As speech material, the Oldenburg sentence test (OLSA; Wagener et al., 1999c,a,b) was used, consisting of five-word semantically unpredictable sentences with the fixed grammatical structure, that is, *Name Verb Numeral Adjective and Noun*. Ten alternatives exist for each word category. One hundred twenty sentences are combined in 45 test lists, each containing 20 sentences. The test lists have a phonemic distribution similar to the German language.

The speech signals for actual measurements were generated by convolving the *dry* one-channel OLSA sentences with head-related impulse responses (HRIRs; Kayser et al., 2009). Here, the recorded multichannel impulse responses from behind-the-ear hearing aids with front and back microphone were used (four-channel). The HRIRs were recorded in a cafeteria, that is, the test subject is virtually seated at a table in a cafeteria, listening to the OLSA speaker directing toward him (or her) from 0° and a distance of 102 cm. The layout of the cafeteria is given in Figure 5 by Baumgärtel et al. (2015b), and the target OLSA speaker is placed at position A.

The evaluation of the algorithms was performed in (a) multitalker babble noise (20-talker babble [20T]), (b) cafeteria ambient noise (CAN), and (c) a single competing talker (SCT) located at an azimuth of 90°. The scenarios differ mainly in their spectro-temporal structure. The speech-shaped multitalker babble noise is stationary, the cafeteria noise has a typical quasistationary modulation of a cafeteria ambiance including noise of dishes and cutlery and snippets of conversations, and the SCT is speech-modulated. Each noise scenario has a duration of 600 s, which is a sufficient length for evaluating one pre-processing scheme with one test list containing 20 sentences. As the same signal material was used for the instrumental evaluation of the pre-processing schemes, further detailed information can be found in Baumgärtel et al. (2015b).

All three noise scenarios, including the unidirectional SCT condition, were scaled to a digital long-term root mean square level of -35 dB full scale averaged over all four channels. The dry (one-channel) OLSA sentences intentionally fluctuate in

level around a nominal value to provide the same intelligibility for each sentence. The convolved OLSA sentences (four-channel) were scaled to a new nominal level of -35 dB full scale to match the noise scenarios and keep the internal level fluctuations intact. The sampling rate of sentence and noise files was 44.1 kHz. These four-channel signals mixed together adaptively at different signal-to-noise ratios (SNRs) formed the input for the processing by the algorithms. The translation from digital full scale levels to SPLs was based on realistic pressure levels measured in a cafeteria ambiance by Kayser et al. (2009). The overall presentation level of the noise signals was 73.4 dB SPL averaged over both output channels.

The measurements of the SRTs were performed following the adaptive procedure by Brand and Kollmeier (2002). For the first stimuli presentation, speech and noise were mixed together at an SNR of 0 dB. Depending on the number of correctly understood words, the speech level varies for the next presentation. The noise level is held constant during the measurement procedure, and the change of presentation level for the subsequent sentence follows

$$\Delta L = -\frac{f(i) \cdot (\text{previous} - \text{target})}{\text{slope}}, \tag{2.1}$$

where the parameter *target* denotes the aimed value for intelligibility at which the procedure should converge. As we aimed at determining the SNR corresponding to 50% intelligibility, the parameter *target* is set to 0.5. Parameter *previous* denotes a value for the intelligibility of the previous sentence. The parameters *slope* and *f(i)*, which controls the rate of convergence, are set to standard values following the recommendation by Brand and Kollmeier (2002) to obtain a reliable bias-free SRT estimate. As all participants were able to use the test software and hardware by themselves, they performed the sentence tests in a closed-set response format, that is, during the test, the participants were able to see all 10 possible words for each of the five-word categories on a computer screen and were asked to select the understood words out of this closed 10 x 5 matrix.

The processing schemes were evaluated successively in each noise condition. The order of the three noise conditions was pseudorandomized for each listener, having a balanced distribution over participants. The order of the eight processing schemes and the NoPre condition tested in a given noise condition was also randomized. Each processing scheme was evaluated with a random test list containing 20 sentences. Before the first SRT measurement on a test session, the participants were instructed and performed two random test lists without pre-processing in the upcoming noise condition for training. The sentence test was implemented inside the AFC Toolbox by Ewert (2013) for MatLab. The necessary real-time processing was provided by

the software SoundMexPro. The evaluated pre-processing schemes were realized inside the MHA developed by Grimm et al. (2006). This test setup runs on an Acer tablet Iconia W700 with external soundcard Maya 44 USB by ESI Audio.

To produce a realistic test scenario, the participants were equipped with behind-the-ear hearing aid dummies Acuris P by Siemens. Here, the hearing aid microphones were turned off, and the dummies only served as headphones. The two-channel output of the hearing aids was passed into the participants' ears by using ear plugs E-A-RTONE 13A. The frequency responses of the hearing aid speakers were equalized in a calibration procedure with a 2cc coupler (IEC 126). After equalization, correct output SPLs were ensured using the MHA output calibration routine with broadband noise. The experiments were performed in a soundproofed test room (ANSI/ASA S3.1-1999, 2008).

2.2.5 Instrumental evaluation

To compare the perceptually measured SRT benefits with the instrumental evaluation of the schemes (see Baumgärtel et al., 2015b), we calculated the Kendall rank correlation coefficient τ between the subject data and the instrumental measures. These instrumental measures consist of (a) intelligibility-weighted SNR (iSNR), (b) short-time objective intelligibility (STOI), and (c) perceptual evaluation of speech quality (PESQ). For the instrumental evaluation, 120 OLSA sentences were mixed with the three noise scenarios 20T, CAN, and SCT at different long-term SNRs corresponding approximations of the averaged measured SRTs within the NoPre condition (*baselines*) in each scenario. These mixtures were processed by the schemes and evaluated by the three instrumental measures. The mixtures without any applied pre-processing strategies were also evaluated by the measures. The averaged better channel improvements regarding NoPre were used to determine the capability of each of these measures to predict the SRT benefits for NH and HI listeners.

2.2.6 Binaural Speech Intelligibility Model

In addition to the instrumental measures described in Baumgärtel et al. (2015b), we evaluated the algorithms in this study by means of the binaural speech intelligibility model (BSIM; Beutelmann et al., 2010). In contrast to the instrumental measures, the BSIM was used to predict the individual SRTs of the NH and the HI listeners.

The first stage of the model applies a gammatone filter bank to analyze the binaural speech and noise signals. Internal noise is derived from the individual audiogram and is added to the external noise to account for hearing impairment. In each frequency

band, an independent equalization-cancellation (EC) mechanism (Durlach, 1963) is applied to compute the maximally achievable SNR. It is achieved by eliminating the noise signal due the destructive interference by subtracting the two channels from each other. This maximal SNR is adapted to imperfect human binaural processing by applying binaural *processing errors*, which restrict the performance of the EC process by preventing the perfect cancellation of the noise signal. The speech and noise signals are processed separately for a reliable estimate of SNR. The resulting SNRs are then used as input for the speech intelligibility index (SII; ANSI 20S3.5-1997, 1997), which varies between 0 (completely unintelligible) and 1 (perfect intelligibility). The resulting SII is transformed into an intelligibility value using a nonlinear transform derived from a mapping function for sentence intelligibility (cf. Table III, Fig. 7, Fletcher and Galt, 1950). The SRT for a given condition is calculated by selecting a fixed reference SII value and varying the SNR until the SII equals this reference value.

BSIM was used to predict the SRTs in all noise conditions with signals for NH and HI listeners without any pre-processing. Because the effectiveness of the algorithms may depend on the input SNR, the processing schemes were evaluated in terms of SII for different input SNRs, ranging from -17.5 dB to 0 dB with a step size of 2.5 dB. To validate the quality of the SII benefit predictions on an individual level, the measured SRT benefits are compared with the predicted SII benefits for each HI listener separately. In line with the evaluation by the other instrumental measures, the empirical data and individual model predictions were compared by means of the Kendall rank correlation coefficient τ.

The speech signal consisted of 20 concatenated OLSA sentences convolved with the same HRIRs as the signals in human experiments. For the model predictions, the signals from the front microphones were taken. The length of the speech stimuli corresponded to the ones that were presented to the listeners. For each noise condition, the extended version of the model was applied (short-time BSIM; Beutelmann et al., 2010). This model version was proposed for speech intelligibility predictions in modulated or time variant interferers. The short-time BSIM analyses the signals in short time frames of 1024 samples at 44100 Hz sampling rate and a frame shift of half the frame length. The effective frame length is about 12 ms. These parameters were set based on the previous findings of Rhebergen et al. (2006) and Beutelmann et al. (2010).

The measured SRTs of listeners with NH in the cafeteria noise and signals without pre-processing were used as the reference condition to normalize the SII. The reference SII was set to 0.09. This ensured that the predicted SRT was very close to the mean

SRT measured in the reference condition. The reference SII was kept constant for other conditions. For speech intelligibility predictions of HI listeners, the individual audiograms were used to simulate hearing impairment. For NH listeners, the use of individual audiograms does not significantly influence the accuracy of speech intelligibility predictions in noise, and the variance observed in measured SRTs cannot be predicted by the model (Beutelmann et al., 2010). Therefore, for predictions of NH listeners, an average audiogram of 0 dB HL is assumed at all frequencies.

Because BSIM requires separate speech and noise signals at the input, the phase-inversion method (Hagerman and Olofsson, 2004) was used for separation of the processed signals.

2.2.7 Statistical analysis

Two separate analyses of variance (ANOVAs) were used to investigate the significance of within- (algorithm, noise condition) and between-subject factors (listener group: NH, HI) and their interactions on listeners performance. The first ANOVA analyzed the measured SRTs without pre-processing (baselines). For the second ANOVA, we calculated individual SRT differences (benefits) for each scheme and each noise condition with respect to the NoPre condition. Using the statistical software IBM SPSS, the data were fed into mixed-model ANOVAs with repeated measurements. Whenever necessary, violations of sphericity were adjusted using the Greenhouse–Geisser correction. To determine the sources of significant effects indicated by the ANOVA, post hoc tests for multiple comparisons were conducted and reported with adjusted criteria for significance by the amount of comparisons.

2.3 Results

2.3.1 Speech Reception Thresholds

The measured SRTs without pre-processing are shown in Figure 2.2. All data values are indicated separately for each of the three noise conditions and the two listener groups, that is, NH and HI listeners. The data from the HI listeners are also shown by index numbers, which are aligned with the information about the hearing-impaired participants given in Table 2.1.

The ANOVA of the data supports that both within-subjects factor noise condition and listener group as between-subjects factor are statistically significant [noise condition: $F(2,40) = 438.3$, $p < 0.001$; listener group: $F(1,20) = 411.7$, $p < 0.001$]. The interaction of both factors is also significant [$F(2,40) = 12.7$, $p < 0.001$]. Post

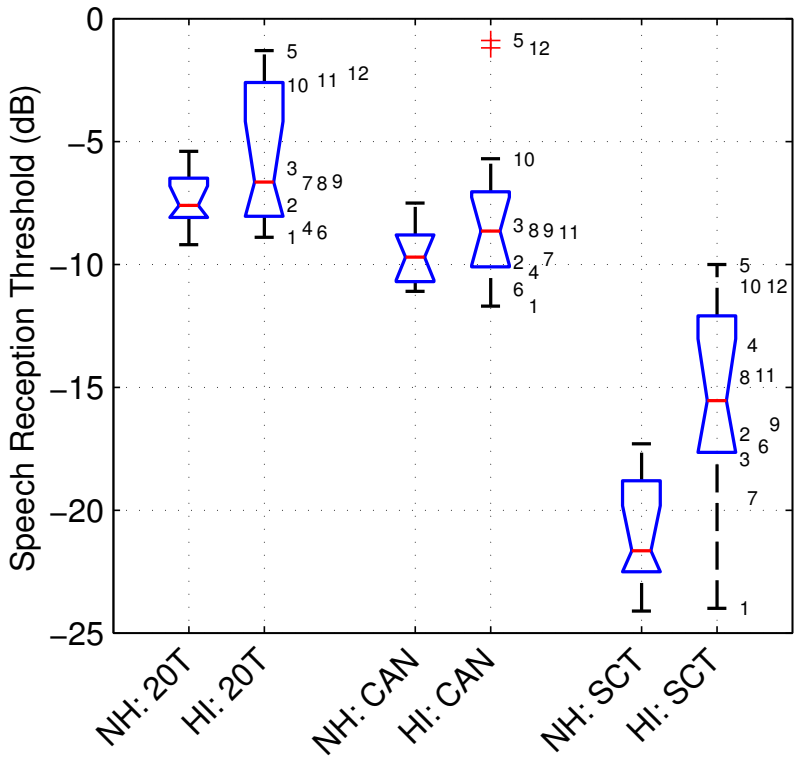

Figure 2.2: *Distribution of measured speech reception thresholds for the no pre-processing condition, pairwise for both listener groups (left: normal hearing, right: impaired hearing) and separated for noise conditions (from left to right: 20 talker babble (20T), cafeteria ambient noise (CAN), single competing talker (SCT)). The boxes have lines at the lower quartile, the median value, and the upper quartile. Whiskers extend from each end of the box to the adjacent values in the data. Outliers (+ sign) are data with values beyond the ends of the whiskers. Individual data from the HI listeners are shown by index numbers (cf. Table 2.1).*

hoc analysis of the interaction shows that all multiple comparisons between the noise conditions differ significantly (all $p_{Bonf} < 0.05$) for both groups, with normal and impaired hearing. Concerning the significant factor noise condition ($p < 0.001$), the multitalker babble noise (20T) exhibits the highest thresholds, followed by the cafeteria ambient noise (CAN) and the single competing talker (SCT). Comparing the thresholds by the subject groups, mean thresholds for NH are lower in all noise conditions than for HI listeners. In the 20T condition, the thresholds for NH are on average 1.8 dB lower than for HI (mean SRTs of NH and HI group are -7.5 dB and -5.7 dB, respectively). An equal difference of 1.8 dB is observed in cafeteria ambient noise (NH: -9.7 dB, HI: -7.9 dB). In the SCP condition, the thresholds of NH listeners are 5.5 dB lower in respect to the HI listeners (NH: -21.0 dB, HI: -15.5 dB). Furthermore, in all noise conditions the variance between the HI listeners is considerable higher than in the group of NH listeners.

The individual thresholds in the NoPre condition, that is, without any active noise reduction algorithm, form the baselines for the further analysis of the signal processing schemes. For each scheme the SRT difference with respect to the individual baseline is calculated (SRT benefit), that is, the benefit in terms of SRTs provided by each scheme is analyzed.

The mean individual SRT benefit averaged across NH participants is shown in Figure 2.3 for each of the eight processing schemes. The corresponding data for HI listeners are plotted in Figure 2.4. The error bars indicate intervals of ± standard deviation from the mean value. The results are presented separately for each of the three noise conditions (20T, CAN, and SCT). Also, the averaged SRT values with corresponding standard deviation for the NoPre condition are shown for each noise condition in the figures legends.

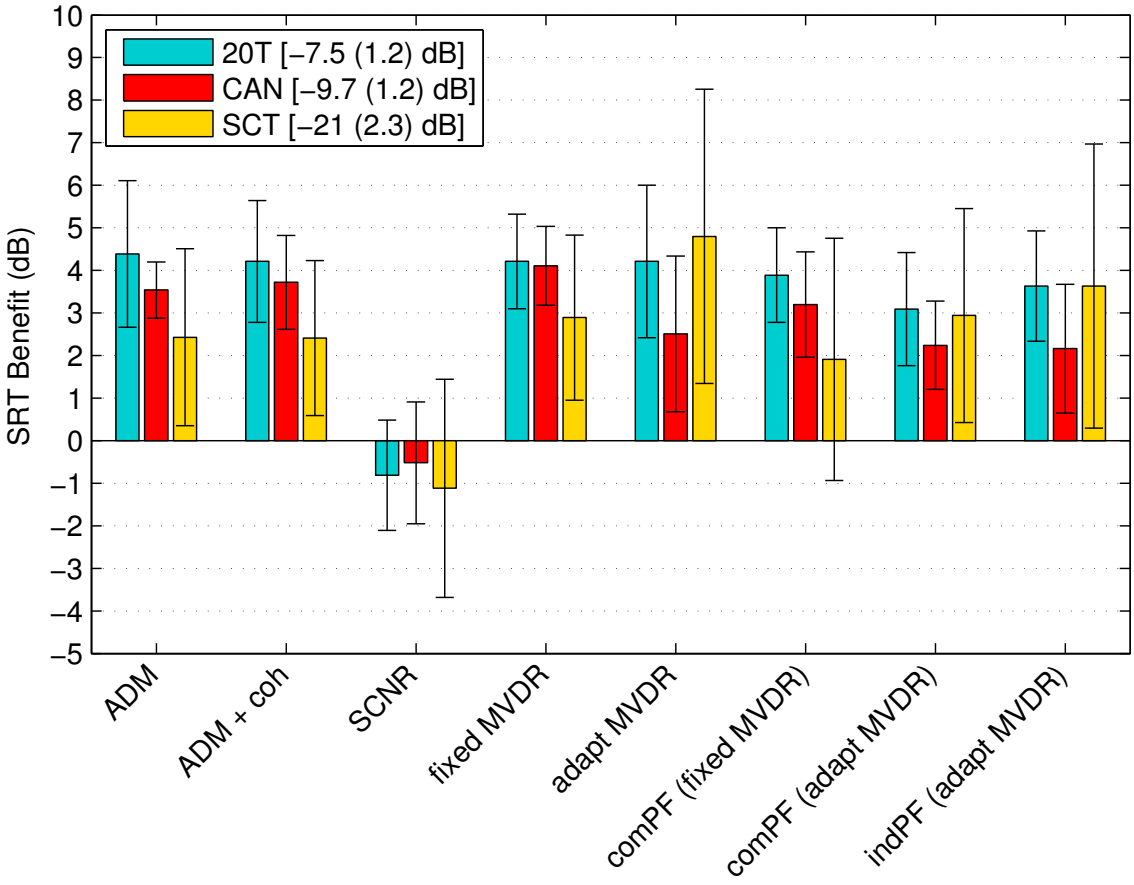

Figure 2.3: *Averaged individual SRT benefit caused by eight signal processing schemes for ten normal hearing participants. The error bars denote an interval of ± standard deviation from the mean value. The results are displayed separately for the three noise conditions 20T, CAN, and SCT. The legend shows the averaged SRT values measured without pre-processing including standard deviation in each noise condition.*

The data from NH participants (Figure 2.3) reveal that all but the SCNR scheme lead to an improvement in SRTs. Excluding the SCNR scheme, the benefits caused by the schemes in the stationary 20T condition are quite similar across the algorithms, ranging from 3.1 dB (com PF based on adapt MVDR) to 4.4 dB (ADM). In the cafeteria condition also all but the SCNR scheme cause an improvement in speech recognition. Here, the benefits range from 2.2 dB (ind PF based on adapt MVDR) to 4.1 dB (fixed MVDR). In the SCT condition, the improvements range from 1.9 dB (com PF based on fixed MVDR) to 4.8 dB (adapt MVDR), again with the exception of the SCNR scheme. The variance of the individual improvements for each scheme are highest in the SCT condition, i.e. the mean standard deviation of SRT benefit over all schemes is 2.5 dB (20T: 1.4 dB; CAN: 1.2 dB).

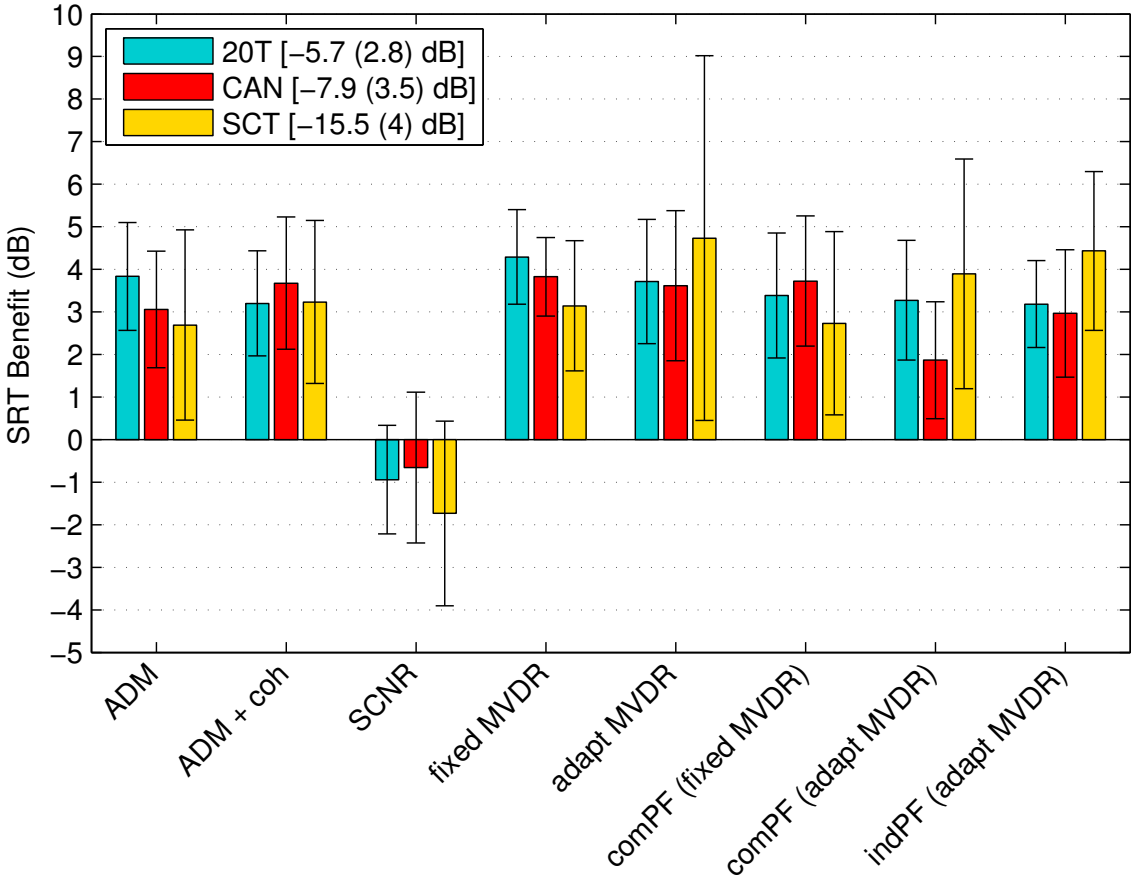

Figure 2.4: *Averaged individual SRT benefit caused by eight signal processing schemes for twelve hearing impaired participants. The error bars denote an interval of ± standard deviation from the mean value. The results are displayed separately for the three noise conditions 20T, CAN, and SCT. The legend shows the averaged SRT values measured without pre-processing including standard deviation in each noise condition.*

The results from HI listeners (see Figure 2.4) show similar trends as the data obtained from NH listeners. All but the SCNR scheme lead to improvements in

speech recognition for all noise conditions. In 20T, the benefit ranges from 3.2 dB (ind PF based on adapt MVDR) to 4.3 dB (fixed MVDR). In CAN, the smallest improvement of 1.7 dB is observed for the com PF based on adapt MVDR scheme; the highest improvement of 3.6 dB is obtained for fixed MVDR. In SCT, the benefits range from 2.7 dB (ADM) to 4.7 dB (adapt MVDR). Again, the SCT condition shows the highest variance in SRT benefit for each scheme (mean standard deviations of SRT benefit are 20T: 1.2 dB; CAN: 1.3 dB; SCT: 2.2 dB).

The between-subjects factor of listening group shows no significant effect on the measured SRT benefits [$F(1,20) = 0.024$, $p > 0.8$]. This indicates that NH and HI listeners benefit equally from the different processing schemes. The within-subjects factor algorithm has a significant effect on the SRT benefit [$F(8,160) = 96.5$, $p < 0.001$]. The factor noise condition alone shows no effect [$F(2,40) = 1.6$, $p > 0.2$]. The interaction of factors algorithm and noise condition is significant [$F(4.4, 87.4) = 4.3$, $p = 0.002$]. We investigated this significant interaction further in a post hoc analysis by means of multiple comparisons of the algorithms in each noise condition separately.

The pairwise comparisons of pre-processing schemes regarding speech intelligibility benefit averaged over NH and HI listeners separated by the three noise conditions are displayed in Tables 2.3 (20T), 2.4 (CAN), and 2.5 (SCT): The numbers denote differences in dB for the measured SRTs averaged across listeners (schemes in rows minus schemes in columns), that is, positive values correspond to a better performance of the scheme in the respective row. Significant differences (criteria adjusted by the amount of comparisons) are marked with * ($p < 0.0014$), ** ($p < 0.00028$), and *** ($p < 0.00003$).

Table 2.3: *Pairwise comparison of processing schemes regarding speech intelligibility benefit averaged over normal-hearing and hearing-impaired listeners within noise condition 20T. The numbers denote differences in dB for the averaged measured speech reception thresholds (schemes in rows minus schemes in columns), i.e. positive values correspond to a better performance of the scheme in the respective row. Significant differences (criteria adjusted by the amount of comparisons) are marked with * ($p < 0.0014$), ** ($p < 0.00028$), and *** ($p < 0.00003$).*

scheme in 20T	NoPre	ADM	ADM + coh	SCNR	fixed MVDR	adapt MVDR	com PF (fixed MVDR)	com PF (adapt MVDR)	ind PF (adapt MVDR)
NoPre	—								
ADM	4.1***	—							
ADM + coh	3.7***	-0.4	—						
SCNR	-0.9	-5.0***	-4.6***	—					
fixed MVDR	4.3***	0.1	0.5	5.1***	—				
adapt MVDR	4.0***	-0.1	0.3	4.8***	-0.3	—			
com PF (fixed MVDR)	3.6***	-0.5	-0.1	4.5***	-0.6	-0.3	—		
com PF (adapt MVDR)	3.2***	-0.9	-0.5	4.1***	-1.1**	-0.8	-0.5	—	
ind PF (adapt MVDR)	3.4***	-0.7	-0.3	4.3***	-0.8**	-0.6	-0.2	0.2	—

Table 2.4: *Pairwise comparison of processing schemes regarding speech intelligibility benefit averaged over normal-hearing and hearing-impaired listeners within noise condition CAN. The numbers denote differences in dB for the averaged measured speech reception thresholds (schemes in rows minus schemes in columns), i.e. positive values correspond to a better performance of the scheme in the respective row. Significant differences (criteria adjusted by the amount of comparisons) are marked with * ($p < 0.0014$), ** ($p < 0.00028$), and *** ($p < 0.00003$).*

scheme in CAN	NoPre	ADM	ADM + coh	SCNR	fixed MVDR	adapt MVDR	com PF (fixed MVDR)	com PF (adapt MVDR)	ind PF (adapt MVDR)
NoPre	—								
ADM	3.3***	—							
ADM + coh	3.7***	0.4	—						
SCNR	-0.6	-3.9***	-4.3***	—					
fixed MVDR	4.0***	0.7	0.3	4.6***	—				
adapt MVDR	3.1***	-0.2	-0.6	3.7***	-0.9	—			
com PF (fixed MVDR)	3.5***	0.2	-0.2	4.1***	-0.5	0.4	—		
com PF (adapt MVDR)	2.1***	-1.2**	-1.6***	2.6***	-1.9***	-1.0	-1.4***	—	
ind PF (adapt MVDR)	2.6***	-0.7	-1.1**	3.2***	-1.4***	-0.5	-0.9	0.5	—

Table 2.5: *Pairwise comparison of processing schemes regarding speech intelligibility benefit averaged over normal-hearing and hearing-impaired listeners within noise condition SCT. The numbers denote differences in dB for the averaged measured speech reception thresholds (schemes in rows minus schemes in columns), i.e. positive values correspond to a better performance of the scheme in the respective row. Significant differences (criteria adjusted by the amount of comparisons) are marked with * ($p < 0.0014$), ** ($p < 0.00028$), and *** ($p < 0.00003$).*

scheme in SCT	NoPre	ADM	ADM + coh	SCNR	fixed MVDR	adapt MVDR	com PF (fixed MVDR)	com PF (adapt MVDR)	ind PF (adapt MVDR)
NoPre	—								
ADM	2.6**	—							
ADM + coh	2.8***	0.3	—						
SCNR	-1.4	-4.0***	-4.2***	—					
fixed MVDR	3.0***	0.5	0.2	4.4***	—				
adapt MVDR	4.8***	2.2	1.9	6.2***	1.8	—			
com PF (fixed MVDR)	2.3*	-0.2	-0.5	3.7***	-0.7	-2.4	—		
com PF (adapt MVDR)	3.4***	0.9	0.6	4.8***	0.4	-1.4	1.1	—	
ind PF (adapt MVDR)	4.0***	1.5	1.2	5.5***	1.0	-0.7	1.7	0.6	—

For the multitalker babble noise (20T), we find that with the exception of SCNR all intelligibility improvements caused by the schemes compared to NoPre are highly significant ($p < 0.00003$). The largest improvement (4.3 dB) is provoked by fixed MVDR. However, this benefit is not significantly higher than four other schemes, that is, ADM (4.1 dB), adapt MVDR (4.0 dB), ADM + coh (3.7 dB), and com PF based on fixed MVDR (3.6 dB). Fixed MVDR shows a significantly larger benefit than ind PF based on adapt MVDR ($\Delta_{\text{benefit}} = 0.8$ dB) and com PF based on adapt MVDR ($\Delta_{\text{benefit}} = 1.1$ dB), both with $p < 0.00028$. SCNR shows no significant difference to NoPre ($p > 0.05$).

Identical to what was observed in the 20T conditions, all intelligibility improvements caused by the pre-processing schemes in the cafeteria ambient noise scenario (CAN) are highly significant with $p < 0.00003$. Again, the largest benefit is provided by fixed MVDR (4.0 dB), which is not significantly different from the four schemes ADM + coh (3.7 dB), com PF based on fixed MVDR (3.5 dB), ADM (3.3 dB), and adapt MVDR (3.1 dB). Regarding assertiveness, the fixed MVDR shows a significantly larger benefit than ind PF based on adapt MVDR ($\Delta_{\text{benefit}} = 1.4$ dB) and com PF based on adapt MVDR ($\Delta_{\text{benefit}} = 1.9$ dB), both with $p < 0.00003$. In contrast to the situation with 20T, the ADM + coh scheme prevails over ind PF based on adapt MVDR ($\Delta_{\text{benefit}} = 1.1$ dB, $p < 0.00028$) and com PF based on adapt MVDR ($\Delta_{\text{benefit}} = 1.6$ dB, $p < 0.00003$). As for the noise condition 20T, the SCNR scheme does not lead to significant changes in SRT with respect to NoPre ($p > 0.05$).

In the SCT condition, all schemes but SCNR lead to significantly improved thresholds with respect to NoPre, whereas ADM brings a benefit of 2.6 dB ($p < 0.00028$) and com PF (fixed MVDR) an improvement of 2.3 dB ($p < 0.0014$). The benefits caused by the remaining schemes, ranging from 2.8 dB (ADM + coh) to 4.8 dB (adapt MVDR), are significant with $p < 0.00003$. No significant differences between the seven schemes leading to a prevailing strategy are found. Again, the SCNR scheme (-1.4 dB) shows no difference to NoPre ($p > 0.05$).

2.3.2 Instrumental evaluation

The correlation between the averaged individual SRT benefits of NH and HI listeners and the group-dependent improvements calculated within the instrumental evaluation with three different measures (Baumgärtel et al., 2015b) is displayed in Figure 2.5. The Kendall rank correlation coefficient data are shown separately for the instrumental measures, iSNR (left panel), STOI (middle panel), and PESQ (right panel). Within each panel, the data is displayed separately for each noise condition, correlation coefficients τ are given in the figure legend.

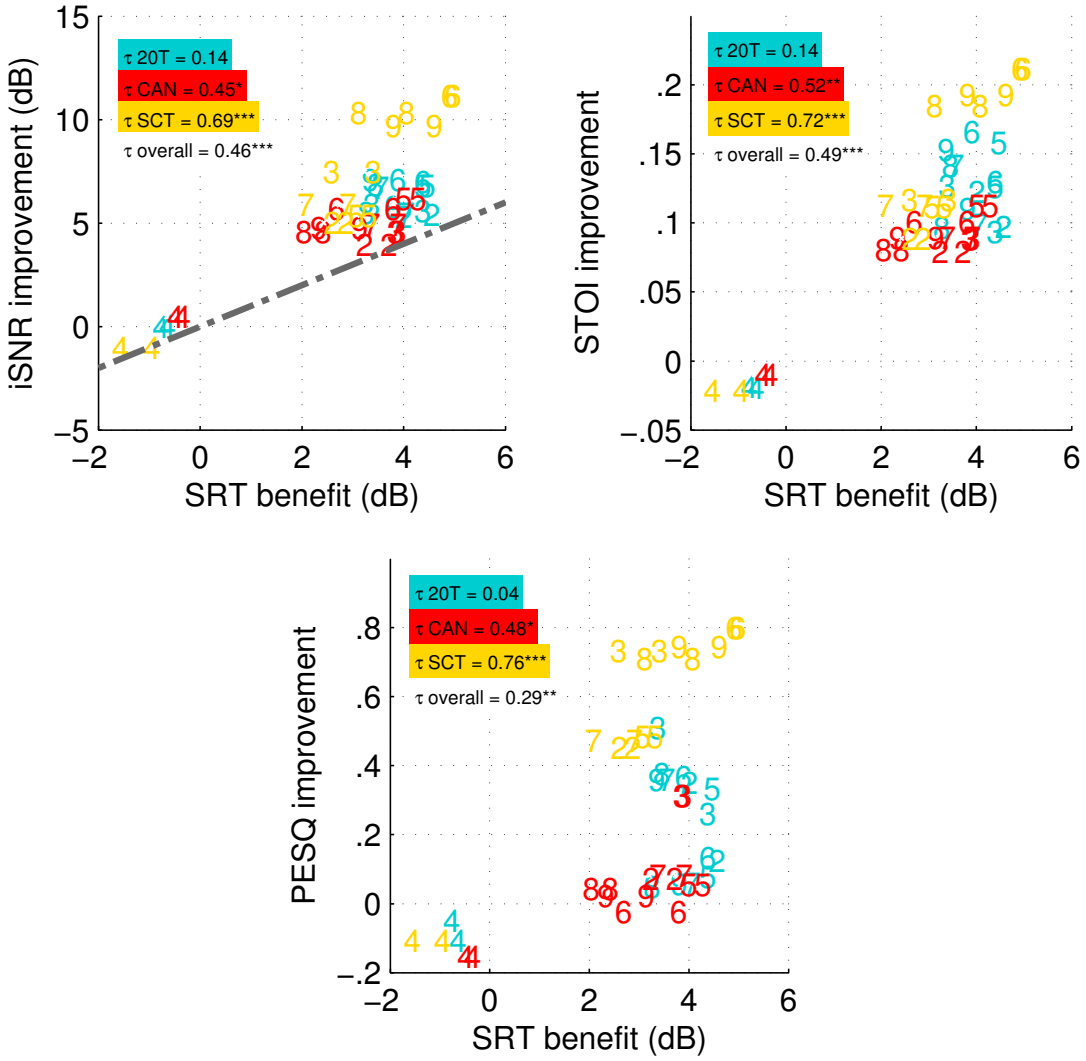

Figure 2.5: *Correlation between the averaged individual SRT benefits for normal-hearing and hearing-impaired participants and the group-dependent improvements (better channel) calculated within the instrumental evaluation with three different measures (Baumgärtel et al. 2015b). The Kendall rank correlation coefficient data is shown separately for the instrumental measures, iSNR (left panel), STOI (middle panel), and PESQ (right panel). Within each panel, the data is displayed separately for each noise condition, correlation coefficients τ are given in the figure legend. Significant correlation coefficients are marked with * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). The dash-dotted line in the left panel (iSNR) represents an idealized linear correlation between the instrumental results and subjective data.*

Averaging over all three noise scenarios, iSNR and STOI show a similar power to predict the subjective SRT benefits obtained with the tested pre-processing schemes ($\tau_{\text{iSNR}} = 0.46$, $p < 0.001$; $\tau_{\text{STOI}} = 0.49$, $p < 0.001$). Using PESQ, the overall correlation is $\tau = 0.29$ ($p < 0.01$). The predictions by all instrumental measures with respect to the different noise conditions have the highest statistical power ($p < 0.001$)

in the SCT scenario (iSNR: $\tau = 0.69$, STOI: $\tau = 0.72$, PESQ: $\tau = 0.76$). Within CAN, the best performance is reached with STOI ($\tau = 0.52$, $p < 0.01$. None of the instrumental measures is able to produce reliable predictions of the algorithm rankings in the 20T condition ($p > 0.05$).

Furthermore, the direct comparison of the subjective SRT benefit and the iSNR benefit predictions (Figure 2.5, left panel) showed a shift towards larger benefits derived from the iSNR measure, that is, all data points are above the dash-dotted line, which represents an idealized linear correlation between the instrumental results and subjective data. The median *gap* between iSNR improvements and SRT benefit, that is, the overestimation using iSNR to predict the true SRT benefit, is 2.4 dB, which is in line with the earlier findings of, for example, Van den Bogaert et al. (2009).

2.3.3 Binaural Speech Intelligibility Model

Figure 2.6 shows the overall correlation between predicted and measured SRTs of NH and HI listeners for three noise conditions and signals without pre-processing. For listeners with NH, predicted SRT is compared to the mean measured SRT averaged across listeners. For HI listeners, the comparisons are done on individual data. The coefficient of determination (R^2), the linear offset (bias) defined as the horizontal or vertical distance between the ideal mapping and the best fit with unity slope, and the root-mean-square prediction error (rms_e) were calculated to evaluate the accuracy of model predictions. The coefficient of determination correspond to fraction of the variance in the data which can be explained by the model. The predictions of BSIM show a statistically significant correlation ($p < 0.001$) to the measured data with the squared correlation coefficient of $R^2 = 0.83$, bias of 0.02, and $rms_e = 3.1$ dB.

For each noise condition and each processing scheme, the SII benefit is shown as a function of input SNR in Figure 2.7. The SII benefit is calculated as a difference in SII between the respective processing schemes and the NoPre condition. The data are averaged across the predictions for all HI listeners. In all noise conditions, the SII benefit increases with increasing SNR for all the algorithms with exception of the SCNR scheme. The SCNR scheme does not show any benefit over the whole range of input SNRs. In SCT, the SCNR scheme shows even a slight deterioration in SII compared with NoPre. Three main groups of algorithms can be distinguished in the 20T and CAN noises. To the first one with the largest SII benefit belong all beamformers with post processing, ind PF (adapt MVDR), com PF (fixed MVDR), and com PF (adapt MVDR). The beamformers without post processing (adapt MVDR and fixed MVDR) and directional microphones (ADM and ADM + coh)

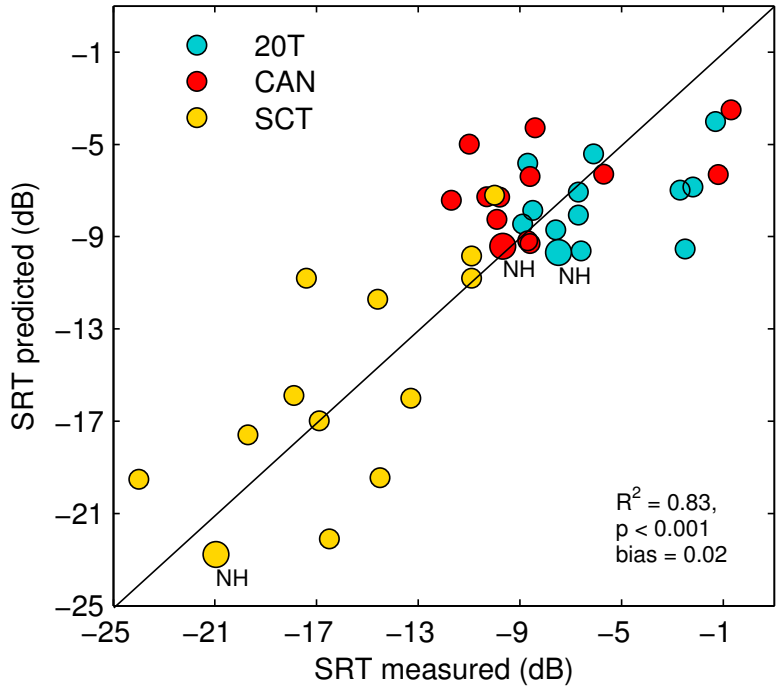

Figure 2.6: *Scatter plot of predicted SRTs against measured SRTs for normal-hearing and hearing-impaired listeners and three noise conditions (indicated by different colors). Solid line is the bisecting line. The coefficient of determination, R^2, was calculated based on individual data of HI listeners and mean data for NH listeners (marked as NH).*

create the second group with a moderate SII benefit. The smallest benefit was predicted for the SCNR scheme, which forms the third group.

In the SCT noise, the largest improvement is observed for the adapt MVDR. The SII benefit of adapt MVDR is larger in the SCT noise that in 20T and CAN noises. The lowest SII benefit is again observed for the SCNR scheme. All other processing schemes show moderate improvements in SII compared with NoPre. Similarly to other two noise conditions, the adaptive beamformers with post processing demonstrate higher benefit than directional microphones (ADM and ADM + coh) and fixed beamformer without post processing (fixed MVDR).

The analysis of model predictions on individual level is shown in Table 2.6. The Kendall's rank correlation coefficients τ are reported for each HI listener separately. Because the data shown in Figure 2.7 indicate that the predicted benefit depends on the noise condition, the accuracy of model predictions was calculated not only as average across all processing schemes and noise conditions but also for each noise condition separately. The overall correlations indicate that the model is able to predict an individual benefit from different algorithms for half of the listeners (significant

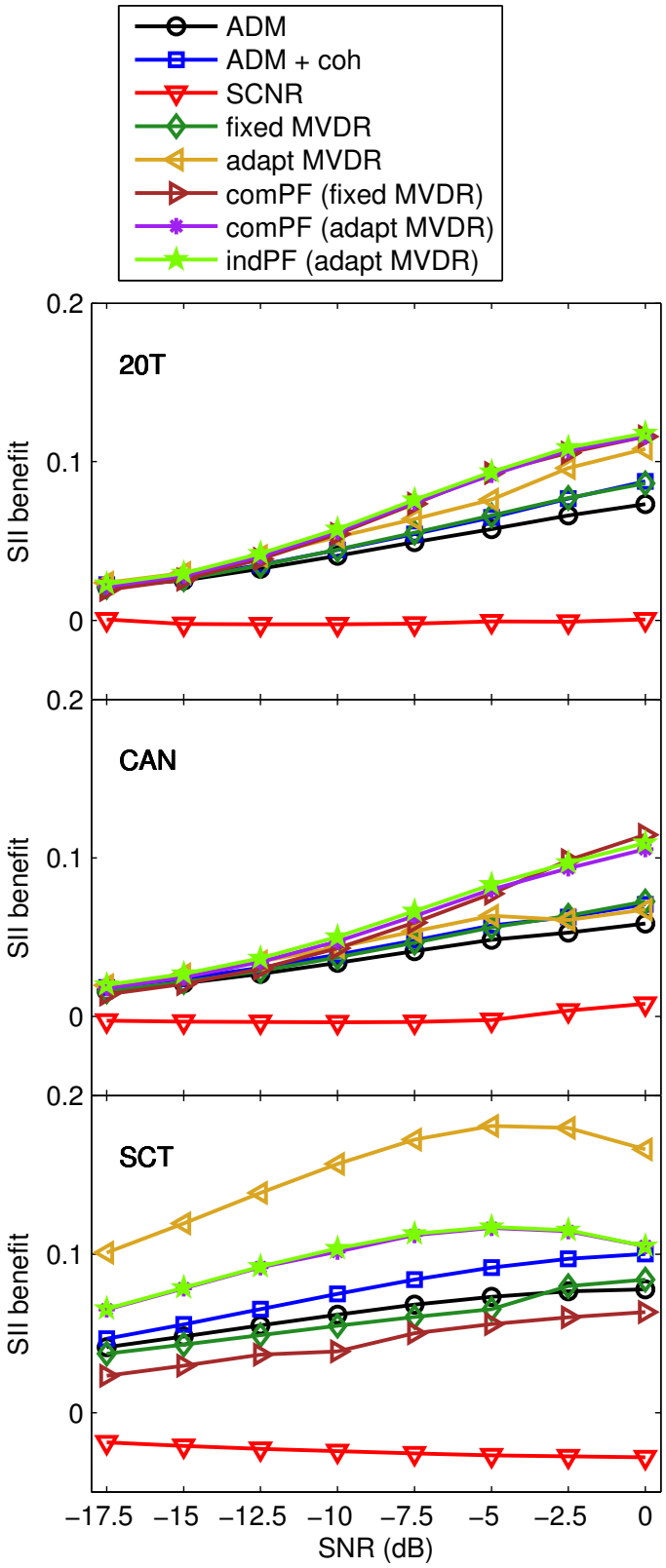

Figure 2.7: *Mean SII benefit averaged across hearing-impaired listeners for each noise condition (20T, CAN, SCT) and each algorithm as a function of input SNR.*

Table 2.6: *Rank correlation coefficients (Kendall's τ) between individual subjective SRT benefits and predicted binaural SII benefits caused by the eight processing schemes for the twelve HI listeners. Correlations are calculated separately for each noise condition, as well as the overall correlation. Significant correlation coefficients are marked with * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). Index numbers of the HI listeners (column 'ID') correspond to audiometric data from Table 2.1 and measured thresholds from Figure 2.2.*

	Noise Conditions			
ID	20T	CAN	SCT	Overall
1	0.21	0.29	0.36	0.46**
2	0.00	0.47	0.86***	0.52***
3	-0.40	-0.11	0.93***	0.31*
4	-0.14	0.07	0.91***	0.04
5	0.62*	0.40	-0.07	0.17
6	-0.07	-0.04	-0.14	-0.03
7	0.18	-0.25	0.76**	0.46**
8	0.33	-0.07	0.52	0.14
9	-0.11	0.29	0.50	0.22
10	0.55	0.33	0.22	0.39**
11	-0.07	-0.25	0.43	0.38**
12	0.29	-0.07	0.00	-0.15

correlations are indicated in Table 2.6 with stars). For the SCT noise, the trends can be very well predicted for a few listeners with τ up to 0.93. On the other side, in the same noise condition, the model fails to make reliable predictions of the individual SRT benefit for other listeners (e.g., ID 5, 6, 10, 12). Furthermore, the observed individual trends cannot be predicted in quasistationary noise conditions (20T and CAN). This suggests that the variance in the measured data for different processing schemes in each individual can not be predicted well based on the individual audiograms. The possible reasons for the discrepancies between measured and predicted benefits are discussed in the next section.

2.4 Discussion

2.4.1 Speech Reception Thresholds

Comparing the measured thresholds without any signal pre-processing scheme (baselines) in the NH and HI groups, thresholds are significantly higher ($p < 0.001$) for the HI group in every noise condition (see Figure 2.2). The HI listeners require on average 1.8 dB to 5.5 dB higher SNRs than NH listeners to achieve 50% intelligibility despite

the individual compensation of hearing loss with a dynamic multiband compressor. This is a well-known finding in hearing research and can be explained by a limited processing capacities of the impaired auditory system, which leads to a limited benefit of hearing aids (Plomp, 1978). It should be stressed that a hearing aid is a supportive device but is not able to restore all the mechanisms affected by the hearing loss.

Both, the lowest thresholds and the highest variance measured in the SCT condition, can be explained by the spectro-temporal characteristics of the noise being a single competing talker. Due to silent gaps during speech pauses, this fluctuating noise exhibits the lowest masking of the target speaker and leads to low thresholds. The silent gaps offer the possibilities for the so-called listening in the dips (Peters et al., 1998). Because the noise fragment presented with a respective sentence is not the same for each listener, the time points at which the the dips are present also differ across listeners, possibly leading to a higher variance. Furthermore, the high variance is also due to the variable individual ability of listeners to attend information in the dips.

The statistical analysis revealed that the between-subjects factor listener group does not significantly affect the measured SRT benefits ($p > 0.8$). Both listener groups, NH and HI, benefit equally from the investigated processing schemes. This finding gives more evidence for a reasonable use of noise reduction schemes also for NH persons.

The fact that between-subjects factor listener group was not significant allowed us to calculate benefits of the schemes averaged over both groups. From the pairwise comparisons of the schemes (cf. Tables 2.3–2.5), we can calculate an estimate of the assertiveness of each scheme by counting the number of significant wins over the other processing strategies. Considering the restricted capacities of a digital hearing aid, it is most likely necessary to choose one pre-processing scheme over others. A sensible selection criterion would be to choose the one processing scheme with the most significant wins measured in this study. We find the fixed beamformer (fixed MVDR) belonging to the schemes with highest assertiveness in all three noise scenarios. In the multitalker babble noise, the fixed MVDR was in the lead (significantly winning four times) and provides the highest benefit of 4.3 dB. In the cafeteria ambient noise, the fixed MVDR (4.0 dB benefit) shares the same assertiveness as the ADM +coh scheme (3.7 dB benefit). Although we do not find an assertive scheme in the single competing talker condition, the fixed beamformer (3.0 dB benefit) belongs to the seven schemes that significantly improve the intelligibility in respect to the no pre-processing condition. However, although exhibiting the highest benefit in 20T and CAN, the benefit of the fixed MVDR scheme is not significantly higher than the

benefits caused by four other schemes. Still, in terms of assertiveness over competing processing strategies generalized across the tested noise conditions, the fixed MVDR scheme emerges as being the best choice.

An alternative selection criterion for the pre-processing schemes could be to consider the implementation expense, that is, computing time and energy demand. The fixed MVDR is the best choice in terms of assertiveness. However, we found no significant difference between the binaural fixed MVDR and the monaural ADM scheme in any of the noise conditions. Thus, the ADM and the fixed MVDR provide comparable SRT benefits. In contrast to the fixed MVDR, the ADM does not require a binaural link across the ears to process the signals. Therefore, if the low costs are of great importance, the ADM can be considered as a reasonable alternative.

To estimate the implication of the SRT benefit values on the speech intelligibility in daily life, we assume that 1 dB increase in SNR produces a 20% increase in intelligibility for NH listeners and 5% for severely HI listeners (Brand and Kollmeier, 2002). Our tested group of HI showed slight-to-moderate hearing loss, that is, it is expected that 1 dB improvement in SNR will result in 5% to 20% (on average about 10%) increase in % correct scores. Thus, the measured benefit of 4.0 dB in the realistic cafeteria ambient noise would lead to an increase of about 80% in intelligibility for normal-hearing and of about 40% for slight-to-moderate hearing-impaired listeners. Relating this to our speech in noise test, these are four intelligible words out of five more for the normal-hearing and two intelligible words out of five more for the hearing-impaired persons. Hence, for both listener groups the gained 3.0 dB to 4.3 dB benefit caused by fixed MVDR promises a considerable increase in terms of everyday life quality.

Due to the high variance in the measurements, the adaptive beamformer does not differ significantly from the fixed beamformer in the SCT noise condition. However, the adaptive beamformer scheme alone (4.8 dB) and the two postfilter variations based on the adaptive beamformer technique (4.0 dB and 3.4 dB) show a trend for a larger benefit than the fixed beamformer (3.0 dB). This possible advantage of the adaptive beamformer schemes could be explained by the characteristic of the noise scenario. The competing talker coming from a single source (90°) allows the beamformer to adapt and to suppress this noise source. In the other more complex noise scenarios 20T and CAN, the adaptation procedure does not operate equally well, that is, the adaptive beamformer does not outperform the fixed beamformer. This is in line with results from other evaluation studies with adaptive beamformer schemes (e.g., Maj et al., 2006).

The SRT measurements with the adaptive beamformer scheme in the single competing talker condition exhibit a noticeable large standard deviation (cf. Figures 2.3 and 2.4). We assume that due to the interactions of the speech signal and the competing talker, the adaptation process of the adaptive beamformer is unstable. This would lead to a high variance in the subjective measurement. The system's instability has a consequences for separation of processed signals using the phase-inversion method (Hagerman and Olofsson, 2004). One of the assumptions of this method is stability of the system. Violation of this assumption can result in inaccurate separation of the speech and noise signals. The portions of the competing talker in the speech signal will lead to an overestimation of the SII benefit (see Figure 2.7).

Concerning the combination of schemes, we find the benefits caused by the additional coherence-based noise reduction serially coming after the adaptive differential microphone not significantly different from the ADM scheme alone. Still, in terms of assertiveness over competing processing schemes, the ADM + coh counts one more win than ADM in the realistic cafeteria ambient noise condition. This can be explained considering the mostly incoherent nature of the cafeteria noise. Here, the noise reduction scheme can perform the separation of the coherent speech signal and unwanted incoherent noise signals very well. In the other two tested noise conditions, the additional coherence-based noise reduction does neither show an advantage nor disadvantage with respect to the ADM scheme alone.

In our evaluation, we found that in none of the tested noise scenarios any of the combinations of the two MVDR beamformer schemes and tested type of postfilter performed better than the corresponding beamformer alone. Furthermore, in the 20T and the CAN condition the adaptive MVDR scheme alone performs significantly better without the common postfilter ($\Delta_{\text{benefit}} = 1.1$ dB in 20T, $p < 0.01$; $\Delta_{\text{benefit}} = 1.9$ dB in CAN, $p < 0.001$) or the individual postfilter ($\Delta_{\text{benefit}} = 0.8$ dB in 20T, $p < 0.01$; $\Delta_{\text{benefit}} = 1.4$ dB in CAN, $p < 0.001$). This perceptual finding is not congruent with our objective calculations of the SII benefit (cf. Figure 2.7) or the instrumental evaluation of the processing schemes by Baumgärtel et al. (2015b) and gives evidence for another mismatch between objective and perceptual evaluation. Possible explanations and solutions are discussed later.

A potential shortcoming in the measurements of the HI is the use of concatenated noise reduction and dynamic compression. When performing noise reduction before dynamic compression, the residual noise will receive more amplification compared with the speech. This might hamper the purpose of using noise reduction (Ngo et al., 2012). Although the operation point of the algorithms was different for the NH and HI listeners, no significant differences were found in the SRT benefit for both listener

groups. Therefore, for the algorithms tested, it may be assumed that as long as HI listeners with moderate hearing loss are considered, no degradation in efficiency due to the serial concatenation of noise reduction and dynamic compression was induced with respect to listeners with normal hearing in the current study.

Having in mind that both listener groups of this study (NH and HI) benefit equally from the investigated processing schemes, we can now consider the third group — the bilateral cochlear implant (CI) users. The CI data are shown in Figure 2 in Baumgärtel et al. (2015b). Comparing the datasets, the benefits for the CI users are approximately 2 dB higher than for NH or HI listeners in the multitalker babble and the cafeteria ambient noise. Also, the CI users profit up to 10 dB more from the adaptive beamformer scheme (with and without postfilter) than NH or HI listeners in the single competing talker condition.

In general, it may be concluded that the general trends observed for CI users are in principle comparable with the two investigated listener groups from this study. Therefore, it can be stated, that the benefits caused by the signal processing schemes hold for listeners with a very different hearing status (normal hearing, hearing aid users, and cochlear implant users).

2.4.2 Instrumental evaluation

The analysis considering the rank correlations between the instrumental measures presented in Baumgärtel et al. (2015b) and the subjective data from this study (cf. Figure 2.5) revealed significance in two of three noise conditions (SCT and CAN). However, the ranking correlations of the processing strategies tested here should be considered with caution. The empirical data of NH and HI listeners indicate that, for example, in the SCT noise, seven out of eight algorithms do not show significant differences in SRT benefit. Therefore, an attempt to make a reliable ranking of the processing strategies may be debatable. None of the instrumental measures described in the study of Baumgärtel et al. (2015b) can predict the trends in all noise conditions. The general discrepancy between instrumental and perceptual results, as also found in earlier studies (e.g., Luts et al., 2010), underlines the importance of subjective evaluations with normal-hearing and hearing-impaired listeners for a comprehensive hearing aid evaluation.

2.4.3 Binaural Speech Intelligibility Model

The BSIM predictions can explain 83% of variance in the measured SRTs without any pre-processing. A relatively high and significant correlation between the measured

and observed data as well as small bias indicate that the pure tone threshold is a main factor for higher SRTs observed for hearing-impaired listeners. In other words, the pure tone audiometry is an efficient measure for describing the sensitivity loss, which is the main factor influencing speech perception. The effect of noise type is well predicted for the difference between stationary and single talker noise. However, a considerable rms_e was observed between predicted and measured SRTs what indicates that despite accurate predictions of the general trends, the variance between listeners with similar audiograms cannot be predicted very well by the model. The accuracy of model predictions shown in this study is comparable with the previous finding of Beutelmann et al. (2010), who compared measured and predicted SRTs in stationary, babble, and single-talker noise conditions for normal-hearing and hearing-impaired listeners. Beutelmann and colleagues used the same model as in the current study to predict speech intelligibility and reported an overall correlation between measured and predicted SRTs of 0.78. The bias and rms_e were -3.4 dB and 3.0, respectively. A comparatively high bias in the study of Beutelmann et al. (2010) was greatest in the babble noise. The authors argued that it may be caused by the spectral difference between babble noise and speech that might be not properly handled by the SII, which frequency band importance function is linear and does not account for the correlations between adjacent or synergistic effects between the frequency bands.

The algorithms were further evaluated in terms of the speech intelligibility index (SII), which was calculated for each hearing-impaired listener and for each scheme over a broad range of SNRs. Based on that, the SII benefit was calculated corresponding to the difference in SII between signals with and without pre-processing. In agreement with the expectations, the SII benefit increased with increasing SNR. Also the differences between the algorithms were more prominent at high SNRs than at low SNRs. For the majority of the algorithms, the SII benefit was observed at the SNRs corresponding to measured SRTs. According to the model predictions, all postfilters based on binaural MVDR beamformer outperform other processing schemes in babble and cafeteria noise. These findings are not consistent with the measured data. The reason for the discrepancies might be the accuracy of the method used for separation of the processed signal into speech and noise. It is known that the SCNR schemes are able to improve the SNR but at the same time may introduce distortions to the speech or noise signal. By separating the processed signals into speech and noise components using the phase-inversion technique (Hagerman and Olofsson, 2004), the distortions will be associated with the speech component. In the following analysis by the BSIM, the whole energy in the speech component (target speaker and also

the introduced distortions) will be considered useful, leading to an overestimation of the speech recognition.

Although the individual SRTs can be well predicted in the NoPre condition (as described earlier), the model fails to make reliable estimates of the benefit from different noise reduction schemes for each individual. The analysis showed only overall (averaged across noise conditions) and in the SCT noise high and significant correlations for a subgroup of HI listeners. A number of studies give evidence for several reasons to be considered to explain the lack of the consistent correlations on individual level as well as the high observed variance. Beside the sensitivity loss, factors like age (Festen and Plomp, 1990; Dubno et al., 2002), reduced sensitivity to temporal fine structure (Lorenzi et al., 2006), narrower frequency range, in which listeners are able to use the interaural phase/time differences (Warzybok et al., 2014; Neher et al., 2011), or cognitive factors (Akeroyd, 2008) were shown to influence speech perception in noise. Thus, better individualization of model predictions for listeners with hearing impairment can be achieved when aspects other than pure sensitivity loss are accounted for. For example, Warzybok et al. (2014) showed that speech intelligibility in binaural conditions can be predicted more accurately when individual abilities in the detection of interaural phase differences are taken into account.

To overcome the shortcoming, a few solutions could be examined. The first one refers to the signal separation method. The phase-inversion method applied here could be replaced by the shadow filtering method (e.g., Fredelake et al., 2012). A second approach could consider a method suitable to estimate the distortions in the speech signal. Based on this estimate, a correction factor could be applied to the speech signal to account for the detrimental part of the speech energy.

Furthermore, another frond end could be used in the model, like an equalization-cancellation stage, that does not require separate noise and speech signals as proposed by Hauth and Brand (2015). However, replacement of the front end alone will not solve the problem discussed here since SII requires a knowledge about the SNRs in different frequency bands. After the EC processing stage, the SNRs will have to be estimated.

The most advanced solution could replace the EC stage as well as the back end in order to make the speech intelligibility predictions based on mixed signals. This could, for example, be achieved by combining the EC processing stage proposed by Hauth and Brand (2015) and the speech intelligibility model of Schädler et al. (2015). The model of Schädler et al. (2015) is based on an automatic speech recognition system and is able to make the predictions using mixed signals. In addition, in

contrast to the SII, the predictions are done without any calibration of the model to the empirical data. This combination of the models could be a future step towards predictions that will not be influenced by the accuracy of the method for separating noise and speech signals after processsing.

2.5 Conclusion

In this study, NH and HI participants tested variations of noise reduction schemes with respect to the possible benefit for speech intelligibility. This subjective evaluation was expanded by objective individual model predictions using the BSIM (Beutelmann et al., 2010). The following conclusions can be drawn:

- Forced to choose one pre-processing scheme over the others, for example, in practical applications, the fixed MVDR beamformer scheme represents the best choice in our comparison study. In all noise scenarios, it was either the best placed or equal to other best placed schemes regarding assertiveness. Depending on the noise scenario, the fixed beamformer improved SRT from 3.0 dB to 4.3 dB. However, considering implementation expense, the monaural ADM scheme offers a reasonable alternative.

- Both tested listener groups (with normal and impaired hearing) benefit equally from the investigated processing schemes. It can be stated that the benefits caused by the signal processing schemes hold for subjects with different hearing status. Thus, the possible benefit of noise reduction schemes does not only apply to hearing aid users, but calls for their promising use for NH persons as well.

- Model predictions using an individualized BSIM can explain up to 83% of the measured variance of the individual SRTs without any pre-processing, that is, the speech reception of NH as well as aided HI listeners without additional noise cancellation.

- At this stage of development, the individualized model scheme was able to estimate the possible benefits of the noise reduction algorithms for a subset of the participants. However, the model failed to give reliable predictions of signal processing benefits for each individual and all noise scenarios. Thus, further developments are necessary that would include listener characteristics other than the audiogram.

3 Modifications of the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) for use in audiology

ABSTRACT

Objective: Two modifications of the standardized MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA), namely MUSHRA simple and MUSHRA drag&drop, were implemented and evaluated together with the original test method. The modifications were designed to maximize the accessibility of MUSHRA for elderly and technically non-experienced listeners, who constitute the typical target group in hearing aid evaluation. *Design*: Three MUSHRA variants were assessed based on subjective and objective measures, e.g. test-retest reliability, discrimination ability, time exposure, and overall preference. With each method, participants repeated the task to rate the quality of several hearing aid algorithms four times. *Study Sample*: Fifty listeners grouped into five subject classes were tested, including elderly and technically non-experienced participants with normal and impaired hearing. Normal-hearing, technically experienced students served as controls. *Results*: Both modifications can be used to obtain compatible rating results. Both were preferred over the classical MUSHRA procedure. Technically experienced listeners performed best with the modification MUSHRA drag&drop. *Conclusions*: The comprehensive comparison of the MUSHRA variants demonstrates that the intuitive modification MUSHRA drag&drop can be generally recommended. However, considering e.g. specific evaluation demands, we suggest a differentiated and careful application of listening test methods.

The content of this chapter was submitted as an Original Paper to the International Journal of Audiology in 05/2016.

3.1 Introduction

For the development of new signal processing schemes, e.g. algorithms for hearing aids or audio codecs, a final stage of evaluation is crucial. In this stage the question is answered how the algorithms perform and whether they bring the anticipated benefit for the targeted group. A comprehensive evaluation can be split into an instrumental and a perceptual part. In the instrumental part, the algorithms will be evaluated using measures, which can include purely objective measures as well as perceptually motivated measures, which use auditory models mimicking the human perception. In the perceptual part, the algorithms are evaluated by real humans. Test subjects are asked for, e.g., the quality of an algorithm or other attributes like listening effort to rate different aspects of performance. Despite ongoing improvements of predictive models (e.g., PEMO-Q; Huber and Kollmeier, 2006), subjective measurements are still indispensable in an evaluation process as well as to train models of human perception to become more accurate and reliable. To gather human data several methods were developed and standardized over the years.

A standardized and commonly used method for the subjective assessment of the intermediate quality level of audio systems is the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA; ITU-R — Radiocommunication Sector of ITU, 2014a). With the MUSHRA listening test, multiple stimuli can be compared and rated simultaneously. The single stimuli are judged related to a known reference stimulus with the highest quality in the test. In addition to the known reference, the same reference stimulus as well as one (or more) anchor stimuli with intentionally low quality are hidden under the test stimuli. These stimuli have the function to define the limits for the quality range of the test and to span the quality space for the test stimuli. A screenshot of a typical MUSHRA listening test is given in Figure 3.1. The user interface shows eleven stimuli buttons. These are the reference stimulus ('Ref.') and ten test stimuli (letters A to J). The test stimuli are evaluated in terms of overall quality on the five-interval continuous quality scale (CQS; values from 0–100) by using the rating sliders above the buttons. The attributes (excellent, good, fair, poor, bad) for the five intervals of the scale are displayed to the right of the test stimuli.

The MUSHRA listening test was originally designed for evaluating audio codecs with expert listeners and not for the evaluation of hearing aid algorithms with untrained users. However, due to the lack of alternative test procedures, the MUSHRA listening test and variations of the original MUSHRA method are commonly used in hearing research and technology. Simonsen and Legarth (2010) used the MUSHRA method to evaluate the sound quality of four different premium hearing aids using

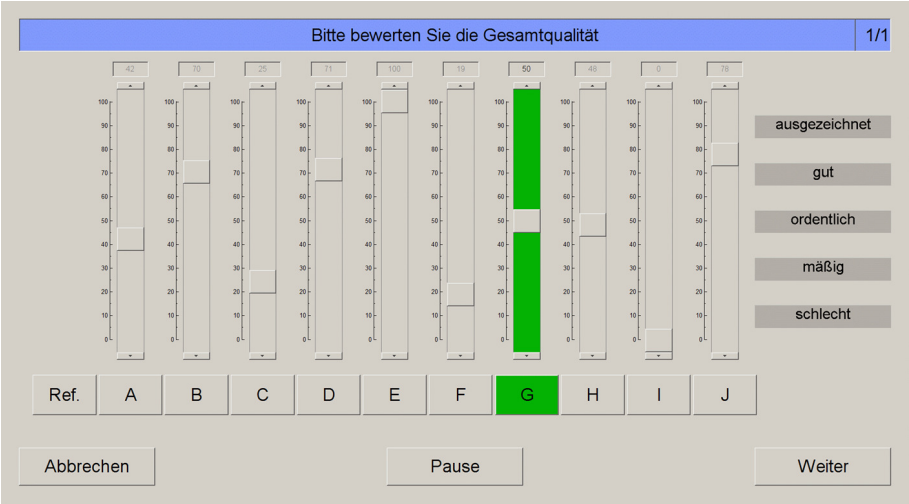

Figure 3.1: *Screenshot of a running experiment with the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA): stimulus 'G' is active at the moment.*

recorded signals. The hearing aids were tested as delivered, i.e. including noise reduction, in seven different sound environments. MUSHRA was also used by Muralimanohar et al. (2013) to measure the overall quality of different hearing aid processing schemes. Roy et al. (2012a) adapted the MUSHRA method for its use with cochlear implant (CI) users. With "CI-MUSHRA", eleven postlingually deafened CI users rated the sound quality of real-world musical stimuli with decreasing amounts of bass information. In a follow-up study, twelve postlingually deafened CI users rated musical stimuli with increasing high-frequency removal (Roy et al., 2012b). Despite the recommendation to use MUSHRA with screened or 'expert' listeners (p. 4; ITU-R — Radiocommunication Sector of ITU, 2014a) with so-called 'golden ears' (p. 99; Hardy et al., 2002), MUSHRA is used in many studies with 'naive listeners'. For example, Parsa et al. (2013) use the MUSHRA listening test with children having normal and impaired hearing. Their study investigates the impact of a nonlinear frequency compression (NFC) algorithm on the perceived sound quality with these 'naive listeners'. In another study MUSHRA is used by hearing impaired listeners with an average age of 62 years (Cubick et al., 2014).

Often the affected people regarding hearing aid usage are elderly and technically non-experienced listeners, which are therefore the preferable subject group for a realistic hearing aid evaluation. Our experience during measurements showed that such test subjects can have difficulties using the original MUSHRA listening test. Next to other challenges like a general decline of psychometric test performance associated with advancing age (e.g., Granick and Friedman, 1967; Binder et al., 1999), the listeners were challenged using the graphical user interface on a touch screen and

the classical MUSHRA listening test implementation seemed to be too complex for them.

The aims of the presented study are the introduction and evaluation of two modifications for the original MUSHRA test method to maximize its accessibility and consequently the applicability of this method. This will give us the opportunity to reliably extend the user group of this standardized method from experienced and trained subjects to the target group of elderly, technically non-experienced people with hearing impairments. A barrier-free evaluation method designed for the affected target group might help measuring the real-world benefits more properly in order to improve the development of new assistive hearing systems.

3.2 MUSHRA Modifications

For the following comparison of variants of the MUSHRA listening test, our implementation of the original MUSHRA test is called MUSHRA classic (cf. Figure 3.1). Our two modifications of the original MUSHRA test method are called MUSHRA simple and MUSHRA drag&drop. The modifications mainly concern the graphical user interface (GUI) and the general handling of the MUSHRA listening test, leaving the underlying methodology untouched as far as possible. To allow for a fair comparison, several underlying technical aspects of the three MUSHRA variants were implemented identically: The participant activates and therefore listens to the stimuli by clicking on the buttons by using the touchscreen display of the tablet computer that was used for the experiments (or with the attached mouse). A click on a button activates the underlying stimulus and deactivates the stimulus played before. For this nearly instantaneous switching, stimuli are cross-faded with a window length of 100 ms. As indicated by the recommended test procedure (p. 9; ITU-R — Radiocommunication Sector of ITU, 2014a), the subjects are constrained to be able to adjust only the score assigned to the item he or she is currently listening to with all MUSHRA variants. Regarding the rating procedure, the attributes (excellent, good, fair, poor, bad) for the five intervals of the recommended continuous quality scale (CQS) are visible in all MUSHRA variants.

3.2.1 MUSHRA simple

Using MUSHRA classic (cf. Figure 3.1) in studies for evaluation of hearing aid signal processing schemes, we experienced that elderly, hearing impaired test subjects seem to have problems with a) using the sliders on a touch screen display and b) comparing too many signals at the same time. This led to the motivation of a simplification of

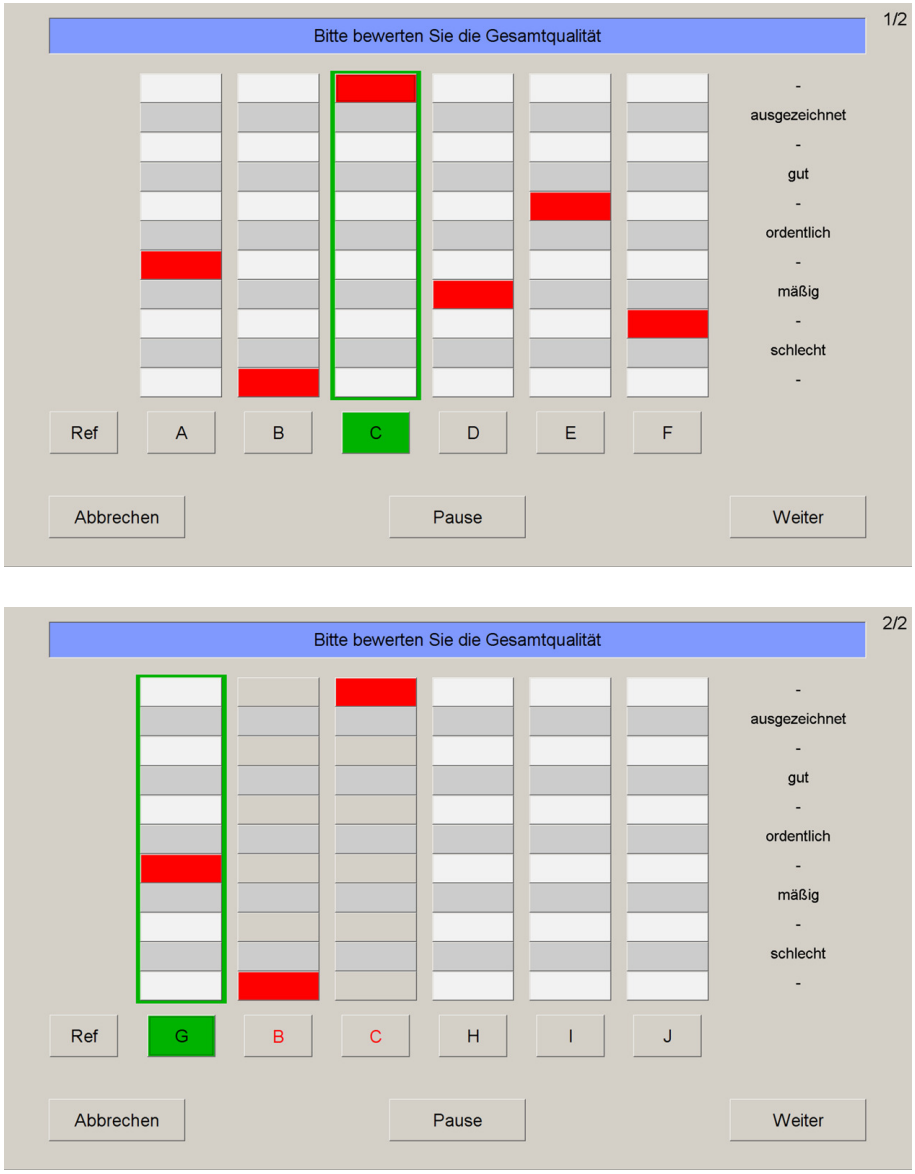

Figure 3.2: *Two screenshots of MUSHRA simple, where each trial is split onto two successive rating screens: The upper panel shows an example of final stimuli ratings for a first rating screen. The lower panel shows a running experiment during a second rating screen. In this example, buttons B (hidden low-pass anchor) and C (hidden reference) were carried over from the first screen with fixed ratings for the second screen. Buttons H, I, J were not rated yet.*

the complex method. In MUSHRA simple, instead of sliders, buttons are introduced, which should be easier to handle and should make the decision of the responses easier because of the limited resolution (eleven steps) compared to the possible values by the rating sliders (0–100). There is no numeric scale visible. The rating for each signal is given by pressing the button corresponding to the rating, in the column over the button with the letter representing the "hearing aid", which is listened to at

the moment. A screenshot of the GUI for MUSHRA simple is shown in Figure 3.2 (left panel).

To avoid having the subject to compare too many signals at the same time (i.e. more than six), a second screen is introduced (cf. Figure 3.2, right panel). On each screen only six stimuli (including hidden reference and hidden anchor) plus the visible reference are shown. The second screen is shown after pressing the "Next"-button (German: "Weiter") inside the first screen (bottom right). The buttons for the hidden reference and the hidden anchor stay fixed at their position from the first screen and their rating is taken over and cannot be changed anymore. But these two stimuli have to be listened to again also inside the second screen in order to complete the experiment.

To compare the gathered ratings of MUSHRA simple with the ratings from MUSHRA classic and MUSHRA drag&drop, the eleven possible discrete rating stages were converted to numerical values $[0, 10, 20, \ldots, 100]$.

3.2.2 MUSHRA drag&drop

As the name suggests, the graphical user interface for MUSHRA drag&drop makes use of the pointing device gesture 'drag-and-drop'. This modification for the MUSHRA test was inspired by the combined discrimination and identification procedure developed by Pfitzinger (e.g., Pfitzinger, 1998, 2003). The procedure was effectively used e.g. for measuring the perceived speech rate (Pfitzinger, 1998) and the assessment of perceived vowel quality (Pfitzinger, 2003).

In the initial state of the experiment, stimuli buttons are placed on the outside of a rating field (cf. Figure 3.3, left panel). Similar to the other MUSHRA versions, the participant activates and therefore listens to the stimuli by clicking on the single buttons. To rate a stimulus, the listener keeps the stimulus button pressed and now drags the stimulus into the the rating field and drops it onto the desired position. In contrast to the classical MUSHRA test and our modification MUSHRA simple, where the rating scales are arranged vertically (bottom: lowest rating; top: highest rating), the rating process with MUSHRA drag&drop works horizontally from left (lowest rating) to right (highest rating). According the continuous quality scale (CQS) used in the recommended MUSHRA test procedure, the rating field from MUSHRA drag&drop is divided into five equal intervals. The adjectives from the CQS are used as headings for the intervals. An example for a final stimuli rating and button positions at the end of a test run is displayed in the right panel of Figure 3.3. The button 'Ref', holding the reference stimulus, is fixed on the rightmost side inside the rating field and can not be moved away from its position. In addition to the

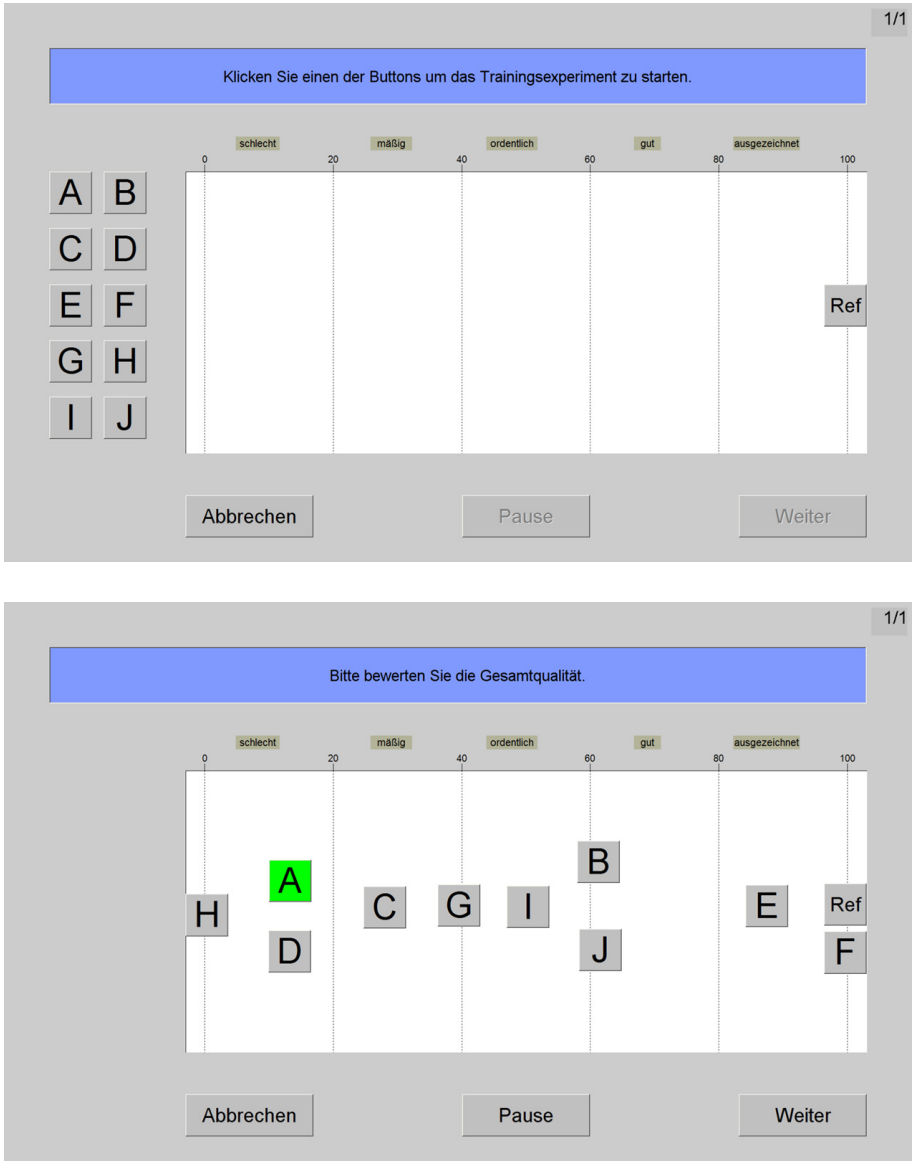

Figure 3.3: *Two screenshots of MUSHRA drag&drop: The upper panel shows the GUI at the beginning of a test run with initial positions of the stimuli buttons outside the rating field. An example for a final stimuli rating and button positions at the end of a test run is displayed in the lower panel.*

written instructions this is another visual cue for the participant about the meaning of the reference stimulus: It is implied that this stimulus has the highest quality rating with a quality score of 100. The other stimulus buttons can be clicked and moved by the participant as long as desired. The two-dimensional rating field allows the participant to give the same rating to several stimuli. The arrangement in terms of height inside the rating field is of no meaning for the final quality analysis. For this, the position of each stimulus button on the CQS is evaluated leading to quality scores from 0 to 100.

MUSHRA drag&drop offers the benefit of an instantaneous visualization of a final stimulus ranking. By successively clicking the stimulus buttons from left to right or from right to left, the assessor can easily check the ratings and alter them where necessary. The horizontal layout of the continuous quality scale combined with a sorting direction of the stimuli from left to right, which corresponds to the direction of writing of the participants, should offer an intuitive and plausible way of control and should lead to an easy-to-learn test method.

3.3 Method

3.3.1 Participants

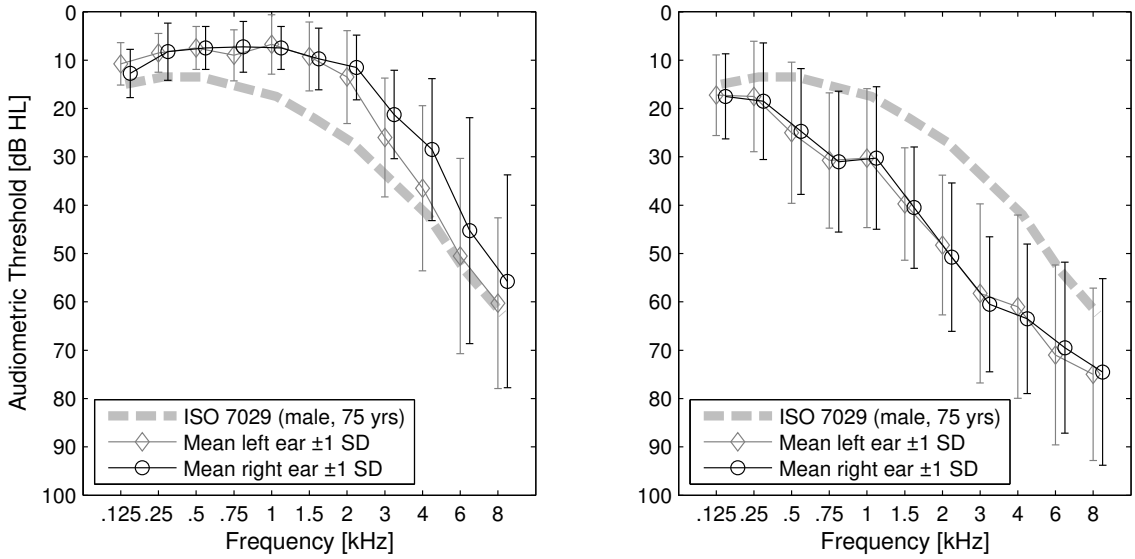

Figure 3.4: *Audiometric data for twenty otologically normal-hearing listeners (left panel) and twenty hearing-impaired listeners (right panel): The mean thresholds (marked for the left ear with diamonds and with circles for the right ear) ± standard deviation are displayed. Additionally, the median values for otologically normal persons (male, 75 years) are shown in both panels (thick dashed lines).*

The evaluation of the modifications for the MUSHRA test was performed with fifty participants in total.

We arranged five subject groups with N= 10 listeners each to investigate our assumptions for the developed modifications: a) the modifications are especially beneficial for elderly, technically non-experienced listeners, and b) this benefit also holds for hearing-impaired listeners.

The five subject groups (G1–G5) differ in the factors age, technical experience, and hearing ability:

G1: young, technically experienced, normal hearing (control group),

G2: elderly, technically experienced, normal hearing,

G3: elderly, technically non-experienced, normal hearing,

G4: elderly, technically experienced, impaired hearing,

G5: elderly, technically non-experienced, impaired hearing.

We defined the levels of the subject factors as follows: young listeners should be younger than 30 years, elderly participants have to be over 60 years. Our definition of normal hearing with respect to impaired hearing follows the description of otologically normal hearing (European Committee for Standardization, 2001), which takes into account the progressive declining sensitivity of human hearing with age. Furthermore the participants with impaired hearing from this study were experienced hearing aid users, whereas the elderly participants with otologically normal hearing do not own or wear hearing aids on a daily basis. The definition concerning the technical experience is described in the following.

For each of the participants from groups G2–G5, information about the willingness of technology use ('technology commitment') was collected using the scale by Neyer et al. (2012). Their scale consists of twelve statements, that are to be rated on a five-point scale. The model takes into account the three sub-scales technology acceptance, technology competence, and technology control, each represented by four items in the questionnaire. The questionnaire was sent by mail to all participants of the current study. Neyer et al. (2012) found that technology commitment is correlated negatively with factor age. As for the factor gender, they discovered significantly higher technology commitment values for men. To distinguish between the different willingness of technology use and to account for the effects of factors age and gender, the technology commitment data from the listeners was divided into age- and gender-related terciles. Tercile limits regarding technology commitment values were defined for four age groups (1: $<$ 50 yrs, 2: 50–64 yrs, 3: 65–79 yrs, 4: $>$ 79 yrs) for both genders. Listeners from the first tercile (group 1) are considered being unwilling of technology use, whereas listeners from the second percentile (group 2) have a moderate relation to technology, and listeners from the third tercile (group 3) are supposed to show a high technology commitment. The definition of the terciles (low, medium, and high technology commitment) was based on a re-analysis of the data set (n= 2.032, people $>$ 50 yrs.) of a representative study in Lower Saxony (Künemund and Tanschus, 2014). To evaluate the MUSHRA modifications regarding the factor technical experience, we defined listeners belonging to the first terciles as

technically non-experienced, listeners belonging to the third terciles as technically experienced. During the break on the first test day (see below), each participant filled the questionnaire regarding technology commitment again. This retest revealed that three of twenty participants with a high technology commitment (regarding the results from the first survey) now exhibit a medium commitment. More interestingly, from twenty listeners invited based on their low technology commitment, eight were showing a medium commitment, and two exhibiting a high commitment in the retest. As the retest was performed directly after the participants used all MUSHRA variants successfully, the higher commitment values might be explained by the time of the inquiry (cf. Forberg and Neyer, submitted 2014).

The 'control group' G1 (young, technically experienced, normal hearing) consisted of students (five male, five female) of natural sciences with self-reported high technical experience and self-reported normal hearing. Their age ranges from 20–27 years (mean = 23.8 years). Audiometric measurements of G1 reveal averaged thresholds over the frequencies 500, 1000, 2000, and 4000 Hz (4PTA) from −2.5 dB to 11.25 dB, which assures normal hearing of the control group G1. The forty elder participants from groups G2–G5 are aged between 62 years and 86 years with a mean age of 73.8. To relate the listeners from this study to otologically normal hearing persons, thresholds (male, 75 years) were calculated using the standardized thresholds from European Committee for Standardization (2001). Groups G2 and G3 together hold twenty elder participants with otologically normal hearing, whereas G2 consists of ten technically experienced listeners (four male, six female) with a mean age of 73.1 years, and G3 of ten technically non-experienced listeners (four male, six female) with a mean age of 73.2 years. The factor hearing impairment is represented in groups G4 and G5. These consist of ten technically experienced subjects (G4: six male, four female, mean age=74.6 years) and ten technically non-experienced listeners (G5: eight male, two female, mean age=74.3 years).

The measured audiometric data for the twenty elder participants with otologically normal hearing (left panel: G2 and G3) and twenty elder listeners with impaired hearing (right panel: G4 and G5) is shown in Figure 3.4. The mean thresholds (marked for the left ear with diamonds and with circles for the right ear) and error-bars indicating ± one standard deviation are displayed. The calculated thresholds from ISO 7029 (male persons, 75 years) are shown in both panels (thick dashed lines).

3.3.2 Signal material & Equipment

By using the original MUSHRA method (MUSHRA classic) and our two modifications MUSHRA simple and MUSHRA drag&drop, all subjects rated the overall quality of seven different noise reduction algorithms designed for the application in hearing aids, such as a directional microphone (Elko and Nguyen Pong, 1995) or a coherence-based noise reduction scheme (Grimm et al., 2009). In addition to the noise reduction algorithms a 'no processing scheme' is part of the stimuli portfolio. The subjects rated the noise reduction algorithms in three different realistic noise scenarios. The signal material for the scenarios consists of scene-specific dialogues in a kitchen, a supermarket and a cafeteria. All noise scenarios were generated using a toolbox for acoustic scene creation and rendering (TASCAR; Grimm et al., 2013, 2014) having a signal duration of 20 seconds, a signal-to-noise ratio of 2 dB, and a sample rate of 44.1 kHz. The scenarios were created exhibiting realistic overall sound pressure levels (cafeteria: 72 dB, kitchen and supermarket: 65 dB).

Being part of the recommended test procedure (ITU-R — Radiocommunication Sector of ITU, 2014a), we defined reference and anchor signals for each noise scenario. The reference signal appears visibly and also hidden in the test, the anchor signal only appears hidden under the test stimuli. The reference signals consist of stimuli with a high signal-to-noise ratio (SNR) of 20 dB. As the remaining stimuli exhibit a SNR of 2 dB, this leads to reference stimuli having the highest quality in the test. The definition for the anchor signals follows the recommendation (p. 7, ITU-R — Radiocommunication Sector of ITU, 2014a). The anchor is generated from a low-pass filtered version of the original signal ('no processing scheme') with a cut-off frequency of 3.5 kHz. The anchor stimuli are supposed to show the worst (or at least a very low) quality in the test. The noise reduction algorithms were implemented in the Master Hearing Aid (MHA; Grimm et al., 2006). Using the MHA, the signal material for the three noise scenarios was processed by the noise reduction schemes to be usable for offline processing. The individual signals for the hearing-impaired listeners from subject groups G4 and G5 were additionally processed by a multiband compressor scheme to compensate their hearing loss. The compressor was applied serially after the noise reduction schemes. Insertion gains for the compressor were calculated with the nonlinear fitting procedure NAL-NL1 (Byrne et al., 2001) based on actual measurements of the individual hearing threshold levels of the hearing-impaired listeners (cf. Figure 3.4, right panel).

The three used MUSHRA versions were implemented using the software MatLab. The necessary signal processing including instantaneous switching of signals was provided by the software SoundMexPro. The test software runs on an Acer Iconia

W700 tablet with external USB soundcard RME Fireface UC providing the D/A conversion. The signals were amplified with the Tucker Davis HB-7 headphone buffer and presented to the listeners with headphones (Sennheiser HDA200). The frequency response of the headphones was free-field-equalized using an appropriate filter inside the software SoundMexPro. After equalization, the desired output sound pressure levels were ensured using a calibration routine inside SoundMexPro with broadband noise. The output pressure levels from the headphones were measured with an ear simulator according to IEC 60318-1. The experiments were performed in a soundproofed test room (ANSI/ASA S3.1-1999, 2008).

3.3.3 Measurement procedure

Each participant performed the quality ratings using the MUSHRA versions twice on two different days. Out of the six possible sequences for the use of the three MUSHRA versions, each participant is assigned with a fixed test method order. All six possible sequences are almost equally distributed over the fifty participants (ten listeners per subject group), to ensure that the sequence of the MUSHRA variants has no effect on the evaluation results. The test protocol provides quality ratings using the first MUSHRA variant (e.g., MUSHRA simple) in the three noise conditions presented randomly, followed by quality ratings in three noise conditions using the second MUSHRA variant (e.g., MUSHRA classic), completed by quality ratings with the third MUSHRA variant (e.g., MUSHRA drag&drop), again in three noise conditions. After having used all MUSHRA variants, the protocol provides a break, which is followed by a repetition of the quality ratings using all three MUSHRA variants in the same sequence as before. Each subject performed these two test series on two different days, i.e. every participant used each MUSHRA variant four times. Considering the three noise conditions, each participant gave 36 ratings per algorithm (three noise conditions, four test runs, three test methods), leading to 1800 judgments per algorithm in total. Before using each MUSHRA variant for the first time on a test day, the participants were given written instructions. This was followed by a training phase for each MUSHRA test method, in which the participant rated the stimuli in one randomly presented noise scenario. The experimenter was present during this training phase to support the listener if necessary with more oral instructions, ensuring that all listeners are able to use each MUSHRA method correctly. The duration of one test day completing two test series maximally took 2.5 hrs. The listeners got paid for their participation and their travel expenses.

Approval for this experiment was obtained from the Carl von Ossietzky Universität Oldenburg ethics committee on 08.01.2009 and 06.09.2013.

3.3.4 Comparison measures

Next to comparing the final results regarding the algorithm quality ratings collected with each of the methods, the evaluation of the MUSHRA modifications is done by comparing the results of several subjective and objective measurements, which will be presented in the following.

3.3.4.1 Subjective measurements

To examine the usability of the MUSHRA methods for all subject groups, we used the **system usability scale** (SUS; Brooke, 1996). The SUS questionnaire consists of ten test items to be rated on the five-point Likert scale (1 : strongly agree — 5 : strongly disagree). SUS scores (ranging from 0 to 100) are calculated from the given answers to the test items. The SUS is a simple and fast method for rating usability. We used the German translation of the SUS (Glende et al., 2011) with slight modifications of the test items as the SUS originally aims at the rating of systems or products. Our modifications of the SUS aim at the rating of computer programs for hearing experiments. The SUS for each MUSHRA method was measured directly after the second usage of this MUSHRA variant on a test day, yielding to a test and retest value for the usability of each method by all listeners. As another subjective measurement a **preference ranking** for the MUSHRA versions was observed. The listeners were asked to assign numbers from 1 to 3 to the MUSHRA versions regarding their personal preference. Together with the values concerning the usability, the preference ranking data also was collected after the second usage of each MUSHRA variant on a test day, leading to two preference rankings (test and retest).

3.3.4.2 Objective Measures

As objective measures to compare the MUSHRA versions we measured the **duration** of each trial (one noise scenario) when using each MUSHRA method. We also measured the amount of **button clicks** for each MUSHRA version, i.e. the amount of clicks for the (visible) reference button, as well as the amount of overall stimuli button clicks was recorded. As another objective measure for the comparison of the MUSHRA versions, we evaluated the **performance of each assessor**. As each assessor from this study used all three MUSHRA versions (in a randomized order), the comparison of MUSHRA versions in terms of assessor performance is legitimate. Since 2014, the recommendation for the MUSHRA test procedure includes an obligatory selection of assessors (ITU-R — Radiocommunication Sector of ITU, 2014a): As ratings from naive assessors would degrade the quality of the data and

the estimates of central tendency, only assessors categorized as experienced assessors for any given test should be included in the final data analysis.

The necessary calculations for a distinction between naive and experienced assessors can be performed by the methods for assessor screening, that are reported by the ITU-R (ITU-R — Radiocommunication Sector of ITU, 2014b). These methods are based on a model for expertise gauge (eGauge; Lorho et al., 2010), which exploits the idea by Schlich (1994) to split the global disagreement of a listener panel into separate contributions of each assessor.

Using the methods for assessor screening by the ITU-R (ITU-R — Radiocommunication Sector of ITU, 2014b), we calculated two metrics for each assessor, **reliability** and **discrimination**. Reliability is a measure of the closeness of repeated ratings of the same test item and discrimination is a measure of the ability to perceive differences between test items. The calculation of the assessor performance metrics is based on analysis of variance (ANOVA). The factors algorithm and noise reduction are merged to create a new factor stimulus. After this reduction of dimensions, a one-way ANOVA with all repeated ratings for each stimulus is computed for each assessor j, leading to a value for the mean square from factor stimulus ($\mathrm{MSS_j}$) and a mean square of the residual term ($\mathrm{MSR_j}$). The variance explained by the factor stimulus is the variance explained by the experimental design. Also the averaged standard deviation of a score $\mathrm{SPAN_j}$ is calculated for each assessor j. SPAN is the mean value of the calculated $\mathrm{SPAN_j}$. The $\mathrm{reliability_j}$ is defined as the ratio between SPAN and the square-root of the residual mean square $\mathrm{MSR_j}$ (cf. equation 3.1). The $\mathrm{discrimination_j}$ is the ratio of the mean square for factor stimulus $\mathrm{MSS_j}$ and the mean square of the residual term $\mathrm{MSR_j}$ and thereby the F-ratio of the computed ANOVA (cf. equation 3.2). This measure can be interpreted as the 'signal-to-noise ratio of the repeated rating of a set of stimuli by an assessor' (p. 192; Lorho et al., 2010).

$$\mathrm{reliability_j} = \frac{\mathrm{SPAN}}{\sqrt{\mathrm{MSR_j}}} \tag{3.1}$$

$$\mathrm{discrimination_j} = \frac{\mathrm{MSS_j}}{\mathrm{MSR_j}} \tag{3.2}$$

For the categorization of assessors (naive versus experienced), a non-parametric permutation test (Dijksterhuis and Heiser, 1995) with 10 iterations per assessor is used as a test of significance for both metrics. The permutation test defines the so-called noise floor of the assessor performance concerning reliability and discrimination. Only assessors performing better than the test levels for both metrics are regarded as experienced assessors. Concerning the comparison of MUSHRA versions in terms

of assessor performance, we compare a) the collected 'raw' data for reliability and discrimination, and b) the percentages of data above the calculated noise floors for each MUSHRA version.

3.3.5 Statistical Analyses

The collected data were analyzed using the statistical software IBM SPSS. Data for algorithm ratings, system usability scale, duration, button clicks, and assessor performances were fed into single mixed-design analysis of variance (ANOVAs) with between- and within-subject factors. Similar to the model for calculating the assessor performance, where algorithms ('systems') and noise conditions ('samples') are merged to one factor 'stimuli' (p. 3, ITU-R — Radiocommunication Sector of ITU, 2014b), we averaged the algorithm ratings over the three tested noise conditions before analyses of variance. Whenever necessary, violations of sphericity were adjusted using the Greenhouse-Geisser correction. For the analysis of the subjective preference ranking of the MUSHRA versions, a non-parametric test (Friedman's ANOVA) was applied. In case of significant effects or interactions, a posthoc analysis with Bonferroni corrections was performed. The results were reported in terms of adjusted significance levels, p_{Bonf}.

3.4 Results

3.4.1 Algorithm ratings

The averaged overall quality ratings for the tested algorithms are shown in Figure 3.5. The mean ratings for each algorithm and each test method were calculated from overall 600 judgments (50 participants, three noise scenarios, four test runs). To compare the ratings between the used test methods, the eleven discrete rating states from MUSHRA simple were transformed to numeric values $[0, 10, 20, \ldots, 100]$. The errorbars denote a range of $\pm$ 1 standard deviation from the mean value. The ratings are displayed separately for the three MUSHRA versions used. The order of the rated stimuli from left to right is (hidden) low-pass anchor with a supposedly low rating, followed by the non-processed stimuli and stimuli processed by seven different noise reduction algorithms, and the (hidden) reference stimuli with the expected highest rating on the right. The displayed averaged quality ratings show very similar final ratings for the algorithms collected with all three tested MUSHRA versions.

However, statistical analysis reveals that within-subjects factor test method significantly affects the quality ratings, $F(2,90) = 29.2$, $p < 0.001$. A post hoc analysis

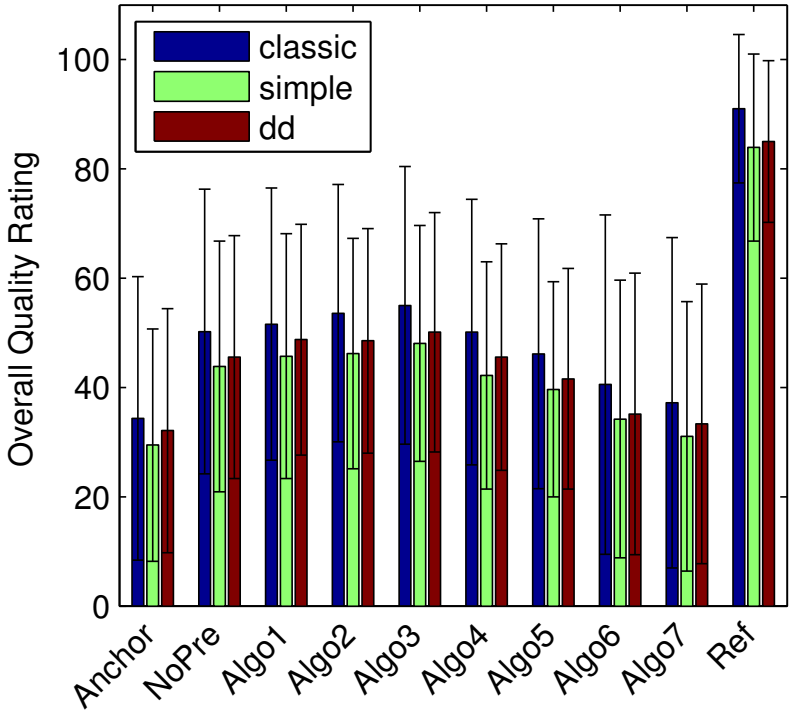

Figure 3.5: *Overall quality ratings averaged over all participants for the tested algorithms and methods (classic: blue, simple: green, drag&drop: red); the errorbars denote a range of ± 1 standard deviation from the mean value.*

indicates that paired comparisons between all MUSHRA variants differ significantly from each other on adjusted significance levels. The mean judgments using MUSHRA classic were 6.6 points higher than MUSHRA simple ($p_{\mathrm{Bonf}} < 0.001$) and 4.4 points higher than MUSHRA drag&drop ($p_{\mathrm{Bonf}} < 0.001$), i.e. the ratings using MUSHRA drag&drop were 2.2 points higher than MUSHRA simple ($p_{\mathrm{Bonf}} < 0.023$). Also the main effects algorithm and test run number significantly affect the quality ratings, $\mathrm{F}_{\mathrm{algorithm}}(2.6,115.4) = 117.7$, $\mathrm{p} < 0.001$ and $\mathrm{F}_{\mathrm{test\ run}}(2.2,99.4) = 5.9$, $\mathrm{p} = 0.003$. As the effect caused by the different algorithms on the ratings is intended, the effect of test run needs a closer examination. On both measurement days, the ratings from the second test run were significantly higher (first day: 2.7 points, $p_{\mathrm{Bonf}} = 0.002$, second day: 2.0 points, $p_{\mathrm{Bonf}} = 0.004$). The differences between the first test runs on both days and the second test runs on both days were not significant. This finding can be explained by a customization effect: Although the listeners performed a training session in one noise condition with each MUSHRA variant (see above), the participants need one complete test run (with all method variants and all noise conditions) on both days to get known to the tested algorithms. To check whether this behavior for factors algorithm and test run changes in dependence on the used MUSHRA variant, we investigated the interactions of factor test method with the

other main effects. Both main factors do not interact statistically significant with factor test method (algorithm: $p > 0.1$, test run: $p > 0.9$).

Considering the between-subjects factor group, no significant effect was found on the quality ratings ($p > 0.3$). Also the interaction of test method with between-subjects factor group was not found to be affecting the quality judgments of the listeners ($p > 0.4$). We found a significant interaction of within-subjects factor algorithm and factor group, $F(10.3,115.4) = 5.4$, $p < 0.001$. Summing up the significant differences between the averaged algorithm ratings within G1, we find 32 of the $\binom{10}{2} = 45$ paired comparisons being significantly different. Within the other groups, less averaged comparisons are significantly different (G2: 18, G3: 12, G4: 17, G5: 9). Within G5, only the nine differences with respect to the hidden high quality reference were found to be significant. From this we can conclude that the participants from control group G1 differentiate more between the algorithms than the listeners from the other subject groups.

3.4.2 System Usability Scale

The statistical analysis concerning the subjective evaluation of the test methods with the system usability scale does not reveal any significant effects. Neither between-subjects factor group, nor within-subjects factors test method or test run affect the calculated values for the system usability scales. The collected SUS values for MUSHRA classic are 85.6 (test) and 86.0 (retest), for MUSHRA simple 86.4 (test) and 90.6 (retest), for MUSHRA drag&drop 86.0 (test) and 88.1 (retest).

3.4.3 Preference Ranking

Averaging over the fifty participants from all groups, we find that factor test method significantly affects the preference rankings in both data acquisitions, test and retest (test: $\chi^2(2) = 7.3$, $p < 0.05$; retest: $\chi^2(2) = 17.9$, $p < 0.001$). The preference order in both tests is the same: MUSHRA simple is ranked best (test: mean rank of 1.75, retest: mean rank of 1.62), MUSHRA drag&drop is second-placed (test: mean rank of 1.98, retest: mean rank of 1.93), and MUSHRA classic is on the third place (test: mean rank of 2.27, retest: mean rank of 2.45). A post hoc analysis with Wilcoxon signed-rank tests and Bonferroni corrections reveals that differences in preference rankings between MUSHRA simple and MUSHRA classic are significant (test: $T = 309$, $r=-0.27$, $p < 0.0167$; retest: $T=248$, $r=-0.38$, $p_{Bonf} < 0.0167$). Differences between MUSHRA drag&drop and MUSHRA classic are significant in the retest ($T=328$, $r=-0.28$, $p_{Bonf} < 0.0167$). No significant differences between preference

ranking for MUSHRA simple and MUSHRA drag&drop were found. Analyzing the data separately for each subject group, we find MUSHRA classic in none of the groups as being the most preferred test method. Moreover, in groups G1–G4, both MUSHRA variants are preferred over the original MUSHRA. In group G5, the modification MUSHRA simple is preferred over MUSHRA classic.

3.4.4 Duration

Analysis shows that within-subjects factor test method influences the duration per trial, $F(1.6,74.0) = 7.6$, $p = 0.001$. The averaged duration values for each test method are 201s (MUSHRA classic), 206s (MUSHRA simple), and 175s (MUSHRA drag&drop). Post hoc analysis reveals that the duration value of MUSHRA drag&drop is significantly smaller than MUSHRA classic ($p_{Bonf} = 0.002$) and than MUSHRA simple ($p_{Bonf} < 0.001$). MUSHRA classic and MUSHRA simple do not differ significantly in duration per trial. Factor test run (four test runs) also affects the duration, $F(2.6,116.9) = 14.1$, $p < 0.001$. The mean values for duration decrease over the test runs from 222s to 177s. Post hoc analysis shows that there are significant differences between test runs on the two different test days and not inside of one test day.

The interaction of test method and test run is also affecting the measured duration values for each trial, $F(4.3,191.4) = 3.4$, $p = 0.009$. Post hoc analysis shows that this general effect of decreasing time over the four test runs is differently pronounced within the different MUSHRA versions. For MUSHRA drag&drop, already the second run on the first testing day is significantly shorter than the first run (44 seconds, $p_{Bonf} = 0.001$), and the measured duration values for test runs 2–4 do not differ significantly anymore. For MUSHRA classic and MUSHRA simple the significant differences between test runs only occur between the two different test days. Between-subjects factor group influences the averaged duration per trial, $F(4,45) = 16.6$, $p < 0.001$. The averaged trial duration values for each group are G1: 119s, G2: 209s, G3: 219s, G4: 177s, G5: 245s. Post hoc analysis shows that control group G1 rates the algorithms significantly faster than all other groups, at which listeners from G5 need twice the time. The duration difference of 68s between G4 and G5 is also significant ($p_{Bonf} = 0.002$). The mentioned effect of factor group does not interact with factor test method ($p>0.3$).

3.4.5 Button Clicks

The factor test method significantly affects the amount of **reference button clicks**, $F(1.3,59) = 16.9$, $p < 0.001$. Using MUSHRA classic or MUSHRA drag&drop the

reference button is clicked 4.2 times per trial on average. Using MUSHRA simple, the reference button is clicked more often (5.2 times), which is significant with $p_{Bonf} < 0.001$. Between-subjects factor group was not found to be influencing the amount of reference button clicks ($p>0.7$). Also, the factor group is not interacting significantly with factor test method ($p>0.07$). The higher amount of reference button clicks with MUSHRA simple is most likely due to the splitting of the stimuli on two successive test screens: The participant has to click and listen to the reference stimulus on both screens to finish the actual trial of the experiment. Different to the other two MUSHRA versions with at least one reference click per trial, the minimum value of reference clicks is two.

Concerning the overall **stimuli button clicks** per trial, factor test method affects the amount of button clicks, $F(2,90) = 6.4$, $p = 0.002$. The averaged amounts of clicks are 22.6 (MUSHRA classic), 24.1 (MUSHRA simple), and 25.2 (MUSHRA drag&drop). Pairwise comparisons with post hoc analysis show that the difference between MUSHRA classic and MUSHRA drag&drop (2.6 clicks) is significant with $p_{Bonf} = 0.004$. Identical to the analysis of the reference button clicks, between-subjects factor group was not found to be significant ($p>0.3$). The interaction of factor test method with factor group also does not affect the amount of overall stimuli button clicks ($p>0.5$).

3.4.6 Assessor Performance

The between-subjects factor test method affects both measured metrics concerning the assessor performance, reliability and discrimination, $F_{reliability}(1.7,76.9) = 8.9$, $p = 0.001$ and $F_{discrimination}(2,90)=20.6$, $p<0.001$. The mean values for reliability separated by test method are 0.35 (MUSHRA classic), 0.25 (MUSHRA simple), and 0.38 (MUSHRA drag&drop). The averaged values for the ability to discriminate between the stimuli are 1.49 (MUSHRA classic), 1.05 (MUSHRA simple), and 1.52 (MUSHRA drag&drop). Posthoc analysis shows that the assessor performance metrics for MUSHRA classic and MUSHRA drag&drop are significantly higher than for MUSHRA simple. Regarding reliability, MUSHRA classic has higher values than MUSHRA simple with $p_{Bonf} < 0.05$, MUSHRA drag&drop has higher values with $p_{Bonf} < 0.001$. The differences regarding discrimination ability between MUSHRA simple and the other two MUSHRA versions are significant with $p_{Bonf} < 0.001$. The differences between MUSHRA classic and MUSHRA drag&drop are not significant.

Between-subjects factor group affects the discrimination ability significantly, $F(4,45) = 3.4$, $p < 0.05$. The measures for reliability are not affected by the factor group ($p>0.1$). Post hoc analysis for the discrimination abilities shows

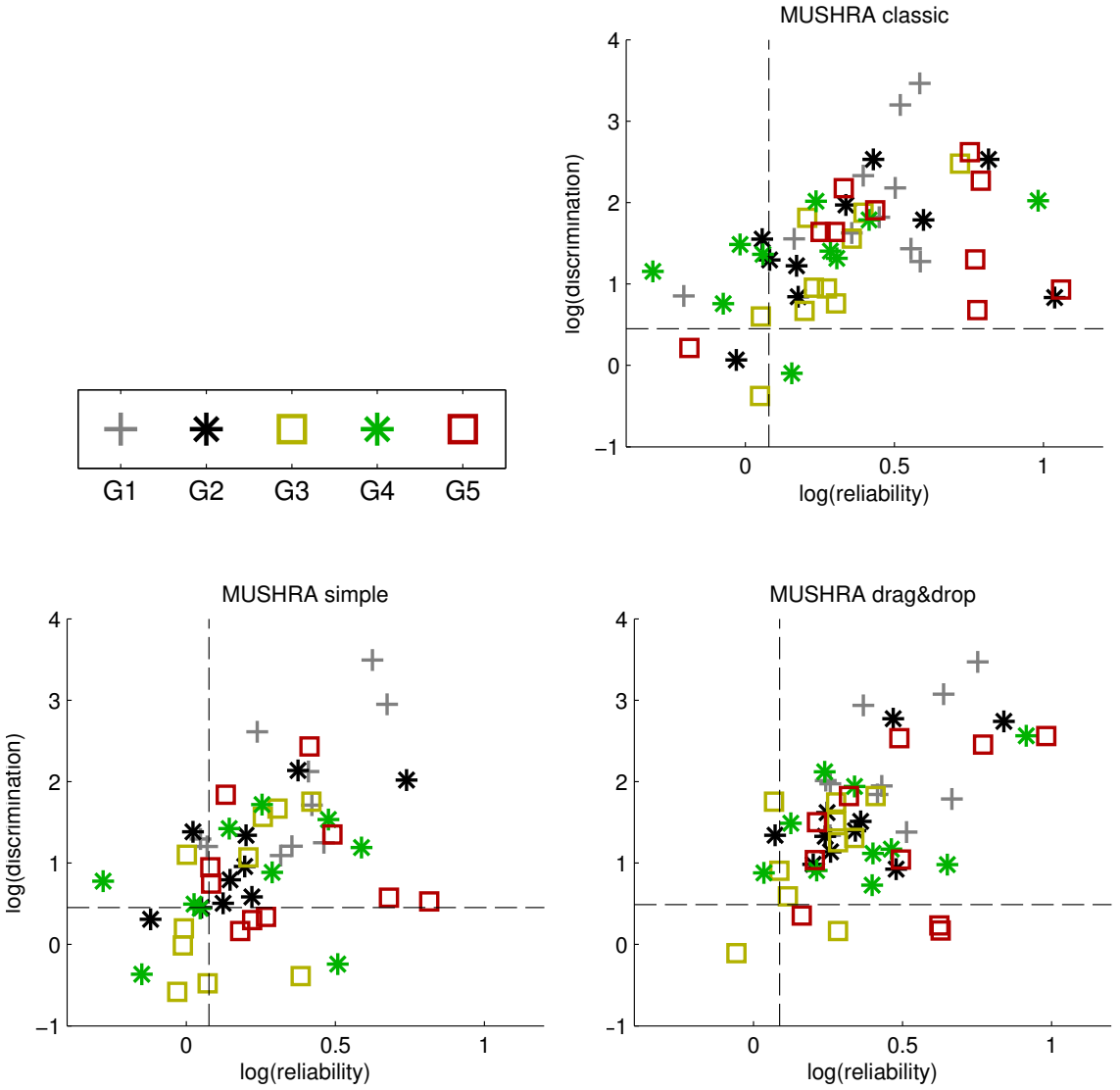

Figure 3.6: *Scatterplot of reliability vs. discrimination for the assessors using MUSHRA classic (top right), MUSHRA simple (bottom left), and MUSHRA drag&drop (bottom right). Plus sign: Control group G1, stars: technically experienced subjects (G2, G4), squares: technically non-experienced subjects (G3, G5). Also displayed (dashed lines) are the noise floors for both metrics: In total, 11/50 (22%) assessors have to be excluded from a mean value analysis when using MUSHRA classic, 19/50 (38%) assessors for MUSHRA simple, and 9/50 (18%) assessors using MUSHRA drag&drop.*

that control group G1 has a significant higher ability to discriminate ($\Delta = 1.1$, $p_{Bonf} < 0.05$) between the test stimuli than group G3. As no significant interactions of test method with factor group were found for both assessor metrics ($p_{reliability}$>0.2, $p_{discrimination}$>0.6), we can state that the differing behavior of the subject groups regarding discrimination is not affected by the used MUSHRA variants.

The analyses based on assessor performance including the calculated noise floors for each used MUSHRA variant are displayed in Figure 3.6. The upper right panel shows a scatterplot of the assessor performance metrics reliability and discrimination when using MUSHRA classic. With respect to the recommendation (p. 4, ITU-R — Radiocommunication Sector of ITU, 2014a), assessors with values below the noise floors should be excluded from the final mean value analysis of the ratings. In this sense, eleven assessors showed problems using the MUSHRA classic user interface; one assessor from the control group G1, two listeners each from G2 and G3, five assessors from G4, and one participant from G5. In the bottom left panel of Figure 3.6 the scatterplot for MUSHRA simple is displayed. Here, 19 assessors showed problems using the MUSHRA simple user interface: two assessors from the control group G1, three from G2, six from G3, five from G4 and another three from G5. The bottom right panel shows the assessor results for MUSHRA drag&drop. Using the drag & drop user interface nine assessors showed problems, which would result in an exclusion from the final data analysis: one assessor from G2, four participants from G3, one from G4, and three from G5.

The percentages of the fifty participants producing data above the noise floors with each listening test method are 78% (MUSHRA classic), 62% (MUSHRA simple), and 82% (MUSHRA drag&drop). Ranking these data, we find the highest amount of data above the noise floors with our modification MUSHRA drag&drop, followed by MUSHRA classic and MUSHRA simple. With the exception of two listeners (one from G3, one from G4), all subjects were able to produce reliable rating results with at least one MUSHRA variant. Three participants could only produce valid rating results with MUSHRA classic (G2: one, G3: two), six subjects only by using MUSHRA drag&drop (G1: one, G2: one, G3: one, G4: four).

An overview of the results for the MUSHRA method evaluation is given in Table 3.1. The table shows the results for the algorithm ratings and the results for the subjective and objective comparison measures separately for each MUSHRA method and each subject group G1–G5. With the exception of the assessor data above noise floors, significant differences between the MUSHRA variants are denoted with ^{a,b,c} on the top level (regarding all groups) for each comparison measure.

3.5 Discussion

3.5.1 Negligible differences between algorithm ratings

We state that the found significantly highest ratings with MUSHRA classic (6.6 points higher than MUSHRA simple, 4.4 points higher than MUSHRA drag&drop) can

Table 3.1: *Overview of the MUSHRA method evaluation: Averaged results for the algorithm ratings and the subjective and objective comparison measures are displayed separately for each MUSHRA variant and each subject group G1–G5. With the exception of the assessor data above noise floors, significant differences between the MUSHRA variants are denoted with* ^{a,b,c} *on the top level (regarding all groups) for each comparison measure.*

Measure	MUSHRA classic	MUSHRA simple	MUSHRA drag&drop
algorithm quality ratings	$51.0^{b,c}$	$44.4^{a,c}$	$46.6^{a,b}$
G1	49.0	45.1	45.4
G2	52.4	42.9	46.4
G3	58.7	51.6	54.2
G4	47.1	43.2	43.1
G5	47.7	39.4	43.9
system usability (test, retest)	85.6 (86.0)	86.4 (90.6)	86.0 (88.1)
G1	83.25 (84.25)	84.25 (92.25)	86.75 (89.75)
G2	90.5 (90.5)	89.5 (90.25)	87.0 (93.25)
G3	82.5 (84.5)	83.5 (86.75)	83.25 (82.25)
G4	89.25 (84.5)	91.25 (95.0)	90.0 (90.25)
G5	82.5 (86.0)	83.5 (88.75)	82.75 (85.0)
preference ranking (test, retest)	2.27 (2.45)	1.75^{a} $(1.62)^{a}$	1.98 $(1.93)^{a}$
G1	2.3 (2.3)	1.8 (1.85)	1.9 (1.85)
G2	2.4 (2.7)	1.8 (1.55)	1.8 (1.75)
G3	2.6 (2.55)	1.8 (1.5)	1.6 (1.95)
G4	2.1 (2.8)	1.8 (1.5)	2.1 (1.7)
G5	1.95 (1.9)	1.55 (1.7)	2.5 (2.4)
duration per trial (s)	201	206	$175^{a,b}$
G1	128	116	115
G2	213	228	184
G3	213	252	193
G4	191	175	165
G5	258	261	217
reference button clicks per trial	4.2	$5.2^{a,c}$	4.2
G1	2.9	3.9	2.8
G2	5.1	5.8	4.1
G3	3.7	5.1	3.4
G4	4.5	5.5	4.7
G5	4.8	5.9	6.0
stimuli button clicks per trial	22.6	24.1	25.2^{a}
G1	28.4	28.9	31.9
G2	22.8	24.1	26.4
G3	21.6	22.2	23.6
G4	19.6	21.5	22.7
G5	20.6	24.0	21.6
reliability	0.35^{b}	0.25	0.38^{b}
G1	0.39	0.36	0.45
G2	0.37	0.2	0.35
G3	0.28	0.16	0.21
G4	0.2	0.19	0.38
G5	0.53	0.34	0.49
discrimination	1.49^{b}	1.05	1.52^{b}
G1	1.97	1.89	2.18
G2	1.46	1.05	1.58
G3	1.13	0.59	1.1
G4	1.32	0.79	1.4
G5	1.54	0.92	1.37
data above noise floors	78%	62%	82%
G1	90%	80%	100%
G2	80%	70%	90%
G3	80%	40%	60%
G4	50%	50%	90%
G5	90%	70%	70%

^{a} value significantly different from MUSHRA classic

^{b} value significantly different from MUSHRA simple

^{c} value significantly different from MUSHRA drag&drop

be explained by the computational implementation of this MUSHRA variant. The initial slider positions are on top of the scale, i.e. clicking a stimulus button for the first time leads to an initial rating of '100'. Starting the further ratings from this value is very likely to result in an anchoring bias (Tversky and Kahneman, 1974) in form of higher overall ratings compared to the other methods. Fortunately, this bias can be eliminated in future implementations of MUSHRA classic and for now be corrected by subtracting constant offset values from the ratings gathered with MUSHRA classic. Moreover, the found rating differences between the MUSHRA variants are small (maximally 6.6 point difference on a 100 point scale). The CQS consists of five semantic categories with a width of 20 points each. Applying a difference of 6.6 points to the mean values of each category does not lead to a change of rating categories.

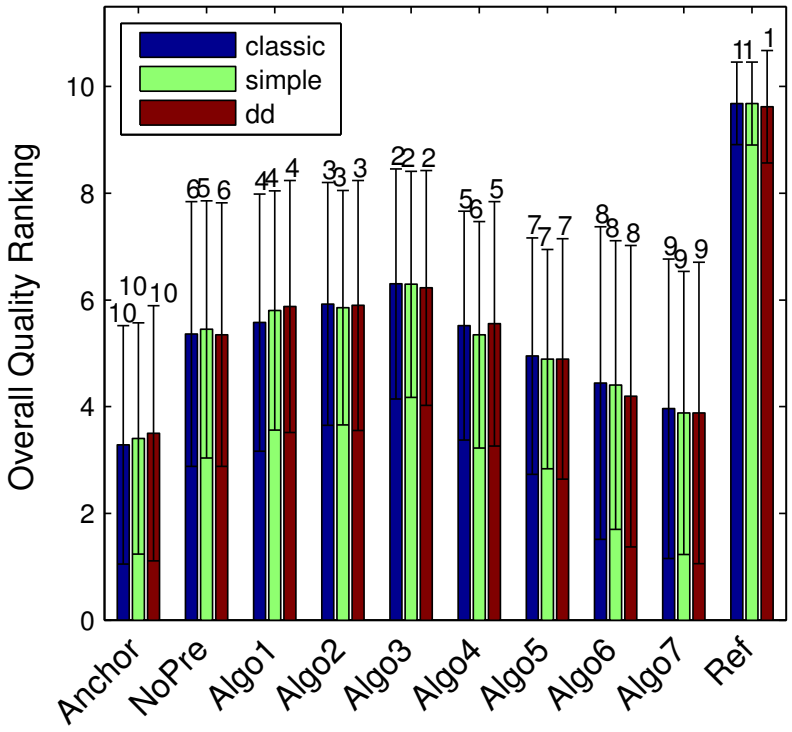

Figure 3.7: *Overall rank-transformed quality ratings averaged over all participants for the tested algorithms and methods (classic: blue, simple: green, drag&drop: red); the errorbars denote a range of ± 1 standard deviation from the mean value. Additionally the final ranks are plotted above the errorbars.*

Additionally, to compare final algorithm **rankings** gathered with each method, the quality ratings given by each participant were rank-transformed and averaged. The averaged algorithm rankings are displayed in Figure 3.7. MUSHRA classic and MUSHRA drag&drop produce the same final algorithm rankings. MUSHRA simple only differs within the middle field of the tested algorithms by swapping rank 5 and

6. It is possible to produce the same or at least nearly identical final quality rankings for the tested algorithms with our two modifications for the MUSHRA listening test as with the original test method. Summing up, the observed differences between the MUSHRA variants are explainable and can be corrected by systematically subtracting constants. Furthermore, the differences can be regarded as small as they are not affecting the algorithm rankings. Hence, the significant rating difference between the MUSHRA variants can be stated as negligible. Moreover, as we did not find a significant interaction of factor test method and between-subjects factor group, we conclude that all tested subject groups, independent of their age, hearing ability or technical commitment (or interaction), are able to produce similar and comparable ratings for the tested algorithms with all three tested MUSHRA variants. After ensuring that all MUSHRA variants lead to similar and comparable results, we discuss the differences between the variants regarding the subjective and objective comparison measurements with respect to participant factors like technical commitment and hearing impairment in the following.

3.5.2 Method comparison: Subjective measurements

Considering the system usability scale, a coherent trend becomes apparent that the averaged usability values from listeners with a high technical commitment (G1, G2, G4) are higher (Δ =4.7) than the values from listeners with a low technical commitment (G3, G5). This trend does not change for any of the implemented MUSHRA variants. Moreover, we find that all implemented MUSHRA variants can be used equally well by the participants from all tested subject groups. This observation holds for the listeners with a high technical commitment (G1, G2, G4); also we do not find any of the MUSHRA variants being unusable by listeners with a low technical commitment (G3, G5). However, only marginally missing statistical significance (p=0.07), the results show a trend for better SUS values for all used methods in the second test run. As the difference between test and retest is largest for MUSHRA simple and another marginally missing statistical significance for the interaction of test method and test run (p=0.07), this possible customization effect for the listeners seems to be most prominent when using MUSHRA simple. We suppose that a larger sample size of participants would strengthen this trend.

With the exception of group G5, we find a general trend of MUSHRA simple being the most preferred variant, followed by MUSHRA drag&drop (second place) and MUSHRA classic (third place). For the group G5, MUSHRA classic (second place) is preferred over MUSHRA drag&drop (third place). As the preference ranking from group G5 is not seen for G3, we can neither conclude that this behavior is due to a

low technical commitment of the participants, nor can we conclude that this is due to a hearing impairment of the listeners, because G4 also does not prefer MUSHRA classic over MUSHRA drag&drop — especially not in the retest. We did not find the participants from G5 commenting negatively about MUSHRA drag&drop. Some listeners from G5 rather noted that they would prefer all variants equally. We suppose that the observed slightly different final preference ranking from G5 can partly be explained by the forced distinct ranking that had to be given by the listeners.

3.5.3 Method comparison: Objective measurements

Regarding the objective comparison of the MUSHRA variants, we found that the participants rate the algorithms significantly faster with MUSHRA drag&drop (∼175 seconds per trial mean time) than using the other two MUSHRA versions (∼201 seconds, ∼206 seconds respective per trial). Extrapolating this small appearing difference (∼30 seconds) when having many trials in a longer test sessions, the difference might become crucial concerning the saving of time and costs. It might be possible to have a participant perform more trials with this MUSHRA variant with a smaller chance of a fatigue effect. As the participants using MUSHRA drag&drop need only one test session to become significantly faster in completing a test trial, we can state that MUSHRA drag&drop possesses a pronounced customization effect.

Analyzing further the found significant effect of factor group, following conclusions can be drawn: a) young listeners (G1) rate approximately 90s faster than elderly listeners (G2–G5), b) listeners with a high commitment for technology (G1, G2, G4) rate faster (roughly 60s) than listeners with a low commitment (G3, G5). The duration difference between young and elderly listeners can probably be explained by the general decline of psychometric test performance associated with advancing age (e.g., Binder et al., 1999). The difference with respect to technical commitment might be due to the fact that listeners with a high commitment for technology are likely more used to the appliance of computer programs and therefore are able to rate the algorithms faster. Moreover we can see from the results that this effect of low technology commitment leading to a longer test duration is more prominent for listeners with hearing impairment (G5) than for normal-hearing listeners (G3). We speculate that this behavior can be explained by a higher motivation of our hearing-impaired listeners to properly judge hearing aid algorithms, which might result in longer duration values per trial. This effect might then add up to the effect due to the low technology commitment. These general findings about the test trial duration concerning subject factors age, technical commitment, hearing ability, and the interaction of technical commitment and hearing ability do not interact with

the used test methods. For each subject group, MUSHRA drag&drop is the fastest method.

Next to the already mentioned fact that the participants click the reference button once more per trial with MUSHRA simple, we find that especially the hearing-impaired listeners (G4, G5) also click the reference button when using MUSHRA drag&drop (5.4) more often than using MUSHRA classic (4.6). This also could be due to the mentioned higher motivation of hearing-impaired listeners to rate hearing aid algorithms, which might become visible when these participants use the MUSHRA drag&drop variant. We also found MUSHRA drag&drop provoking the most overall stimuli button clicks per trial (25.2 clicks). This means that using MUSHRA drag&drop, the participants have the most switches between stimuli and still need the least amount of time when rating the algorithms within a trial.

Concerning the assessor performance metrics, we surprisingly did not find a significant interaction of the MUSHRA variants with between-subjects factor group affecting the assessor performance measures. In this regard, we could not support our hypothesis about a problematic use of MUSHRA classic for elderly and technically non-experienced listeners. Still, we found between-subjects factor group to be affecting the metrics for the discrimination ability. Separating the groups regarding factor age (G1: young, G2–G5: old), we find that the control group G1 shows higher discrimination values (2.02) than the elderly listeners (1.19). This finding is in line with the significant interaction of group and algorithm found in the analysis of algorithm ratings above: The participants from the control group G1 differentiate more in their stimuli ratings than the other elderly listeners from G2–G5.

We found one of our modifications MUSHRA simple showing the lowest values for both assessor metrics, reliability and discrimination ability. A possible explanation could be the reduction of the continuous scale (0...100) into eleven discrete stages (0,10,20,...,100). With only eleven possible answers, it might be harder to produce similar results within four test runs than having more degrees of freedom for the rating process. The eleven discrete stages might also be the reason for low metrics describing the discrimination ability: By rating ten algorithms with only eleven possible answers, it is harder for the listener to prove that he or she can discriminate between the stimuli. Although it was not mentioned by the participants as being problematic, the splitting of the test stimuli on two successive screens could make it more challenging to produce an overall coherent 'rating picture'. Despite these lowest values for both assessor metrics, the gathered final algorithm ratings and rankings between MUSHRA simple and the other two variants are still comparable (cf. Figure 3.5 and Figure 3.7).

Regarding the assessor performance data above the calculated noise floors, we found the highest amount of data with our modification MUSHRA drag&drop (82%), followed by MUSHRA classic (78%) and MUSHRA simple (62%). Separating the data for control group G1 and experimental group (consisting of G2–G5), we still find this ranking for both subgroups: With MUSHRA drag&drop, the participants produce the most data above noise floors (control group: 100%, experimental group: 77.5%), with MUSHRA simple the least amount of data (control group: 80%, experimental group: 57.5%). Separating the assessor performance data including noise floors for participants with high technology commitment (G1, G2, G4) and low technology commitment (G3, G5), we find a different ranking. Surprisingly, the participants with low technology commitment produce the most data above the noise floors when using MUSHRA classic (85%). With MUSHRA drag&drop, they produce 65%, with MUSHRA simple 55% data above the noise floors. However, participants with a high technology commitment produce the most amount of valuable data regarding the algorithm ratings with MUSHRA drag&drop (93.3%), followed by MUSHRA classic (73.3%) and MUSHRA simple (66.7%).

3.5.4 Differentiated choice of MUSHRA variants

So far we described on the one hand the equality between the MUSHRA variants regarding the algorithm rating results. We found only negligible differences in the collected data, which therefore leads to very similar and comparable results regarding the final algorithm ratings and rankings for all tested subject groups. This demonstrates the possible application of the introduced MUSHRA modifications, i.e. the modification have been shown to be bias-free with respect to the original MUSHRA version.

On the other hand we described significant differences between the variants regarding subjective and objective measurements. Both of these findings (equality and differences) give rise to a differentiated choice of listening test methods. According the aimed subject group and the experimenter's aims, one of the MUSHRA variants could be advantageous over the others and be most suitable. Also the available hardware (type of touchscreen) can play a role in choosing a proper test method.

By introducing two modifications of the classical MUSHRA test method, this study addresses the lack of alternative test procedures for hearing aid evaluation. Having now a selection of several listening test methods with the same underlying test method theory yielding comparable results, we recommend considering the efficient link of test subjects and test methods. As we found that technically experienced

listeners perform best with the modification MUSHRA drag&drop, we recommend using this variant over the classical MUSHRA method.

Against our hypothesis, elderly technically non-experienced hearing-impaired subjects (listener group G5) performed best with original MUSHRA. However, we found a trend for these participants to prefer MUSHRA simple over the original MUSHRA. Letting these listeners rate algorithms with their preferred MUSHRA simple might be beneficial.

We are interested in both, a) the further development of barrier-free variants for established evaluation tools, that can be accessed without challenges by non-expert listeners, and b) designing new evaluation methods.

Such a new evaluation method could aim at a maximization of the discrimination ability of the participants, which is quantified during the measurement and adaptively affects the evaluation procedure. This could yield a paired-comparison method (Kuk, 2002; Amlani and Schafer, 2009), including a new way for adaptive stimuli selection for comparisons. Also, a further modification of MUSHRA drag&drop could be used as a patient-centered hearing aid fitting tool.

3.6 Conclusion

In this study we introduced and evaluated two modifications of the classical MUSHRA listening test method. The modifications of the standardized test method aim to increase the intuitiveness and accessibility for different target listener groups, including elder and technically non-experienced people with hearing loss, who constitute the typical target group in hearing aid research.

Overall the three tested MUSHRA implementations led to equivalent ratings regarding the performed hearing aid evaluation, reflecting the interchangeability of the tested user interfaces. Thus, the compatibility of the new modifications combined with the advantages that MUSHRA drag&drop demonstrated within the performed comprehensive comparison of the MUSHRA variants, the intuitive user interface MUSHRA drag&drop can be regarded as the general recommendation.

However, if considering a specific target group or a specific demand regarding a tested measure or both, the one or the other MUSHRA implementation may provide considerable advantages. Hence, in these cases we suggest a differentiated and careful application of listening test methods, e.g. if testing a group of elderly, technically non-experienced persons with impaired hearing.

4 Development and evaluation of an unsupervised model-based multidimensional parameter optimization for hearing aid algorithms

ABSTRACT

Objective: A model-based approach extracting optimized parameter settings for hearing aid algorithms with a multidimensional parameter space was introduced and evaluated. The approach might be generalized to any algorithm by combining objective performance measures. *Design:* The settings of a coherence-based noise reduction scheme were optimized by equally weighing four objective measures. The outcome of the parameter optimization was compared to two manual algorithm adjustments by experts. The adequacy of each setting was compared with an extensive set of instrumental and perceptual measures, i.e. speech reception thresholds (SRTs) and subjective attributes (listening effort, localization, naturalness, preference), which were tested within specifically designed scenarios for each attribute. *Study Sample:* 31 hearing-impaired listeners with moderate sensorineural hearing loss participated in the evaluation. *Results:* Within the instrumental evaluation, the algorithm performed best when adjusted by the model-based optimization routine. The subjective assessment indicated that both expert-driven algorithm adjustments led to more adequate settings. *Conclusions:* The optimization approach did not reach the subjective performance of expert-driven adjustments to the chosen algorithm. Further research must show if using a different weighting and different measures can

The content of this chapter was submitted as an Original Paper to the International Journal of Audiology in 05/2016.

improve the introduced approach. Using attribute-specific noise scenarios proved to be useful to determine the algorithm benefit within controlled laboratory measurements.

4.1 Introduction

4.1.1 Multidimensional hearing aid fitting

One major goal of modern digital hearing aids is to restore the ability to communicate in order to help hearing impaired listeners to fully participate in our society. To achieve this goal, the majority of algorithms in modern digital hearing aids are in general controlled by more than one parameter influencing the signal processing. These variable parameters lead to a multitude of possibilities for adjusting the respective algorithms, which offers the opportunity to fit the algorithms to each individual patient. However, the question remains how to effectively solve the multidimensional problem of finding the optimal fitting for each patient in all kinds of different listening situations. To tackle this question, both, algorithm developers and clinicians offer expert-based adjustments of the respective algorithm parameters: In hearing aid development, optimal algorithm parameters are commonly determined and evaluated by auditory-model based instrumental performance measures as e.g., the 'Hearing-Aid Speech Quality Index' (HASQI; Kates and Arehart, 2014), the standardized 'Perceptual Evaluation of Speech Quality' (PESQ; ITU-T — Telecommunication Standardization Sector of ITU, 2001), and the quality measure 'PEMO-Q' (Huber and Kollmeier, 2006). Also purely technical distance measures as the signal-to-noise ratio (SNR) are used, or the optimization is heuristically based on informal listening by the algorithm developers (Rohdenburg et al., 2006). In clinical practice, the crucial hearing aid fitting task, in which the multidimensional hearing aid algorithms are matched to the individual hearing aid users, is performed by means of the commercial fitting software provided by the respective hearing aid manufacturer in combination with the expertise of the audiologist. Prescriptive fittings are performed which incorporate e.g. the audiometric hearing thresholds or suprathreshold measures like loudness discomfort levels of the patient. Examples for prescriptive fittings are NAL-NL1 (Byrne et al., 2001), DSL [i/o] (Cornelisse et al., 1995), and LGOB (Allen et al., 1990). A detailed overview of various fitting strategies is given by Dillon (2012).

Evidence has shown that the prescriptive initial fitting is rarely the best fit for the individual patient. Leijon et al. (1984) found that prescriptive fitting methods overestimated the preferred gains of a group of moderately hearing-impaired, elderly, untrained hearing aid users by about 10 dB. Keidser and Dillon (2006) investigated

preferred gains relative to NAL-NL1 prescribed gains of 189 hearing-impaired people participating in five different research studies. While 49% perceived the prescribed gains as just right (±3 dB), 46% found them too loud and preferred less gain. In the comprehensive review article by Wong (2011) regarding the self-fitting of hearing aids, it is stated that in general, prescribed gains have been found to be slightly higher than the average preferred hearing aid responses. Jenstad et al. (2003) state that a prescriptive fitting will not guarantee patient satisfaction, particularly with the many available parameters on modern hearing aids for which values are not specified by fitting formulas. For example, the speed of automatic noise reduction parameter is not individually prescribed at all (Dillon et al., 2006).

Most clinicians agree that the first prescribed setting serves as a reasonable starting point for a subsequent fine-tuning of the hearing aids, which can be regarded as the second stage of the fitting procedure. Within the commercial fitting software, the patient's complaints are linked with a parameter adjustment that should resolve these complaints (Jenstad et al., 2003). Boymans and Dreschler (2012) mention drawbacks for the complaint-driven approach, e.g. that the individual fine-tuning process is not well defined. Amlani and Schafer (2009) also notice a general lack of procedural guidelines necessary to fit many of the parameters in hearing aids. Altogether, the crucial fitting process can be very cost-intensive and time-consuming, e.g. incorporating many repeated visits at the audiologist (e.g., Boymans and Dreschler, 2012; Abrams et al., 2011).

To support the professionals (e.g. audiologists or algorithm developers) in determining the optimal hearing aid parameter settings for the individual patient within a reasonable amount of time, an approach to perform an unsupervised extraction of optimized parameter combinations for complex hearing aid algorithms was developed within this study. The main concept of the introduced optimization approach follows a systematic dimensionality reduction of possible algorithm settings. Therefore, the results of various instrumental performance measures are combined to estimate those parameter combinations (for the respective algorithm), which are the most promising candidates to be tested with real patients. The approach allows a use of pure technical measures like the signal-to-noise ratio (SNR) and auditory-model based performance measures that are complimentary to each other.

The optimization approach was exemplary applied to a recently developed binaural coherence-based noise reduction algorithm (Grimm et al., 2009; Luts et al., 2010; Baumgärtel et al., 2015b). One optimized parameter combination without the direct supervision of a professional was extracted. To evaluate the outcome of this unsupervised parameter optimization, it was compared with two expert-based

adjustments of the algorithm parameter settings. For the subjective evaluation with hearing impaired, a common multiband compressor scheme was used to compensate for the individual hearing loss of the listeners. Insertion gains were calculated using the NAL-RP fitting rule (Byrne et al., 1991).

4.1.2 Real-world benefit of hearing devices

The evaluation procedures for assessing hearing aid performance and benefit represent a vast and almost independent realm (Kiessling et al., 2006). One major evaluation objective of interest is to assess the real-world benefit of hearing aids, which, in research, can be captured within field studies (see e.g., Bentler et al., 2008). Discrepancies between the results of different evaluation objectives were found, namely a) between computer-based instrumental and perceptual measures (see e.g., Luts et al., 2010), b) between objective human measures and their subjective ratings (see e.g., Luts et al., 2010; Marzinzik and Kollmeier, 1999; Walden et al., 2000), and c) between laboratory and field trial results (see e.g., Bentler, 2005). From this, it can be concluded that a narrow evaluation can be misleading.

We conducted instrumental and perceptual measurements to characterize the effects of the different settings of the noise reduction algorithm as precise as possible. Perceptual measurements comprised objective speech reception threshold (SRT) measurements and four subjective attributes, i.e. listening effort, localization, naturalness, and preference. Realistic attribute-specific listening scenarios were designed for each tested subjective attribute. The aim of designing these scenarios was to approach the requested real-world benefit within controlled laboratory measurements as closely as possible.

4.2 Algorithm Parameters

4.2.1 Coherence-based noise reduction algorithm

The noise reduction scheme (cf., Grimm et al., 2009; Luts et al., 2010; Baumgärtel et al., 2015b) utilizes the concept of coherence, i.e., the similarity between the signals captured at the left and right ear, to separate the desired speech signal from undesired noisy components. Therefore it is assumed that coherent signal components belong to the desired target signal, e.g., a single speaker talking to the listener, and incoherent signal components belong to the undesired noisy background. A block diagram illustrating the coherence-based noise reduction algorithm is presented in Figure 4.1. The parameter optimization from this study yield onto the adjustment of the two

main algorithm parameters τ and α, which are printed in bold letters within the Figure.

The algorithm works in the short-time Fourier Transform (STFT) domain, where STFT bins are grouped into 15 non-overlapping third-octave frequency bands k with center frequencies ranging from 250 Hz to 8 kHz. The interaural phase difference (IPD) is used as an estimate for the coherence. The coherence $C(k,l)$ in each frequency band k and time segment l is estimated from the vector strength of the complex IPD $c_{\mathrm{IPD}}(k,l)$, as defined in Grimm et al. (2009):

$$C(k,l) = \left|\langle c_{\mathrm{IPD}}(k,l)\rangle_{\tau}\right|. \tag{4.1}$$

The coherence value is estimated using a running average $\langle\cdot\rangle_{\tau}$ with time constant τ. After a linear mapping of the coherence estimates to the interval $[0,1]$, the gain in each frequency band is computed by applying an efficiency exponent α, i.e.:

$$G(k,l) = \widehat{C}(k,l)^{\alpha}. \tag{4.2}$$

Thus the two independent parameters τ and α determine the overall performance of the suggested algorithm and hence build the subject in the optimization process.

4.2.2 Manually optimized parameter settings

The first set of parameters for the algorithm uses a time constant that is short enough to follow the modulations of speech ($\tau = 40$ ms), so that the target speech signal is not significantly degraded (Wittkop and Hohmann, 2003). For the second parameter, the efficiency exponent α, it has to be taken into account that high values provide efficient filtering, but also lead to more audible artifacts. Hence a balancing value of $\alpha = 1$ is used (Luts et al., 2010).

The second setting of the algorithm parameters was based on evidence, which has shown that different frequency areas have a different relevance for speech intelligibility. Thus, the filter efficiency was maximized in those frequency areas which are relevant for speech intelligibility, and the filter settings were optimized for maximum quality outside those frequency areas. Therefore, both main parameters were optimized in terms of a frequency-dependent setting leading to $\alpha(k)$ and $\tau(k)$, where k denotes the center frequency of the k^{th} frequency band. The efficiency exponent $\alpha(k)$ was adapted to roughly follow the band importance function suggested for the calculation of the speech intelligibility index (SII; ANSI 20S3.5-1997, 1997). The coherence estimation (see Equation 4.1) was already introduced by Dietz et al. (2011) as input for estimating the direction of arrival of concurrent speakers within binaural

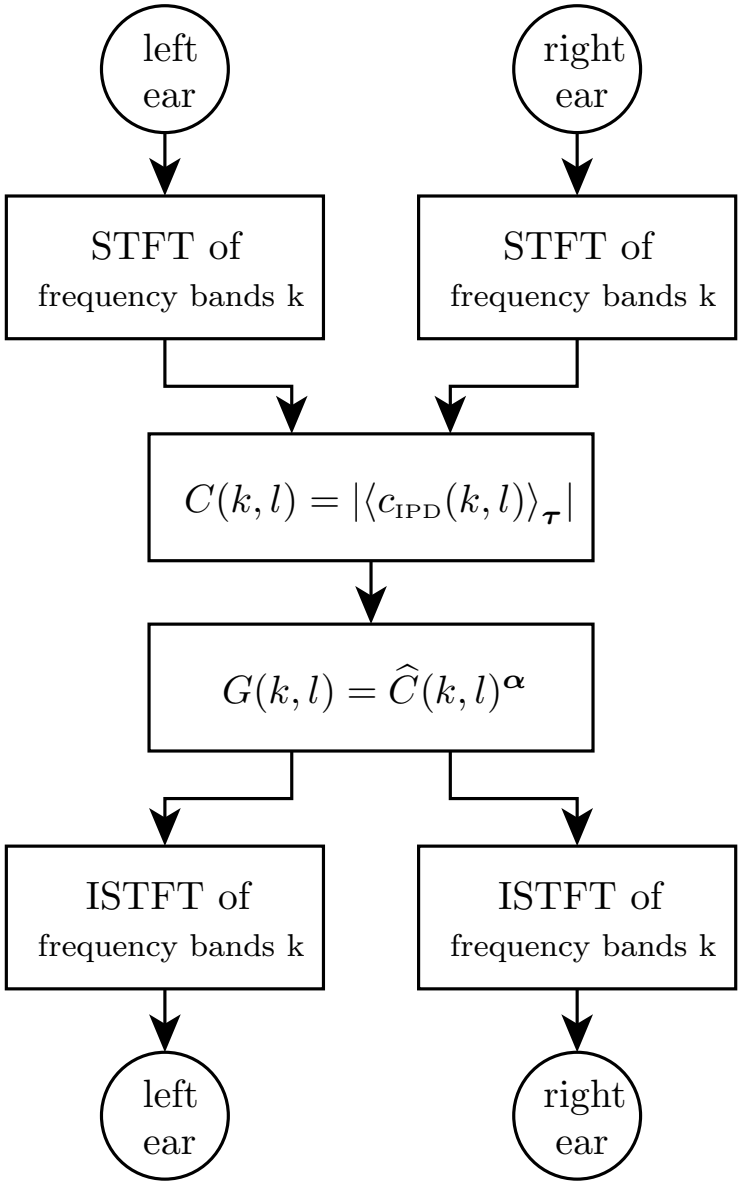

Figure 4.1: *Block diagram illustrating the coherence-based noise reduction algorithm. The optimized parameters τ and α are printed in bold letters.*

signals. In their study, they suggest a frequency-dependent smoothing time constant of $\tau(k) = \frac{5}{f_k}$). Since this choice led to an agreeable setting of the time constant with only few audible processing artifacts as confirmed by informal listening by the algorithm developer, it was used in this study as second setting of the algorithm parameters.

The frequency-dependent values for algorithm parameters τ and α are displayed in Figure 4.2.

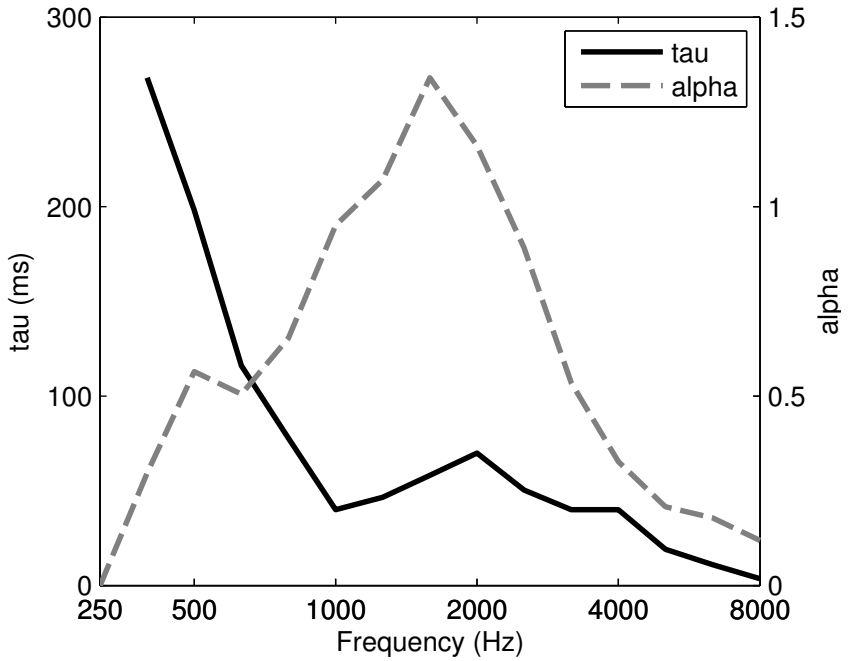

Figure 4.2: *Manual optimization II: Frequency-dependent values for algorithm parameters τ and α.*

4.2.3 Automatically optimized parameter setting

The automated parameter optimization strategy combines different instrumental measures and extracts the optimal setting from the minimal intersection between the calculated iso-performance contours within the multidimensional parameter space. The optimization strategy is a further development of the procedure described by Völker et al. (2013).

Regarding signal material, a sentence from the Oldenburger Satztest (OLSA; Wagener et al., 1999c,a,b) was taken as speech component. Considering possible transient states by the algorithm, short interval of silence before and after the sentence was added. The OLSA sentence was convolved with a binaural room impulse response (BRIR). The BRIR was recorded with the front microphones of a pair of behind-the-ear (BTE) hearing aid shells mounted to an artificial head in a cafeteria ambience (Kayser et al., 2009). The noise component consisted of an extract from a recording of ambient cafeteria sounds. The extract was selected manually such that the noise level was constant and no singular crockery rattle sounds with high peak amplitudes occurred. This binaural ambient recording (Kayser et al., 2009) was part of the database of the used binaural room impulse responses. The same microphone setup as for the BRIR was used for the recording (front microphones of BTE hearing aids). The components were mixed together at a global SNR of

0 dB (averaged over both channels). The total stimulus length of mixed speech- and noise-components was 3.5 seconds.

Regarding a reasonable definition of the two-dimensional parameter space, exponential functions were used to generate combinations for α (ranging from 0 to 15) and τ (ranging from 0 to 1000 ms) with non-equidistant spacing in the parameter space. The signal material was systematically processed by the noise reduction algorithm using the defined combinations of parameter settings, which results in 42 x 32 = 1344 processed stimuli.

The used measures within this study consisted of one purely technical measure, i.e. the signal-to-noise ratio (SNR) and three auditory model-based characterizations, reflecting two major factors, i.e. speech intelligibility and overall quality:

- signal-to-noise ratio (SNR),
- speech intelligibility index (SII; ANSI 20S3.5-1997, 1997),
- binaural speech intelligibility index (BINSII; Beutelmann and Brand, 2006), and
- perceived similarity measure to a (high quality) reference signal (PSM; Huber and Kollmeier, 2006).

The shadow filtering method realized within the MHA (Grimm et al., 2006) was used to measure the processed speech and noise components separately, which is needed for calculations of the SNR, SII, and BINSII predictions. For monaural measures SNR and SII, the mean values between both channels (left and right) were calculated. Regarding PSM, unprocessed speech and noise components were mixed together at an SNR of +20 dB serving as high quality reference signal. As PSM is a solely monaural measure, the binaural reference and test signals were mixed down to mono for the calculations.

Percentile contours within the algorithm parameter space were calculated by the automated parameter optimization procedure for each of the used performance measures. Beginning with calculating 0% percentiles, the percentile values were uniformly increased until a minimal intersection between the iso-performance contours was formed, i.e. the intersection contained only one unique optimized setting. Figure 4.3 shows iso-instrumental-measure contours of the used performance measures for the automated optimization of the algorithm parameters τ and α. The iso-performance contours were calculated by the 89% percentiles, which led to a minimal intersection containing one unique optimized setting (α = 3.87, τ = 39.8 ms). Together with the setting resulting from the first manual parameter optimization,

the unsupervised optimized setting is marked in the figure. It can be seen that the automated optimized setting involves a shift for parameter α to a higher value. However, the value for parameter τ stays almost unchanged with respect to the manually optimized setting I.

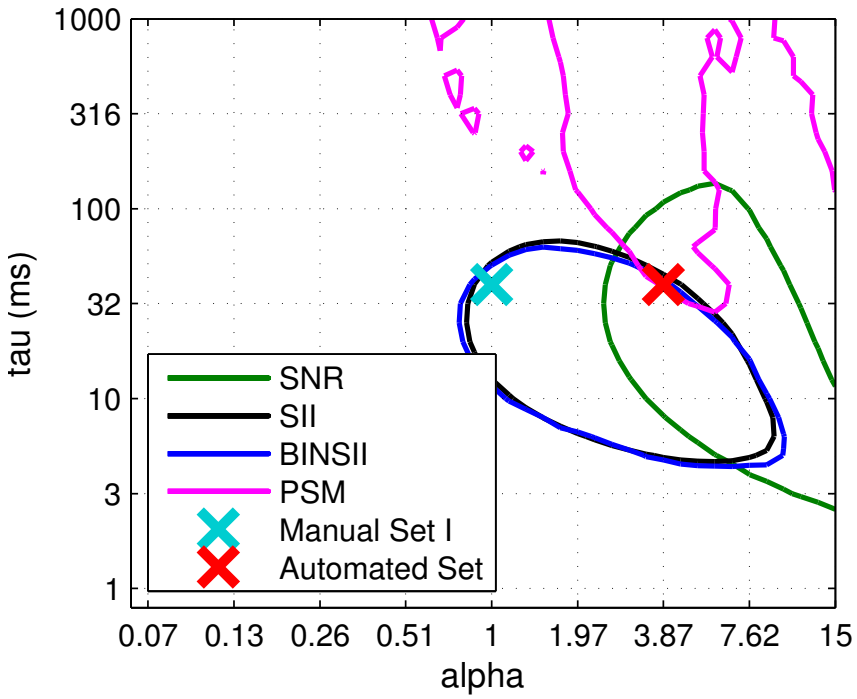

Figure 4.3: *Automatic parameter optimization: Iso-instrumental-measure contours of four performance measures given as 89% percentiles over the evaluated algorithm parameters τ and α for the coherence-based noise reduction algorithm. The two crosses mark the manual optimized setting I (blue) and the automated optimized setting (red), which was found by the only common point of the highest iso-contours for the four instrumental measures.*

4.2.4 Overview of different parameter settings

An overview of the three different parameter settings for the coherence-based noise reduction algorithm evaluated in this study is given in Table 4.1.

Table 4.1: *Three different parameter settings for the coherence-based noise reduction algorithm.*

Setting	Parameter α	Parameter τ
Manual optimization I	1	40 ms
Manual optimization II	frequency-dependent, cf. Figure 4.2	
Automatic parameter optimization	3.87	39.8 ms

4.3 Method

4.3.1 Participants

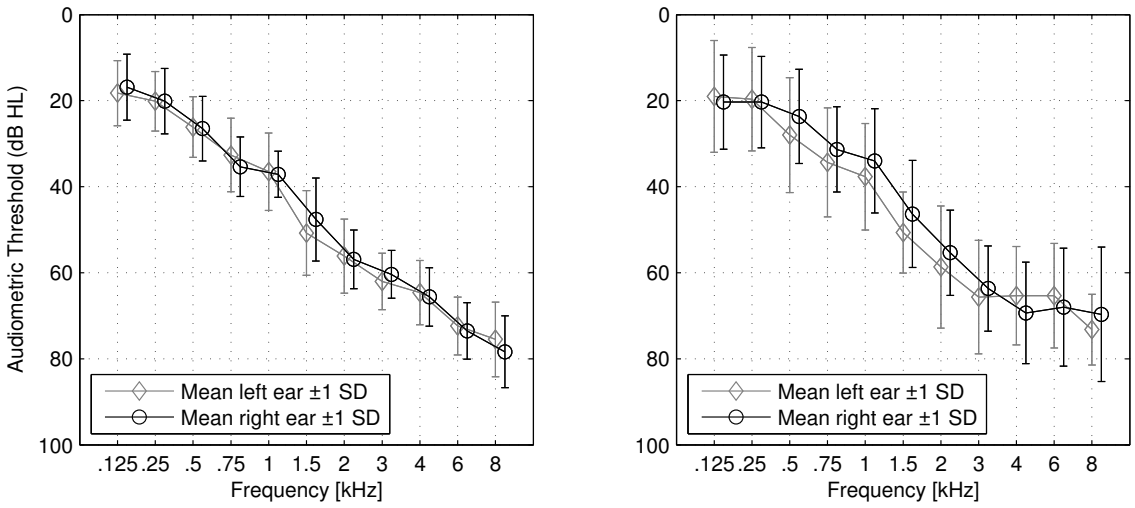

Figure 4.4: *Averaged hearing thresholds ± standard deviation for 31 participants are displayed separately for both ears (marked as diamonds for the left ear, marked as circles for the right ear). The left panel shows the data of 16 listeners from test center I, right panel shows 15 participants from test center II.*

In total, 31 hearing-impaired listeners with moderate bilateral sloping sensorineural hearing loss from two test centers participated in the study. The listeners were aged from 60 to 82 years (mean: 71.9 years). Audiometric data for the participants are shown in Figure 4.4. The averaged hearing thresholds ± standard deviation are displayed separately for both ears (marked as diamonds for the left ear, marked as circles for the right ear). The left panel shows the data of 16 listeners from test center I, right panel shows 15 participants from test center II. For test center I, the averaged thresholds over the frequencies 500, 1000, 2000, and 4000 Hz (4PTA) at the better ear range from 35.5 dB to 51.5 dB (mean: 44 dB). For test center II, the 4PTA values range from 33.75 dB to 57.5 dB (mean: 44 dB). For all but one listener from test center I (Δ_{4PTA}=12 dB) and two listeners from test center II (Δ_{4PTA}=14 dB and Δ_{4PTA}=18 dB), the absolute differences between 4PTA at both ears is below 8 dB HL, which indicates symmetrical hearing losses for the majority of the participants.

The test protocol for the participants included audiometric measurements and paired comparisons regarding different subjective attributes on the first test day. On the second test day, the subjective attributes within paired comparisons were re-tested, and speech reception threshold (SRT) measurements were performed. The listeners were remunerated for their participation and their travel expenses. Approval

for this experiment was obtained from the Carl von Ossietzky Universität Oldenburg ethics committee on 08.01.2009 and 06.09.2013.

4.3.2 Compensation of hearing loss

To compensate for the individual hearing loss of the listeners, a common multiband compressor scheme was used. Together with the evaluated noise reduction scheme from this study, the compressor was implemented in the master hearing aid (MHA; Grimm et al., 2006). The compressor divides the left and right input signals into nine overlapping filter bands and measures the sound pressure levels in each band. Dependent on the actual input band levels and the individual hearing threshold of the listener, insertion gains were calculated using the NAL-RP fitting rule (Byrne et al., 1991). The obtained first fit of the procedure was used. The multiband compressor applied insertion gains serially after the pre-processing noise reduction. The used time constants for attack and release were 20 ms and 100 ms, respectively.

For the measurements, the participants were equipped with custom-made behind-the-ear hearing aid dummies Acuris P by Siemens for the research project 'Model-based Hearing Aids'. The two-channel output of the hearing aids was passed into the participants' ears by using the ear plugs E-A-RTONE 13A.

4.3.3 Speech reception threshold measurements

The evaluation of the three different settings for the coherence-based noise reduction scheme comprised adaptive measurements of speech reception thresholds. The sentences from the Oldenburger Satztest (OLSA; Wagener et al., 1999c,a,b) were used as signal material. As diffuse noise material, the cafeteria recording of the database with head-related impulse responses (Kayser et al., 2009) was used. The noise level was kept constant at 65 dB SPL, the speech level varied during the adaptive SRT measurements (Brand and Kollmeier, 2002), which converged on the 50% intelligibility level. Regarding the test setup, the listeners were seated in a soundproofed test room surrounded by eight equidistant speakers. The 8-channel noise sound was emitted from all speakers, the speech sound was additionally emitted from the speaker at +45 degrees (or -45 degrees) facing the better ear (with respect to 4PTA) of the listener. For each of the four algorithm conditions, a random list of 20 sentences was used to determine the SRT. The order of the algorithms was randomized over the participants. The participants were instructed to repeat the understood words after each sentence presentation, whereupon the experimenter entered the correctly understood words into the measurement computer. As training,

measurements with two lists of sentences were performed in the condition without noise reduction (NoPre).

4.3.4 Subjective attributes

In addition to the SRT measurements, the three different settings for the coherence-based noise reduction scheme and the NoPre condition were also evaluated with respect to subjective attributes listening effort, sound localization, naturalness, and preference. For each attribute, the participants performed a full pairwise comparison ('round-robin tournament') of the four different conditions (i.e., six comparisons), where the listeners chose the winning scheme in each comparison responding to the following questions:

- With which hearing aid is it easier to follow the conversation ('listening effort')?
- With which hearing aid is it easier to locate the ringing telephone / the conversation partners ('sound localization')?
- With which hearing aid does the ambiance sound more natural ('naturalness')?
- Which hearing aid would you prefer over the other ('preference')?

For the evaluation of all tested attributes, a realistic listening scenario consisting of a conversation in a crowded cafeteria was generated. In addition to the cafeteria scenario, four further scenarios were designed especially for testing each single attribute in isolation. Regarding listening effort, a conversation scenario within a home kitchen was generated. A scenario in a supermarket comprising of a conversation between customer and cashier at the cash register and an additional moving ringing mobile phone served as a setting for the localization task. Regarding naturalness, a nature wood land scene with sounds of e.g. birds and flowing water was created. The preference for one or the other noise reduction scheme is also retrieved in a jazz club scenario consisting of music played by a band and typical background noise. All scenarios were calibrated to exhibit realistic and lifelike overall sound pressure levels. The levels were 74.5 dB (cafeteria), 67.9 dB (kitchen), 72.5 dB (supermarket), 49.5 dB (nature scene), and 81.7 dB (jazz club), respectively. In scenarios including a conversation (cafeteria, kitchen, supermarket), realistic signal-to-noise ratios were used.

The scenarios, with a duration of 20 seconds, were generated using the toolbox for acoustic scene creation and rendering (TASCAR; Grimm et al., 2013, 2014). The rendered 8-channel sound files were presented loop-wise from the eight equidistant

speakers to the listener. During the pairwise comparisons of the schemes, the listener could switch for an unlimited amount of times nearly instantaneously between the alternatives and decide for a winning scheme of the actual comparison. Both, the order of the tested attributes and the order of the six algorithm comparisons within each attribute were randomized for each participant.

4.3.5 Instrumental evaluation

To allow for an additional objective comparison of the three different settings for the coherence-based noise reduction scheme (cf. Table 4.1), the performance of the settings was also assessed in terms of an instrumental evaluation. This evaluation involved measurements with respect to the same four performance measures (SNR, SII, BINSII, and PSM), which were used by the introduced approach described above. For this, the used signal material for speech and noise for the automated optimization procedure was mixed together at different SNRs ranging from -10 to 10 dB. These mixtures were processed by the noise reduction algorithm with the three different settings and evaluated by the performance measures. For each parameter setting (Manual Set I, Manual Set II, Automated Set), the enhancements regarding the no pre-processing condition (NoPre) was calculated.

4.3.6 Statistical analyses

The collected data regarding SRT measurements and paired comparisons were analyzed using the statistical software IBM SPSS. To investigate the significance of within-subjects factor *algorithm setting*, between-subjects factor *test center* and their interaction on listeners SRT performance, a mixed-model analysis of variance (ANOVA) with repeated measurements was employed. Normality distribution of each data set was verified with the Kolmogorov–Smirnov test. Levene's test was used to assure the homogeneity of variances. Whenever necessary, violations of sphericity were adjusted using the Greenhouse-Geisser correction. In order to determine the sources of significant effects indicated by the ANOVA, post hoc tests for multiple comparisons were conducted and reported in terms of Bonferroni adjusted significance levels, p_{Bonf}.

The effects of within-subjects factor *algorithm setting* onto the requested subjective attributes within the paired comparisons were examined with non-parametric tests (Friedman's ANOVA). In case of significant effects, Wilcoxon signed-rank tests with Bonferroni adjusted criteria for significance ($\alpha/6 = 0.0083$) were performed as post hoc tests.

4.4 Results

4.4.1 Instrumental evaluation

The results of the instrumental evaluation of the three different parameter settings (Manual Set I, Manual Set II, Automated Set) for the coherence-based noise reduction scheme are shown in Figure 4.5. The enhancements with respect to the NoPre condition are displayed for the performance measures SNR (top left), SII (top right), BINSII (bottom left), and PSM (bottom right) over the input SNRs from -10 dB to 10 dB.

Regarding the performance measure SNR (Figure 4.5, top left), it can be seen that all algorithm settings led to an enhancement of SNR, which behaves nearly constant over the tested input SNRs for all settings. The SNR was maximally enhanced by 6.3 dB by Automated Set at an input SNR of 5 dB. The averaged enhancements and standard deviations were 4.5±0.13 dB (Manual Set I), 3.1±0.14 dB (Manual Set II), and 6.1±0.18 dB (Automated Set). Automated Set outperformed Manual Set I by 1.6 dB and Manual Set II by 3 dB on average.

The enhancement of SII due to the processing strategies is shown in the upper right panel of Figure 4.5. Similar to the SNR performance measure, none of the settings led to negative values, i.e., no reduction of speech intelligibility was introduced by the coherence-based noise reduction algorithm. In the range from -10 to -4 dB all settings showed no difference with respect to the NoPre condition and therefore do not enhance SII. Starting from approximately -4 dB, SII was enhanced by all parameter settings. Depending on the parameter setting, the SII benefits rose with increasing input SNR. At the input SNR of 10 dB, Manual Set II reached an SII increase of 0.07, Manual Set I an increase of 0.1, and Automated Set an SII enhancement of 0.13. Automated Set outperformed Manual Set II for SNRs above -4 dB and Manual Set I for SNRs above 0 dB.

With the exception that Manual Set I introduced very small reductions of binaural speech intelligibility in the SNR range from -10 to -5 dB, the enhancements with respect to BINSII (Figure 4.5, bottom left) behaved very similar to the SII enhancement. Beginning from the SNR of -4 dB, all parameter settings improved BINSII. The improvements rise with input SNR, differently for each setting. Automated Set outperformed Manual Set II for SNRs above -4 dB and Manual Set I for SNRs above 0 dB.

Regarding PSM (bottom right panel), it can be seen that PSM was enhanced by Manual Set I and Manual Set II over the whole tested SNR range. Automated Set improved PSM over a wide SNR range and led to a degradation of PSM at SNRs

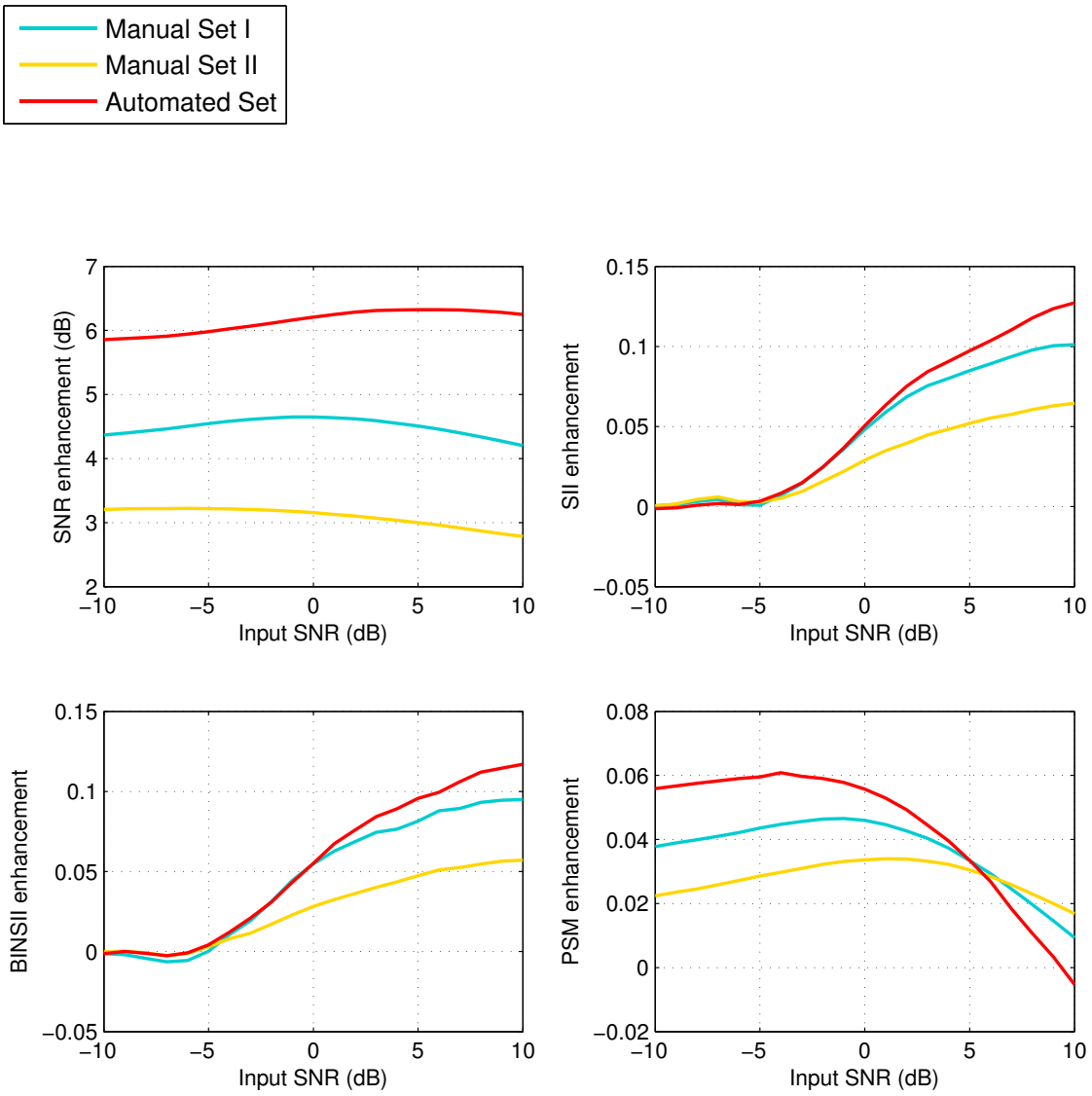

Figure 4.5: *Enhancement of instrumental measures SNR (top left), SII (top right), BINSII (bottom left), and PSM (bottom right) regarding three settings of the coherence-based noise reduction algorithm (Manual Set I, Manual Set II, Automated Set) with respect to the no pre-processing condition (NoPre) over input SNRs from -10 to 10 dB.*

above 9 dB. In the range from -15 to 5 dB, Automated Set produced the largest PSM enhancement (mean = 0.05), followed by Manual Set I (mean = 0.04) and Manual Set II (mean = 0.03). For SNRs above 5 dB, Manual Set II performed best, followed by Manual Set I and Automated Set.

4.4.2 Perceptual evaluation

4.4.2.1 Speech reception thresholds

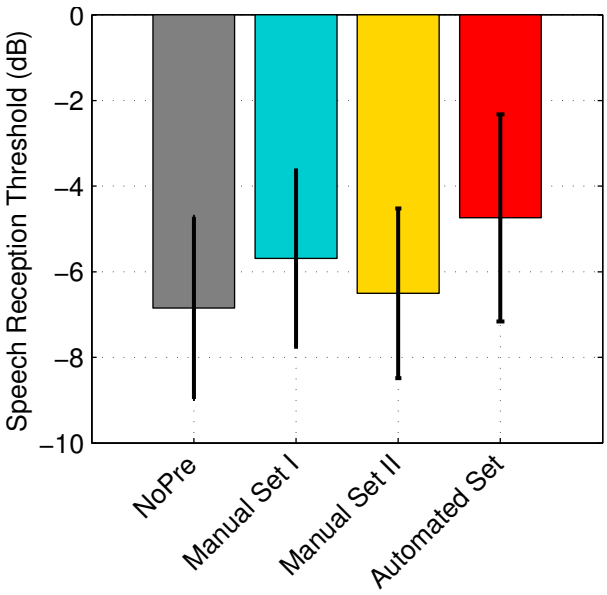

Figure 4.6: *Averaged individual SRTs of 31 participants from two test centers using three different settings of the coherence-based noise reduction algorithm (Manual Set I, Manual Set II, Automated Set) and a no pre-processing scheme (NoPre) as a reference.*

The statistical analysis supports that the between-subjects factor *test center* is not affecting the measured SRTs ($p > 0.6$). This allowed us to pool the results of the participants from both test centers in the following. In contrast, the within-subjects factor *algorithm setting* showed a significant effect, $F(3,87) = 26.8$, $p < 0.001$. The averaged SRTs of the 31 listeners using the three different settings of the coherence-based noise reduction algorithm and a no pre-processing scheme (NoPre) as a reference are displayed in Figure 4.6. The mean thresholds were -6.8 dB (NoPre), -5.7 dB (Manual Set I), -6.5 dB (Manual Set II), and -4.7 dB (Automated Set).

Post hoc analysis shows that the SRT differences regarding NoPre were significant for Manual Set I ($|\Delta_{SRT}| = 1.1$ dB, $p_{Bonf} < 0.001$) and Automated Set ($|\Delta_{SRT}| = 2.1$ dB, $p_{Bonf} < 0.001$). The thresholds using NoPre and Manual Set II did not differ significantly ($p_{Bonf} > 0.8$). The SRT differences regarding Manual Set I were significant with $p_{Bonf} = 0.01$ for Automated Set ($|\Delta_{SRT}| = 1.0$ dB) and Manual Set II ($|\Delta_{SRT}| = 0.8$ dB, $p_{Bonf} = 0.005$). Furthermore, the difference between Manual Set I and Automated Set ($|\Delta_{SRT}| = 1.8$ dB) was significant with $p_{Bonf} < 0.001$. Based on these comparisons, we built a statistically significance ranking of the different algorithm conditions: NoPre and Manual Set II share the first place with the lowest

thresholds, followed by Manual Set I on the second place and Automated Set on the third place with the highest SRT value.

4.4.2.2 Subjective attributes

Given the 31 participants from both test centers, who were tested twice (test and retest), we collected 62 data sets for each of the requested attributes in the different scenarios within full pairwise comparisons. Inconsistencies in the judgments due to circular triads were detected, which were neither found to behave systematically for a set of listeners nor a certain questioned attribute nor only occurring within a specific noise scenario. In the following, we excluded the judgments containing either one or two circular triads, as no definite rankings could be calculated from them.

Figure 4.7 shows the amount of relative wins of each pre-processing strategy (NoPre, Manual Set I, Manual Set II, Automated Set) regarding the questioned subjective attributes listening effort, sound localization, naturalness, and preference within the different sound scenarios. The definite rankings collected from 31 participants from two test centers are displayed (test and retest). For each of the eight attribute-scenario pairs the amount of consistent data sets N is also given in the Figure. Significance strength for each test condition is denoted next to their labels (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Significant differences between the algorithm settings within each test condition are denoted within the Figure (*: $p < 0.0083$).

The subjective attribute listening effort (in the cafeteria scenario) held 33 consistent data sets (of 62, i.e. 53.2%). Statistical analysis showed that within-subjects factor *algorithm setting* did not affect the judgment ratings, $\chi^2(3) = 1.5$, $p > 0.6$. Listening effort in the kitchen scenario exhibited 67.7% consistent data sets, demonstrating a significant effect of factor *algorithm setting*, $\chi^2(3) = 61.9$, $p < 0.001$. Multiple comparisons showed that NoPre had significantly more relative wins (82.5%) than all other three settings (all $p < 0.0083$). Furthermore, Manual Set I and Manual Set II had more wins than Automated Set (both with $p < 0.0083$). No differences between Manual Set I and Manual Set II were found ($p > 0.5$). The attribute sound localization (within the cafeteria scenario) with $N = 27$ had the lowest amount of consistent data sets (43.5%) of the eight attribute-scenario pairs. Still, factor *algorithm setting* had a significant effect on the ratings, $\chi^2(3) = 9.4$, $p < 0.05$, which was provoked by the comparison between Manual Set I and Automated Set: Manual Set I (69.1%) had significant more wins than Automated Set (33.3%), $p < 0.0083$. The other five comparisons showed no differences. Within the supermarket scenario, the amount of consistent data was $N = 46$ (74.2%). The different algorithm settings significantly influenced the sound localization ratings, $\chi^2(3) = 58.8$, $p < 0.001$. Here,

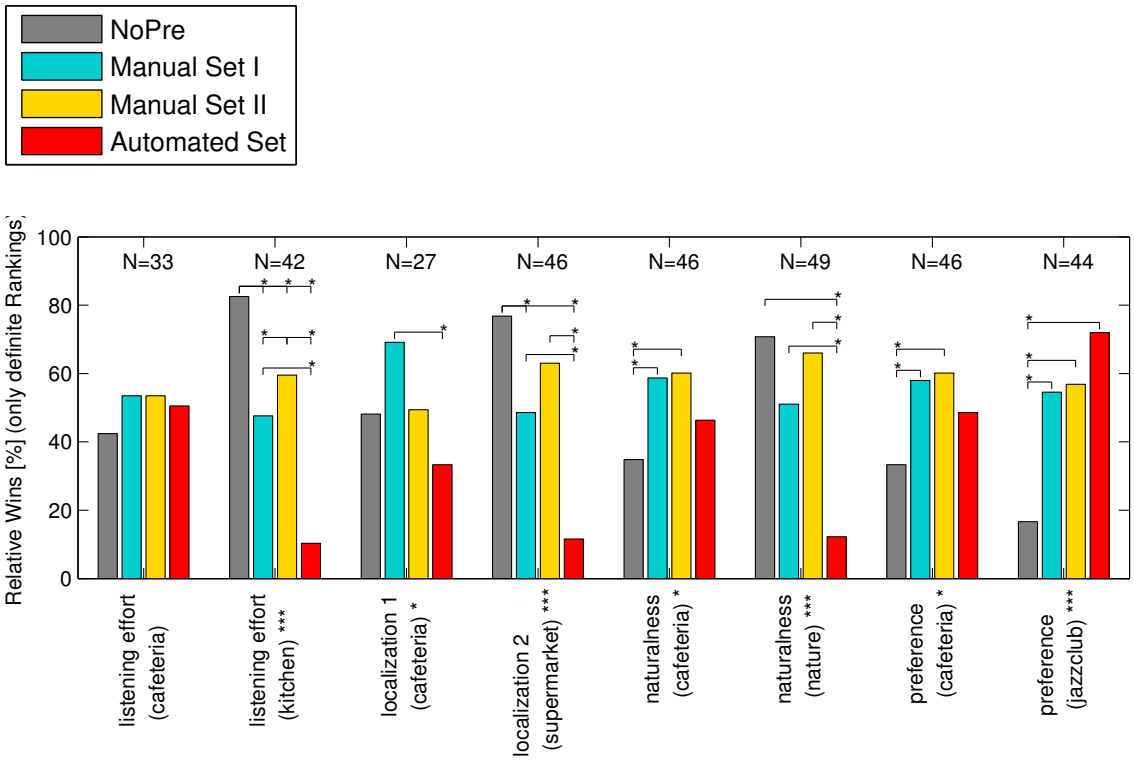

Figure 4.7: *Relative wins of each pre-processing strategy (NoPre, Manual Set I, Manual Set II, Automated Set) regarding the questioned subjective attributes listening effort, sound localization, naturalness, and preference within the different sound scenarios. The definite rankings collected from 31 participants from two test centers are displayed (test and retest). For each of the eight attribute-scenario pairs the amount of consistent data sets N is also given in the Figure. Significance strength for each test condition is denoted next to their labels (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Significant differences between the algorithm settings within each test condition are denoted within the Figure (*: $p < 0.0083$).*

NoPre (76.8%) had significant more relative wins than Manual Set I and Automated Set (both with $p < 0.0083$). Manual Set I and Manual Set II also had more wins than Automated Set ($p < 0.0083$). Comparing Manual Set II with NoPre and Manual Set I showed no significant differences ($p > 0.5$). Regarding naturalness (within the cafeteria scenario), 74.2% of valid ratings were given, which led to a significant effect, $\chi^2(3) = 10.5$, $p < 0.05$. Post hoc analysis showed that Manual Set I and Manual Set II had significant more wins than NoPre (both $p < 0.0083$). In the attribute-specific nature scene the listeners gave the highest amount of consistent ratings (79%). Within-subjects factor *algorithm setting* affected the ratings significantly, $\chi^2(3) = 55.9$, $p < 0.001$. Multiple comparisons showed that all algorithm conditions had significant more wins than Automated Set (all $p < 0.0083$). The remaining three comparisons were not significant ($p > 0.05$). 74,1% of the collected preference ratings (in the cafeteria scenario) were consistent, which led to a significant effect of the factor

algorithm setting, $\chi^2(3) = 11.1$, $p < 0.05$. Identical to the tested naturalness in the cafeteria, Manual Set I and Manual Set II had significant more wins than NoPre (both $p < 0.0083$), and the remaining comparisons were not significant ($p > 0.05$). The preference ratings in the jazz club scenario were also significantly affected by factor *algorithm setting*, $\chi^2(3) = 39.5$, $p < 0.001$. Here, all three multiple comparisons of NoPre (16.7%) with Manual Set I (54.5%), Manual Set II (56.8%), and Automated Set (72%) showed that the latter had significant more wins (all $p < 0.0083$). Differences between the other algorithm conditions were not significant ($p > 0.05$).

4.5 Discussion

4.5.1 Instrumental evaluation

The results from the instrumental evaluation mostly showed a clear and distinct ranking regarding the performance of the different algorithm settings: Automated Set performed best, followed by Manual Set I on the second place and Manual Set II on the third place. Especially regarding the constant SNR benefit by Automated Set over the tested input SNRs, we can conclude that although the optimization procedure was based on the single SNR of 0 dB, the automated approach still led to a robust setting over a large range of input SNRs (cf. Figure 4.5, top left panel).

Furthermore, it can be observed that the performance measures SII and BINSII showed almost identical results. However, a comparison of the raw values of SII and BINSII (not shown) reveals that BINSII was constantly higher by approximately 0.06. This equally held for NoPre, Manual Set I, Manual Set II, and Automated Set. Thus, we found that the algorithm cannot enhance the monaural SII differently than BINSII with respect to NoPre in each of the measures. This may be due to the fact that the binaural cues are kept unchanged during processing by the coherence-based algorithm (cf. Figure 4.1).

The distinct ranking regarding performance did not hold for low SNRs (-10 to -4 dB) with respect to SII and BINSII, where all settings showed no enhancements regarding NoPre. This is especially surprising for Automated Set, which led to a large global SNR improvement (approximately 6 dB), yet the SII/BINSII was not enhanced. However, a closer examination of the spectral characteristics of the signals (NoPre versus Automated Set) at an input SNR of -5 dB may provide an answer. Due to the aggressive noise reduction by Automated Set ($\alpha = 3.87$), which leads to a strong low pass filtering of the complete signal, the global SNR was determined within the low frequency bands. These bands featured high band levels and high band SNRs, leading to a global SNR of approximately +1 dB. However, The global

SNR of NoPre was also determined by middle frequency bands with lower band SNRs, leading to a global SNR of -5 dB). While both signals shared the same SII/BINSII (no enhancement), the signals differed in global SNR by 6 dB, i.e., an improvement in global SNR did not necessarily have to be reflected in an SII/BINSII improvement. This finding supports the claim that an evaluation by global SNR alone can be misleading as it does not account for the degradation of signal quality caused by e.g. low-pass filtering (cf. p. 308; Heute, 2008).

4.5.2 Perceptual evaluation

Considering the SRT measurements, we found that no setting of the algorithm was able to improve SRT. Manual Set II performed best, i.e. by not degrading SRT with respect to NoPre, followed by Manual Set I (SRT degradation of 1.1 dB) on second place and Automated Set (degrading SRT by 2.1 dB) on third place. This subjective SRT ranking is the exact opposite of what was found due most of the instrumental evaluation, where Automated Set was best, followed by Manual Set I and Manual Set II. While it is reassuring that Automated Set performed best within the same instrumental measures on which the automated optimization strategy was based, all the more it is surprising that Automated Set performed worst regarding SRT. This discrepancy between instrumental and perceptual results, as also found in earlier studies (e.g., Luts et al., 2010), underlines the inability of the instrumental measures to provide a valid estimate of the real benefit. However, the described divergence could not be seen in the instrumental evaluation with PSM at high SNRs (cf. Figure 4.5, lower right panel). In this measure the ranking of the measured SRT data is reflected. A further automated parameter optimization based on the PSM performance measure alone at an input SNR of 10 dB may deliver other promising algorithm settings.

The evaluation with respect to the subjective attributes listening effort, localization, naturalness, and preference within paired comparisons showed that it was difficult for the participants to give definite rankings of the four tested alternative settings. An explanation for this could be that the presented alternatives were perceptually too similar to give a clear ranking. Further measurements with a control group, consisting of younger normal-hearing persons, who are also experienced in listening tests, may gain more insight into the specific causes for the observed inconsistencies.

However, after excluding inconsistencies in the judgments, the cleaned data is interpretable. Regarding the statistical significance of each subjective attribute, it could be stated that evaluating the attributes within the cafeteria scenario led to less statistical significance than within the attribute-specific scenarios. Within

the cafeteria, three out of four attributes were significant with $p < 0.05$, and most of the inconsistent rankings were found in this scenario (listening effort and localization). Within the attribute-specific scenarios however, all four attributes were highly significant with $p < 0.001$. As the challenging cafeteria scenario contained a conversation, we speculate that the participants judged the stimuli unintentionally rather by their speech intelligibility than by the questioned attributes. The speech intelligibility was detrimentally affected by the highly reverberant ($T_{60} = 1250$ ms; Kayser et al., 2009) scenario. As all stimuli were similarly affected by the reverberation, the discrimination of the already hard to be distinguished alternatives became even harder in this scenario. This could not only be seen in the less statistical significance within the cafeteria, but also in the similar shapes regarding the relative wins (cf. Figure 4.7). Especially the attributes naturalness and preference exhibit a nearly identical distribution.

With the exception of the preference rating, the evaluated attributes within the attribute-specific scenarios also showed similar distributions and trends. The listening effort within the kitchen scenario is the only attribute, for which an explicit 'winning' algorithm condition could be found (NoPre) that asserts over all others by having significantly more relative wins. While not being the single winner in terms of statistical assertiveness, NoPre also showed the highest amount of relative wins regarding localization and naturalness. Automated Set had the lowest amount of relative wins. However, within the preference ratings in the jazz club, Automated Set had the highest amount of wins and asserts over NoPre. It is unclear why the coherence-based noise reduction scheme did not lead to a lower listening effort with respect to NoPre as it was found by Luts et al. (2010). As Luts et al. (2010) performed a listening effort scaling on a 13 point scale rather than pairwise comparisons, we speculate that the differences were due to this measurement protocol deviation. Regarding the preference ratings within the jazz club scenario, where Automated Set had the highest amount of relative wins, loudness seemed to be the crucial factor. As the intention was to evaluate under preferably most realistic conditions, the jazz club scenario exhibited the highest overall sound pressure level of the noise scenarios (81.7 dB SPL). Here, the noise reduction scheme with the most aggressive setting (Automated Set: $\alpha = 3.87$) suppressed the most of the signal components which led to less annoyance and the highest preference (72%), whereas the unprocessed 'loudest' stimuli had the lowest amount of wins (16.7%) presumably due to the highest annoyance.

Due to the higher discrimination ability of the participants within the attribute-specific scenarios (higher statistical significance), the evaluation using these realistic

scenarios proved to be helpful to gain insight into a possible real-world benefit of the actual noise reduction scheme within controlled measurements in the laboratory.

4.5.3 Automated model-based parameter optimization

It has to be stated that the direct comparison of the algorithm performance gained by the optimized settings (Automated Set) and the second manual setting (Manual Set II) from this study is delicate as the introduced and evaluated unsupervised parameter optimization strategy provides only scalar values for each parameter, whereas Manual Set II uses frequency-dependent parameter vectors and therefore offers a larger potential of adjustments. In a follow-up study, the optimization strategy could be extended to perform a subband-based optimization (cf. Rohdenburg et al., 2006) to provide optimized parameter vectors, so that the best parameter combination would be found for each frequency band.

However, although the better performance within the instrumental evaluation, the unsupervised parameter optimization did not perform better in terms of the perceptual evaluation than Manual Fit I, which also uses scalar values for each parameter. Possible interpretations of this discrepancy between the results from the computer-based instrumental and perceptual measures are given in the following.

At first, it is possible that the four used performance measures were not able to adequately capture the signal changes induced by the processing of the noise reduction algorithm, i.e. an interaction between algorithm and measures might have caused the discrepancy within the results. More generally, it is possible that at least one of the used performance measures was not able to model the human perception under the test conditions. As it was stated above, the purely technical measure global SNR can be misleading. Here, the global SNR measure (cf. Figure 4.3) tends to more aggressive settings ($\alpha > 2$), and an exclusion of this measure could be beneficial. Other auditory-model based measurements like e.g. PESQ (ITU-T — Telecommunication Standardization Sector of ITU, 2001) and HASQI (Kates and Arehart, 2014), which are less forgiving against induced artifacts, could be taken into account instead.

Utilizing the presented selection of performance measures, the used weighting could be responsible for the discrepancies within the results of this study. Here, the performance measures were equally weighted to find a compromise setting, but a different weighting could be of advantage, e.g. reflecting different preferences.

Further developments for the introduced optimization approach could aim at extracting more than one setting. These settings could lie on a reasonable way within the multidimensional parameter space and could form the input for the subjective

fine-tuning of hearing aids. Therefore these settings could be evaluated within a multi-stimulus test procedure like MUSHRA (ITU-R — Radiocommunication Sector of ITU, 2014a), which does not allow for cyclic triads.

4.6 Conclusion

The aims of this study were to introduce and evaluate an approach for a model-based parameter optimization for complex multidimensional hearing aid algorithms. The optimization approach was applied to a binaural coherence-based noise reduction algorithm. Together with two further manual parameter adjustments of the algorithm, the three settings were compared within instrumental and perceptual measurements. For the latter, among SRT measurements, subjective attributes were measured in attribute-specific noise scenarios.

The statistical significance of the perceptual evaluation within the developed attribute-specific listening scenarios was higher for all four tested attributes than within the standard listening scenario, the cafeteria. Thus, the evaluation using these realistic scenarios is potentially more helpful gaining insights into a real-world benefit with controlled measurements in the laboratory than the 'standard scenarios'.

In summary, it could be demonstrated that the automated model-based optimization routine led to the best-performing setting regarding instrumental evaluation measurements. However, with respect to the perceptual evaluation, both expert-based settings of the exemplary algorithm performed better than the automated setting for the majority of evaluated attributes. Possible explanations could be a disadvantageous selection of objective performance measures or the used equal weighting of the latter within this study.

Despite the found discrepancies between instrumental and perceptual evaluation for the exemplary algorithm here, we suggest that the model-based optimization approach is a feasible tool for generating promising algorithm parameters to be used for e.g. a further perceptual evaluation. The optimization approach systematically reduces the amount of hearing aid settings and can support the audiologists in finding the optimal setting for the individual patients.

5 Hearing aid fitting and fine-tuning based on estimated individual traits

ABSTRACT

Objective: A generalized concept for a hearing aid fitting and fine-tuning, which is based on estimated individual traits, is presented together with first implementations within this report. *Design:* To estimate the individual traits, auditory model-based performance measures are used in combination to generate promising candidates within the algorithm‘s parameter space for a subsequent subjective rating. For the subjective assessment, a fast and intuitive multi-stimulus test denoted as CoDiCl (Combined Discrimination and Classification) is presented to capture user preferences for an optimized setting. *Study sample:* The estimation of individual traits is exemplary shown on a multidimensional coherence-based noise reduction algorithm. The dimensionality reduction was performed using differently weighted combinations of speech intelligibility index (SII) and perceived similarity measure (PSM). *Results:* Nine reasonable alternative algorithm setting candidates were extracted from a defined path of probation for a subsequent subjective rating to potentially differentiate between patients with different attitudes towards noise suppression and introduced distortions (i.e.‘noise haters’ and ‘distortion haters’). *Conclusions:* By iteratively improving the agreement between subjective and objective assessment, an objective estimation of subjective traits using appropriate weightings of objective measures may become possible. This will potentially help to efficiently fit modern multidimensional hearing aid algorithms to the individual patient.

5.1 Introduction

Modern digital hearing aids incorporate a large selection of complex signal processing algorithms like multiband compression, feedback management, directional micro-

phones, and noise reduction that are each controlled by a complex set of parameters. In combination, these algorithms create the possibility of various different parameter sets individually tailored to the needs of the individual user. The overall aim of a successful hearing aid fitting is therefore to match the multidimensional parameter space of the hearing aid algorithms with the hearing aid users' needs and expectations. These are determined by the individual processing deficiency caused by the hearing impairment and the individual user preferences. The current report outlines a concept of achieving this match by generating optimized parameter combinations that should differentiate across listeners with different individual traits, thus putting as much of the uncertainty about the individual preferences as possible into the variation across different presets while converting the available (objective) knowledge about algorithmic performance into the optimization of each preset.

The hearing aid fitting process usually contains two stages, a prescriptive and a fine-tuning stage. Within the first stage, a basic fit of the hearing aid is given by prescriptive, rule-based procedures that incorporate e.g., audiometric hearing thresholds or suprathreshold measures like loudness discomfort levels of the patient and provide a (frequency- and input level dependent) gain prescription based on average performance data. Examples for these procedures are NAL-NL1 (Byrne et al., 2001) and DSL [i/o] (Cornelisse et al., 1995). A detailed overview of various fitting strategies is given by Dillon (2012). However, literature shows that the prescriptive initial fitting rarely proofed to be the optimal and satisfying fit for the individual patient (e.g., Leijon et al., 1984; Keidser and Dillon, 2006). Wong (2011) stated that in general, prescribed gains have been found to be slightly higher than the average preferred hearing aid responses.

Most clinicians agree that the first prescribed setting however serves as a reasonable starting point for a necessary subsequent fine-tuning of the hearing aid, which can be regarded as the second stage of the fitting process. As the fine-tuning aims at the final matching of the hearing aids to the individual needs and preferences of the patient, this process is crucial with regard to user satisfaction. Confirmed experience characterizes the fine-tuning process as very time-consuming, possibly involving many repeated visits at the audiologist (e.g., Boymans and Dreschler, 2012; Abrams et al., 2011). While the initial first fit is relatively well defined by generic rules, the individual fine-tuning process is not (Boymans and Dreschler, 2012). Usually, the audiologist uses a complaint-driven fine-tuning procedure being part of the proprietary fitting software of the respective hearing aid manufacturer (Jenstad et al., 2003). This means that a successful fine-tuning process is very much dependent on the expertise of the audiologist and the commercial fitting software.

Regarding personal sound preferences, individual needs and demands towards a hearing system, the potential users and patients are very different. Therefore we assume that the individual sound preference judgments are based on different individual weightings of prototypical *subjective traits*. It has been shown that there are considerable interactions between such fundamental perceptual aspects, as e.g. an existing trade-off between intelligibility and listening comfort (Brons et al., 2014). Additional trade-offs were found within individual preferences of hearing aid users. For example, while some listeners are more tolerant towards distortion or artifacts which are generated by e.g. an aggressive application of noise reduction, others rather accept residual noise as long as no artifacts are provoked (Marzinzik, 2000). It appears that one prototypical subjective trait is 'noise hating' while another trait is 'distortion hating', whereupon both traits are not obligatory oppositional, i.e. a prototypic 'distortion hater' is not mandatory a 'noise lover'.

Although the exact specifications of all relevant subjective traits are not known, we assume that these traits exist and that the optimal individual hearing aid setting or fitting can be predicted from the knowledge of an unspecified amount of subjective traits that are weighted and combined in an individual way. Furthermore we assume that this individual weighting of subjective traits remains constant as long as the person resides in the same (or similar) acoustical context.

The presented concept for an optimized hearing aid fitting process within this report is illustrated in Figure 5.1. On the left, the subjective part **(Patient)** including the subjective traits is displayed. The individual assessment of the acoustic situation is found by a linear combination of the traits weighted with the individual perceptual weights a_i. On the right side, the objective part **(Model Scale)** with the objective perceptual measures is shown. Different weightings of the instrumental weights b_i lead to different objective assessments for the same acoustical situation. Now the long-term goal is to reach an agreement between the objective and subjective assessment. Such an agreement might be found by individualized trait-based models, i.e. by developing objective measures that represent the individual traits as well as possible: The individual best fitting instrumental weights $b_1 \cdots b_n$ estimate quite well the perceptual weights $a_1 \cdots a_n$. Alternatively, the set of subjective traits can be characterized by an appropriate selection and combination of objective measures, such that the best-fitting perceptual weights $a_1 \cdots a_n$ estimate quite well the required objective perceptual weights $b_1 \cdots b_n$. Hence, within a first iterative *learning phase* of the presented concept, it is attempted to cover the prototypical subjective traits by objective perceptual measures as precisely as possible. For this, only such objective measures are utilized, which reflect the assumed subjective traits as accurately as

possible, and only such subjective traits are to be used, which can be characterized by objective measurements. In a subsequent *production phase*, selected objective measures are used to generate optimized settings. These settings should a) reflect local optimum values of the objective parameters (i.e., optimizing intelligibility or sound quality or other global performance measures) and b) be targeted towards discriminating across different individual weightings of these hypothetical traits, i.e. several alternatives are presented that are based on prototypical weightings of these traits. The user then is confronted with a manageable amount of alternative settings that follow a *path of probation* and can decide across objectively optimized settings that only vary across the different assumed possible realizations of individual weightings of traits. Note that these judgments are then performed with an intuitive test procedure independently from an assumed personal preference profile.

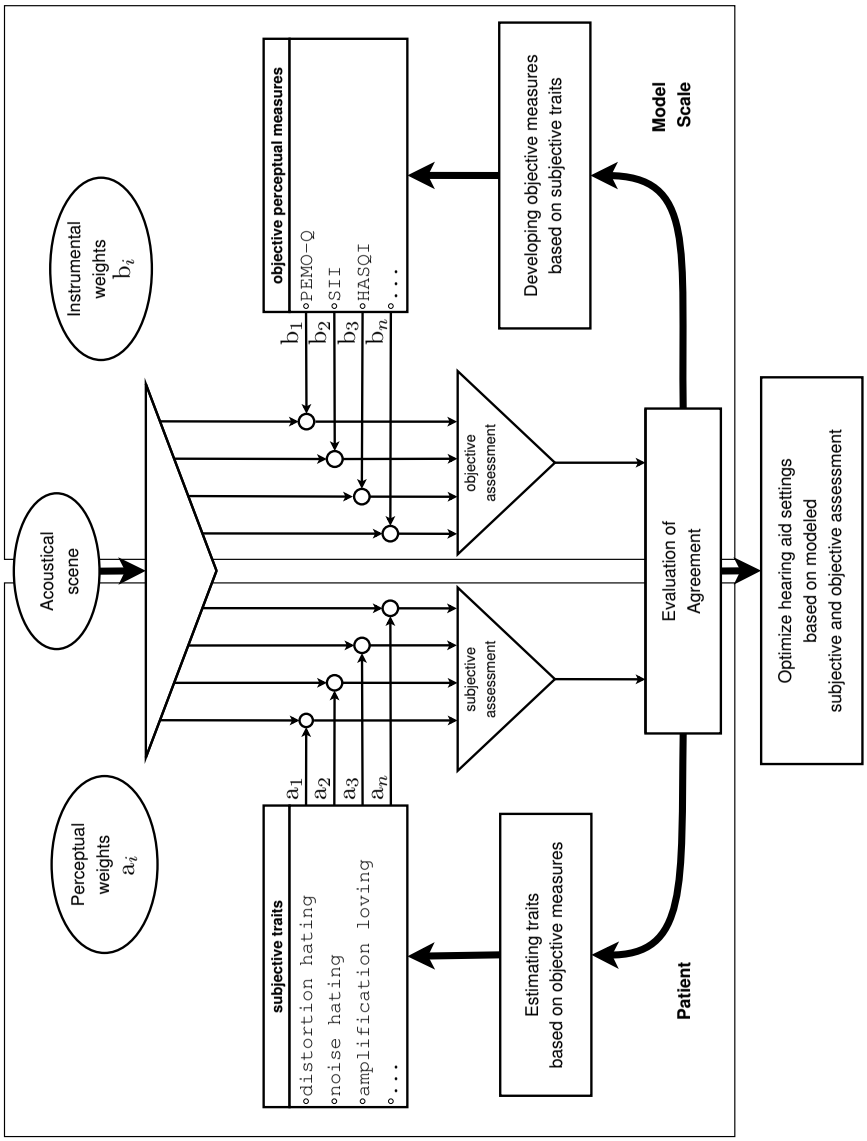

Figure 5.1: *Concept for an optimization of the hearing aid fitting procedure based on modeled subjective and objective assessment.*

5.2 Objective assessment: Extraction of parameter combinations reflecting prototypes and trade-offs regarding subjective traits

The method to extract several optimized parameter combinations that are supposed to sample the patient‘s best-assessed settings for a hypothetical set of individual trait realizations is a further development and generalization of the procedure introduced by Völker et al. (2016a). The method described by Völker et al. (2016a) is designed to generate one optimized setting for multidimensional algorithms, especially in cases when no evidence-based prescriptive formula is available to fit the actual algorithm. The procedure calculates performance contours of instrumental measures within the algorithm parameter space and finds one optimized setting in the minimal intersection of the contours. In their study, the authors used four performance measures that contribute equally to the optimized setting, i.e., the extracted parameter setting can be regarded as a balanced compromise setting of the used performance measures. The used measures consisted of a technical measure, the signal-to-noise ratio (SNR), and three auditory-model based measures, the speech intelligibility index (SII; ANSI 20S3.5-1997, 1997), binaural speech intelligibility index (BINSII; Beutelmann and Brand, 2006), and the perceived similarity measure (PSM; Huber and Kollmeier, 2006).

As stated above, the aim of the concept introduced within this report was to find a manageable amount of presets estimating possible individual realizations of the subjective traits and let the patients interactively find their favorite setting by themselves. Regarding this manageable amount of stimuli, we follow the recommendation for the *MUlti Stimulus test with Hidden Reference and Anchor*, in which it is stated that no more than 12 signals should be included in any trial (MUSHRA; ITU-R — Radiocommunication Sector of ITU, 2014a).

Therefore the original method from Völker et al. (2016a) was extended to locate a reasonable *path of probation* through the algorithm parameter space and to extract several settings thereon to be presented to the patient.

During the first iterative *learning phase* of the presented concept, a sufficient agreement between objective and subjective assessment was established by selecting only such objective measures characterizing the assumed prototypical subjective traits. Now following the concept of calculating minimal intersections of percentile contours within the algorithm parameter space, each of the used (optimized) objective performance measures contributes with different weightings to the extraction of optimized settings reflecting realizations of the subjective traits. Possible initial

weightings, i.e. initial values for the percentile calculation of each performance measure could be values ranging from 0 to 100% with a step size of 20% leading to 6 initial percentile values. The combination with the other participating measures is such that the sum of the percentile values of the involved measures is always 100%. In case of three objective measures, this would lead to $\sum_{k=1}^{6} k = 21$ different initial weightings and combinations of the measures. In case of a step size of 5% and three objective measures, this would lead to $\sum_{k=1}^{21} k = 231$ different combinations of the measures, which would not meet the demand regarding a manageable amount of alternative settings to be presented to the patient (i.e., no more than 12 signals). A potential solution to still reduce the amount of extracted settings is to use different step sizes for the different involved objective performance measures. With respect to the individual importance of the one or the other prototypical subjective trait, the step size for the objective measure estimating the respective different trait realizations can be changed. A greater step size (and therefore less extracted settings) of a measure can be used if the estimated trait is partially marginal for the individual preferences. To determine the individual importance of subjective traits, e.g. questionnaires can be used before the extraction of parameter settings.

5.2.1 Coherence-based noise reduction algorithm

To introduce the practical value of the suggested concept for a trait-based hearing aid fitting and fine-tuning, the estimation of subjective traits and extraction of prototype- and trade-off settings is exemplary performed on a recently developed binaural coherence-based noise reduction algorithm (Luts et al., 2010; Baumgärtel et al., 2015b).

The scheme calculates the coherence (i.e., the similarity) between the signals captured at the left and right ear in several frequency bands. Depending on the detected coherence, gains are applied to the bands. As it is assumed that coherent signal parts primary contain the desired speech signal, the respective gains are not modified. Incoherent signal parts, containing noise, however, are suppressed by the algorithm.

The algorithm works in the short-time Fourier Transform (STFT) domain, where STFT bins are grouped into 15 non-overlapping third-octave frequency bands k with center frequencies ranging from 250 Hz to 8 kHz. The interaural phase difference (IPD) is used as an estimate for the coherence. The coherence $C(k, l)$ in each frequency band k and time segment l is estimated from the vector strength of the complex IPD $c_{\mathrm{IPD}}(k, l)$, as defined in Grimm et al. (2009):

$$C(k,l) = \left|\langle c_{\mathrm{IPD}}(k,l)\rangle_{\tau}\right|. \tag{5.1}$$

The coherence value is estimated using a running average $\langle\cdot\rangle_{\tau}$ with time constant τ. After a linear mapping of the coherence estimates to the interval $[0,1]$, the gain in each frequency band is computed by applying an efficiency exponent α, i.e.:

$$G(k,l) = \widehat{C}(k,l)^{\alpha}. \tag{5.2}$$

The algorithm main parameters affecting performance are i) the efficiency exponent α which reflects the aggressiveness of the noise reduction, and ii) the time constant τ, which represents the smoothing duration of the running average estimating the coherence.

5.2.2 Exemplary preliminary application of algorithm parameter extraction

To demonstrate the feasibility of the algorithm parameter extraction, the methodology was reduced to two objective performance measures within this report. It is well-known that speech intelligibility and speech quality encompass fundamental aspects within communication technology (e.g., Kondo, 2012). Kates and Arehart (2010) state that 'intelligibility is the most important consideration in amplification for the hearing impaired'. However, Kochkin (2005) found that sound quality correlates with hearing aid user satisfaction. For this reason, the end points of the *path of probation* within the algorithm parameter space were defined by using two auditory-model based instrumental performance measures reflecting either speech intelligibility (SII; ANSI 20S3.5-1997, 1997) or quality (PSM; Huber and Kollmeier, 2006).

The extraction of optimized settings follows the concept of calculating percentile contours for each of the performance measures as described in Völker et al. (2016a). These contours are displayed on top of the algorithm parameter space, a two-dimensional (α vs. τ) space in case of the coherence-based noise reduction algorithm used in this study (cf. Figure 5.2). For example, the 70% percentile contour calculated with PSM displayed on top of the parameter space encloses all combinations of α and τ which represent the best 30% of results regarding the evaluation with quality measure PSM. By calculating and displaying percentile contours for more than one performance measure, intersections are formed containing parameter combinations equidistantly spaced in the objective parameter space. By using different percentile values for the calculation, different instrumental weightings b_i of the involved measures can be realized.

Differently weighted percentile contours for the two measures SII and PSM were calculated forming intersections in the two-dimensional algorithm parameter space. While percentiles from 100% to 0% with a decreasing step size of 5% were computed for measure SII, the percentiles for PSM were increased from 0% to 100%, leading to 21 initial weightings of the percentile contours. If these initial weightings (e.g., 90% SII, 10% PSM) hold more than one setting, **both** percentiles are subsequently increased by factor 0.01 until the intersection is minimal, i.e., containing only one setting.

The different weightings, each delivering one optimized setting, are listed in Table 4.1. Finally, nine unique settings were identified (cf. first column) as some calculated weightings led to the same optimal parameter combinations. The resulting proposed *path of probation*, exemplary obtained following the described method within the two-dimensional parameter space (α vs. τ) of the coherence-based noise reduction algorithm, is shown in Figure 5.2 (dashed line). The nine extracted unique settings to be used for a subsequent subjective rating are marked by black crosses.

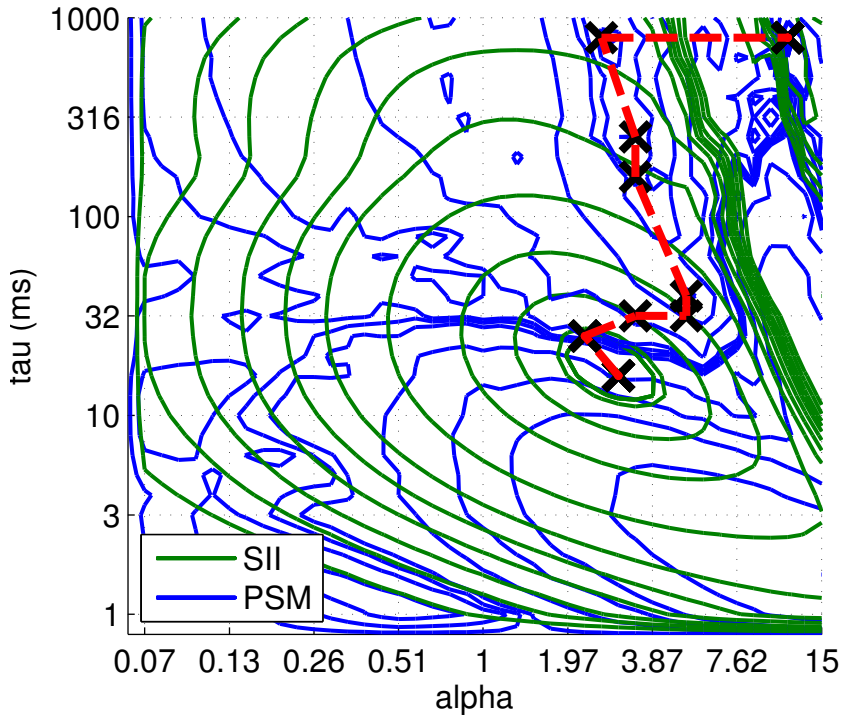

Figure 5.2: *Model-based path of probation (dashed line) within the two-dimensional parameter space (alpha, tau) of a coherence-based noise reduction algorithm generated by minimal intersections of percentile contours of differently weighted performance measures SII (green) and PSM (blue). The markers show nine extracted unique settings to be used for further subjective assessment.*

Table 5.1: *List of extracted unique optimized settings for the common coherence-based noise reduction algorithm.*

Setting	Weighting [in %] (SII / PSM)	Parameter α	Parameter τ
1	100 / 0	2.95	15.84 ms
	99.89 / 5.25	2.95	15.84 ms
	99.89 / 11.09	2.95	15.84 ms
	99.89 / 17.62	2.95	15.84 ms
	99.89 / 24.97	2.95	15.84 ms
2	98.92 / 32.97	2.25	25.11 ms
	98.92 / 42.39	2.25	25.11 ms
	98.92 / 53.26	2.25	25.11 ms
	98.62 / 65.74	2.25	25.11 ms
3	95.50 / 78.14	3.38	31.62 ms
4	89.47 / 89.47	5.07	31.62 ms
5	77.95 / 95.27	5.07	39.81 ms
6	64.92 / 97.38	3.38	158.48 ms
7	53.42 / 99.22	3.38	251.18 ms
8	42.68 / 99.59	2.58	794.32 ms
9	33.29 / 99.89	11.44	794.32 ms
	24.97 / 99.89	11.44	794.32 ms
	17.62 / 99.89	11.44	794.32 ms
	11.09 / 99.89	11.44	794.32 ms
	5.25 / 99.89	11.44	794.32 ms
	0 / 100	11.44	794.32 ms

5.3 Subjective Assessment: Combined Discrimination and Classification Task (CoDiCl)

As a part of the introduced approach for an optimized hearing aid fitting and fine-tuning within this report, a method for the subjective assessment is presented in the following. A necessary requirement for the rating method is to not affect the measurement at all or to exert a controllable effect on the measurement results. The utilized method is motivated by a modification of the standardized *MUlti Stimulus test with Hidden Reference and Anchor* (MUSHRA; ITU-R — Radiocommunication Sector of ITU, 2014a) implemented and evaluated by Völker et al. (2016b). Their modification MUSHRA drag&drop was inspired by the proposed test procedure of Pfitzinger (2003). The modification uses a different graphical user interface allowing the users to sort the stimuli more intuitively from left to right with the help of the 'drag & drop' technique. It has been shown that the modification MUSHRA drag&drop can be used alternatively to the original MUSHRA test, leading to compatible final rating results. In terms of the objective evaluation,

MUSHRA drag&drop was faster than original MUSHRA. Furthermore, the assessor performance (reliability and discrimination ability) measured with eGauge (Lorho et al., 2010; ITU-R — Radiocommunication Sector of ITU, 2014b) was highest for the majority of participants when using MUSHRA drag&drop.

In contrast to the MUSHRA test, the method in this report is not making use of references or anchors and can therefore be regarded as an independent method, namely a combined discrimination and classification task (CoDiCl). With this method, considered parameter combinations can be presented by the audiologist to the hearing impaired persons, who can then intuitively find and rate their favorite setting based on their personal sound preferences.

The presented rating method is an alternative to the method of paired comparisons (cf., Amlani and Schafer, 2009). In contrast to a paired comparison task which can produce cyclic triads, it is not possible to produce such inconsistent ratings within the introduced discrimination and classification task. Also in terms of time duration, the presented method can be of advantage with respect to a full pairwise comparison of alternative settings. Regarding a potential challenge for elderly hearing aid users to handle the computer-based drag&drop technique, it has already been shown that this group of persons is able to successfully use this technique (cf., Völker et al., 2016b; Abrams et al., 2011).

A screenshot of the utilized user interface is given in Figure 5.3. The figure shows an exemplary final rating regarding the overall quality of the nine stimuli (A–I) gathered by the procedure to extract multiple optimized settings described above (cf. Table 4.1 and Figure 5.2). The stimulus ‘E’ is active and the sound-scene is played loop-wise.

5.4 Discussion

Within this report, the concept of a hearing aid fitting and fine-tuning based on estimated weightings of prototypical subjective traits to predict the individual preference was introduced. Within a first iterative *learning phase*, it is attempted to cover the relevant subjective traits by objective perceptual measures as precisely as possible. In a subsequent *production phase*, selected objective measures are used to generate settings that should reflect the assumed traits as well as possible and should be targeted towards discriminating across different individual weightings of these traits. The user then is confronted with a manageable amount of alternative settings and can autonomously decide regarding personal preference with an intuitive test procedure.

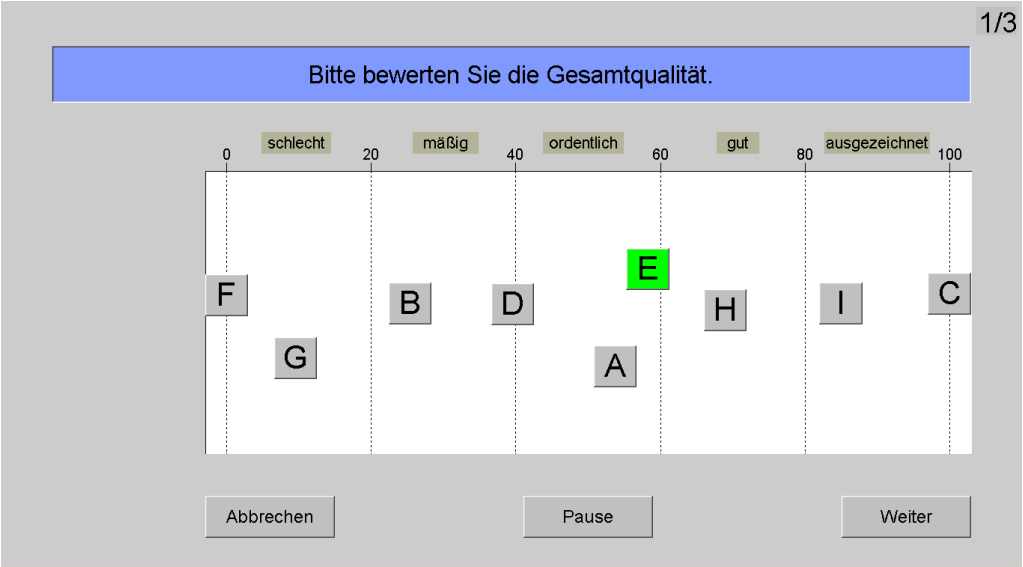

Figure 5.3: *Graphical user interface of the introduced fitting tool, a combined discrimination and classification task (CoDiCl). A possible final rating regarding the overall quality of nine stimuli (A–I) is shown. The stimulus 'E' is active and playing loop-wise at the moment.*

Following the presented concept in the long term can help to match the multidimensional settings of hearing aid algorithms to the complex hearing aid users with their individual hearing losses and all their individual preferences and needs.

It is known that trade-offs between perceptual aspects like intelligibility and listening comfort (Brons et al., 2014) are possible and that there are different individual preferences for those trade-offs (Abrams et al., 2011; Marzinzik, 2000). Therefore especially the usage of opposed auditory model-based performance measures reflecting different fundamental perceptual aspects in the definition for a *path of probation* within the algorithm parameter space and the extraction of alternative settings is promising. By letting the patient find the preferred setting by himself with an intuitive and effective selection method, the user satisfaction might be increased as the patient is more directly involved, which offers therefore a positive psychological impact (cf. 'IKEA effect', Norton et al., 2012).

Further developments of the concept should encompass the generation of an individualized *path of probation* through the parameter space. This can be realized by individualizing model-based performance measures, e.g., incorporating individual hearing thresholds into the SII calculations (ANSI 20S3.5-1997, 1997) or the binaural speech intelligibility model (BSIM; Beutelmann and Brand, 2006).

As it is unlikely to achieve perfect orthogonality in terms of speech intelligibility and quality (cf., Kiessling et al., 2006), it can be beneficial to use several model-based measures representing quality and intelligibility. With the help of subsequent

principal component analysis (PCA) or factor analysis, the dimensionality of the used measures can be reduced to few components that are orthogonal to the greatest possible extent. Generating now a *path of probation* through the algorithm parameter space by different weightings of these principal components can be advantageous for capturing the individual trade-off between perceptual aspects. Furthermore, this reasonable path within the parameter space is a model-based approach similar to the concept of meta-controllers (Kiessling et al., 2006) for hearing aids. By adjusting the (one-dimensional) meta-controller, e.g. the hearing aid user can control the multidimensional hearing aid from the maximum of the first principal component to the maximum of the second principal component within the parameter space. However, a comparison between the meta-controller concept and the *path of probation* concept is yet open: It is unclear if the hearing aid user would find the same preferred setting by manually adjusting the meta-controller than e.g. with the introduced subjective rating procedure of given presets along the *path of probation*.

To gain more insight into the relations between perceptual aspects and individual user preferences, the introduced subjective rating tool can be extended with respect to the utilized dimensionality within the test procedure. After testing e.g., the perceptual aspects intelligibility and quality separately, both dimensions might be evaluated simultaneously by utilizing both axes of the combined discrimination and classification procedure.

6 General Discussion & Conclusion

The long-term goal of the present thesis was to contribute to the fitting of modern hearing devices with a multidimensional parameter space to their individual users with their individual needs and preferences. The thesis thereby provides a) objective methods to generate meaningful parameter settings for the prescription and subsequent fine-tuning of hearing aid algorithms and b) suitable perceptual methods for evaluating the respective hearing aid settings.

The comprehensive comparison of various pre-processing schemes within this thesis (see chapter 2) showed that the instrumental measures were not able to predict the perceptual data in all tested noise conditions and further development is necessary to predict the benefits obtained from the algorithms at an individual level. Alongside with other studies (e.g., Luts et al., 2010), it was shown that evaluating hearing aid algorithms by instrumental measures alone would possibly not reflect the true individual benefit, and the presented study underlines the importance of perceptual measurements for a valid hearing aid evaluation.

One important measure of the individual hearing aid benefit is the assessment of the perceived quality within a given auditory scene. Two preferably barrier-free modifications for the standardized perceptual test procedure 'MUlti Stimulus test with Hidden Reference and Anchor' (MUSHRA) were designed and evaluated within chapter 3. As the evaluation of noise reduction algorithms with the modification 'MUSHRA drag&drop' led to the highest percentage of reliable data from the 50 participants (82%), 'MUSHRA drag&drop' can be generally recommended for perceptual measurements.

But even the most intuitive and barrier-free perceptual test procedure would not solve the complex problem of fitting modern multidimensional hearing devices to their individual users, because a) these test procedures are time-consuming and therefore cost-intensive, b) the amount of alternative algorithms, and c) the amount of the particular alternative algorithm settings is too high as to be tested successively by humans. This means that we are dependent on instrumental measures as they can potentially be used to detect the best-suited algorithm including the best particular

setting of the sophisticated hearing aid algorithms or at least reduce the number of meaningful alternatives to be tested with real patients.

As the study from chapter 2 had shown that single instrumental measures were not able to predict the different algorithm outcomes satisfactorily, another approach was tested in the study from chapter 4 to master the mentioned multidimensionality challenge of hearing aid algorithms: Several instrumental measures were combined to reduce the number of meaningful parameter combinations of a noise reduction algorithm and therefore optimize the algorithm outcome. It was shown that the combination of instrumental measures delivering one compromise setting was not ideal as the equal weighting of the four used instrumental measures did not correlate with the preferred and best-performing settings for the individual users. A possible explanation for the results could be the existence of different individual weightings of so-called prototypical subjective traits, which was not considered in the parameter optimization procedure. Assumed prototypical subjective traits are e.g. 'distortion hating' and 'noise hating'.

Based on these findings, a way to construct a balanced, optimized 'path of probation' between the computed optimum values derived for two (or more) instrumental measures reflecting assumed prototypical subjective traits within the multidimensional parameter space was proposed. This led to the feasibility study presented in chapter 5, where two instrumental measures reflecting assumed prototypical subjective traits were used to extract a reasonable amount of algorithm presets reflecting various weightings of these two traits. This manageable amount of settings can then be evaluated perceptually by the hearing aid user with a further development of one of the intuitive MUSHRA modifications from the study in chapter 3.

The feasibility study in chapter 5 also illustrates how the fitting procedure of modern multidimensional hearing devices to their individual users with their individual needs and preferences might work in the future (cf. Figure 5.1 on page 93): Within a first iterative *learning phase*, it is attempted to cover the relevant subjective traits by objective perceptual measures as precisely as possible leading to a convergence between objective and subjective assessment. In a subsequent *production phase*, selected (trait-based) objective measures are used to generate settings that should reflect the assumed traits as well as possible and should be targeted towards discriminating across different individual weightings of these traits. The user then is confronted with a manageable amount of alternative settings and can autonomously decide regarding personal preference with an intuitive test procedure. We suppose that due to the direct patient involvement and decision-making in the fitting process a higher user satisfaction is reached.

At this particular time, a successful prediction of preferred individual weightings of subjective traits is not known to the author. Furthermore, it is not even known if the concept of an individual weighting of subjective traits is valid at all. Still, it may be a good working concept for our current purposes. At next, 'critical' experiments should be performed to test the validity of this working concept. One could imagine that a prediction will be possible after collecting and analyzing data from many different test subjects. Their preferred settings might correlate with the individual preference regarding e.g., listening to high-fidelity music, that means a prediction of the individual traits and their respective weighting would be possible on the basis of a questionnaire requesting e.g. sound preferences.

For a successful hearing aid evaluation, both branches — instrumental and perceptual evaluation (see Figure 1.1 on page 4) — appear to be necessary, but are momentarily not sufficient to characterize human behavior. Both branches complement one another, i.e. to improve and to train the instrumental model-based measures, it is crucial to have barrier-free methods for perceptual measurements for a valid data collection. On the other hand, to improve the validity and mutual cross-consistency of experimental methods, better theoretical prediction methods need to be developed for creating 'crucial' experiments. Realistic and attribute-specific test scenarios (cf. chapter 4) are the indispensable basis for such experiments.

Bibliography

Abrams, H., Edwards, B., Valentine, S., and Fitz, K. A patient-adjusted fine-tuning approach for optimizing the hearing aid response. *Hearing Review*, 18(3):18–27, 2011.

Akeroyd, M. A. Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology*, 47(Suppl. 2):S53–S71, 2008.

Allen, J. B., Hall, J. L., and Jeng, P. S. Loudness growth in 1/2-octave bands (LGOB) — a procedure for the assessment of loudness. *Journal of the Acoustical Society of America*, 88(2):745–753, 1990.

Amlani, A. M. and Schafer, E. C. Application of paired-comparison methods to hearing aids. *Trends in Amplification*, 13(4):241–259, 2009. doi: 10.1177/1084713809352908.

ANSI 20S3.5-1997. *Methods for the Calculation of the Speech Intelligibility Index*. American National Standard, 1997.

ANSI/ASA S3.1-1999. *Maximum permissible ambient noise levels for audiometric test rooms*. American National Standards Institute, Washington, D.C., r2008 edition, 2008.

Baumgärtel, R. M., Hu, H., Krawczyk-Becker, M., Marquardt, D., Herzke, T., Coleman, G., Adiloğlu, K., Bomke, K., Plotz, K., Gerkmann, T., Doclo, S., Kollmeier, B., Hohmann, V., and Dietz, M. Comparing binaural pre-processing strategies II: Speech Intelligibility of Bilateral Cochlear Implant Users. *Trends in Hearing*, 19:1–18, 2015a.

Baumgärtel, R. M., Krawczyk-Becker, M., Marquardt, D., Völker, C., Hu, H., Herzke, T., Coleman, G., Adiloğlu, K., Ernst, S. M. A., Gerkmann, T., Doclo, S., Kollmeier, B., Hohmann, V., and Dietz, M. Comparing binaural pre-processing strategies I: Instrumental evaluation. *Trends in Hearing*, 19:1–16, 2015b.

Bentler, R. A., Niebuhr, D. P., Getta, J. P., and Anderson, C. V. Longitudinal study of hearing aid effectiveness. I: Objective measures. *Journal of Speech and Hearing Research*, 36(4):808–819, 1993a.

Bentler, R. A., Niebuhr, D. P., Getta, J. P., and Anderson, C. V. Longitudinal study of hearing aid effectiveness. II: Subjective measures. *Journal of Speech and Hearing Research*, 36(4):820–831, 1993b.

Bentler, R., Wu, Y.-H., Kettel, J., and Hurtig, R. Digital noise reduction: Outcomes from laboratory and field studies. *International Journal of Audiology*, 47:447–460, 2008.

Bentler, R. A. Effectiveness of directional microphones and noise reduction schemes in hearing aids: A systematic review of the evidence. *Journal of the American Academy of Audiology*, 16:473–484, 2005.

Beutelmann, R. and Brand, T. Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 120(1):331–342, 2006.

Beutelmann, R., Brand, T., and Kollmeier, B. Revision, extension, and evaluation of a binaural speech intelligibility model. *Journal of the Acoustical Society of America*, 127(4):2479–2497, 2010.

Binder, E. F., Storandt, M., and Birge, S. J. The relation between psychometric test performance and physical performance in older adults. *Journal of Geronology: MEDICAL SCIENCES*, 54A(8):M428–M432, 1999.

Boymans, M. and Dreschler, W. A. Audiologist-driven versus patient-driven fine tuning of hearing instruments. *Trends in Amplification*, 16(1):49–58, 2012.

Brand, T. and Kollmeier, B. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility. *Journal of the Acoustical Society of America*, 111(6):2801–2810, 2002.

Breithaupt, C., Gerkmann, T., and Martin, R. A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing. In *Inproceedings ICASSP '08*, 2008.

Brons, I., Houben, R., and Dreschler, W. A. Effects of noise reduction on speech intelligibility, perceived listening effort, and personal preference in hearing-impaired listeners. *Trends in Hearing*, 18:1–10, 2014. doi: 10.1177/2331216514553924.

Brooke, J. *Usability Evaluation in Industry*, chapter SUS - A quick and dirty usability scale, pages 189–194. Taylor & Francis, London, 1996.

Byrne, D., Parkinson, A., and Newall, P. *The Vanderbilt Hearing Aid Report II*, chapter Modified hearing aid selection procedures for severe/profound hearing losses. York Press, Parkton, MD, 1991.

Byrne, D., Dillon, H., Ching, T., Katsch, R., and Keidser, G. NAL-NL1 procedure for fitting nonlinear hearing aids: characteristics and comparisons with other procedures. *Journal of the American Academy of Audiology*, 12:37–51, 2001.

Cord, M. T., Surr, R. K., Walden, B. E., and Dyrlund, O. Relationship between laboratory measures of directional advantage and everyday success with directional microphone hearing aids. *Journal of the American Academy of Audiology*, 15: 353–364, 2004.

Cornelis, B., Moonen, M., and Wouters, J. Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel wiener filtering based noise reduction. *Journal of the Acoustical Society of America*, 131(6): 4743–4755, 2012.

Cornelisse, L. E., Seewald, R. C., and Jamieson, D. G. The input/output formula: A theoretical approach to the fitting of personal amplification devices. *Journal of the Acoustical Society of America*, 97(3):1854–1864, 1995.

Cubick, J., Santurette, S., Dau, T., and Laugesen, S. Influence of high-frequency audibility on the perceived distance of sounds. In *Forum Acusticum 2014, Krakow*. European Acoustics Association, 2014.

Dietz, M., Ewert, S. D. E., and Hohmann, V. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, 53(5):592–605, 2011.

Dijksterhuis, G. B. and Heiser, W. J. The role of permutation tests in exploratory multivariate data analysis. *Food Quality and Preference*, 6(4):263–270, 1995.

Dillon, H. *Hearing Aids*. Thieme, second edition, 2012.

Dillon, H., Zakis, J., McDermott, H., Keidser, G., Dreschler, W., and Convery, E. The trainable hearing aid: What will it do for clients and clinicians? *Hearing Journal*, 59:30–36, 2006.

Doclo, S., Gannot, S., Moonen, M., and Spriet, A. *Handbook on Array Processing and Sensor Networks*, chapter Acoustic beamforming for hearing aid applications, pages 269–302. Wiley, 2010.

Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing. *Journal of the Acoustical Society of America*, 111(6):2897–2907, 2002.

Durlach, N. I. Equalization and cancellation theory of binaural masking-level differences. *Journal of the Acoustical Society of America*, 35(8):1206–1218, 1963.

Elko, G. W. and Nguyen Pong, A.-T. A simple adaptive first-order differential microphone. In *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 169–172, 15-18 Oct 1995 1995. URL `http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=482983`.

European Committee for Standardization. *Acoustics — Statistical distribution of hearing thresholds as a function of age (ISO 7029:2000)*. European Committee for Standardization, 2001.

Ewert, S. D. AFC - a modular framework for running psychoacoustic experiments and computational perception models. In *Proceedings of the International Conference on Acoustics AIA-DAGA*, pages 1326–1329, Merano, Italy, 2013.

Festen, J. M. and Plomp, R. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of America*, 88(4):1725–1736, 1990.

Fletcher, H. and Galt, R. H. The perception of speech and its relation to telephony. *Journal of the Acoustical Society of America*, 22(2):89–151, 1950. doi: http://dx.doi.org/10.1121/1.1906605.

Forberg, A. and Neyer, F. J. Technology commitment can be enhanced: Evidence from an experimental study with older adults. *European Journal of Ageing*, submitted 2014.

Fredelake, S., Holube, I., Schlueter, A., and Hansen, M. Measurement and prediction of the acceptable noise level for single-microphone noise reduction algorithms. *International Journal of Audiology*, 51:299–308, 2012.

Gerkmann, T., Breithaupt, C., and Martin, R. Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. *IEEE Transactions on Audio, Speech and Language Processing*, 16(5):910–919, 2008.

Glende, S., Nedopil, C., Podtschaske, B., Stahl, M., and Friesdorf, W. *Erfolgreiche Lösungen durch Nutzerintegration. Ergebnisse der Studie "Nutzerabhängige Innovationsbarrieren im Bereich Altersgerechter Assistenzsysteme"*. VDE-Verlag, Berlin, Offenbach, 2011.

Granick, S. and Friedman, A. S. The effect of education on the decline of psychometric test performance with age. *Journal of Gerontology*, 22(2):191–195, 1967.

Griffiths, L. J. and Jim, C. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1):27–34, Jan 1982.

Grimm, G., Wendt, T., Hohmann, V., and Ewert, S. Implementation and perceptual evaluation of a simulation method for coupled rooms in higher order ambisonics. In *Proc. of the EAA Joint Symposium on Auralization and Ambisonics*, pages 27–32, Berlin, 2014.

Grimm, G., Herzke, T., Berg, D., and Hohmann, V. The Master Hearing Aid: A PC-based platform for algorithm development and evaluation. *Acta Acustica United with Acustica*, 92:618–628, 2006.

Grimm, G., Hohmann, V., and Kollmeier, B. Increase and subjective evaluation of feedback stability in hearing aids by a binaural coherence-based noise reduction scheme. *IEEE Transactions on Audio, Speech and Language Processing*, 17(7): 1408–1419, 2009. doi: 10.1109/tasl.2009.2020531.

Grimm, G., Coleman, G., and Hohmann, V. Realistic spatially complex acoustic scenes for space-aware hearing aids and computational acoustic scene analysis. In *16. Jahrestagung der Deutschen Gesellschaft für Audiologie*, Rostock, Germany, 2013.

Hagerman, B. and Olofsson, A. A method to measure the effect of noise reduction algorithms using simultaneous speech and noise. *Acta Acustica United with Acustica*, 90:356–361, 2004.

Hardy, D., Malléus, G., and Méreur, J.-N., editors. *Networks: Internet, Telephony, Multimedia*. Springer, 2002.

Hauth, C. and Brand, T. Ein blindes Modell zur Vorhersage der binauralen Sprachverständlichkeit. In *DAGA - 40. Jahrestagung für Akustik*, pages 1035–1038, Oldenburg, 2015.

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. An algorithm to improve speech recognition in noise for hearing-impaired listeners. *Journal of the Acoustical Society of America*, 134(4):3029–3038, 2013. doi: 10.1121/1.4820893.

Heute, U. *Speech and Audio Processing in Adverse Environments*, chapter Telephone-Speech Quality, pages 287–337. Springer, 2008.

Hougaard, S. and Ruf, S. Eurotrak i: A consumer survey about hearing aids in germany, france and the uk. *Hearing Review*, 18(2):12–28, 2011.

Huber, R. and Kollmeier, B. PEMO-Q — a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6):1902–1911, 11 2006.

Humes, L. E. Understanding the speech-understanding problems of the hearing impaired. *Journal of the American Academy of Audiology*, 2:59–69, 1991.

ITU-R — Radiocommunication Sector of ITU. *Recommendation ITU-R BS.1534-2 — Method for the subjective assessment of intermediate quality level of audio systems.* BS Series, Broadcasting service (sound). International Telecommunication Union, 2014a.

ITU-R — Radiocommunication Sector of ITU. *Report ITU-R BS.2300-0 — Methods for Assessor Screening.* BS Series, Broadcasting service (sound). International Telecommunication Union, April 2014b.

ITU-T — Telecommunication Standardization Sector of ITU. *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs — ITU T Recommendation P.862.* International Telecommunication Union, 2001.

Jenstad, L. M., Van Tasell, D. J., and Ewert, C. Hearing aid troubleshooting based on patients' descriptions. *Journal of the American Academy of Audiology*, 14(7): 347–360, 2003.

Kates, J. M. and Arehart, K. H. The hearing-aid speech quality index (hasqi). *J. Audio Eng. Soc.*, 58(5):363–381, 2010.

Kates, J. M. and Arehart, K. H. The hearing-aid speech quality index (HASQI) Version 2. *Journal of the Audio Engineering Society*, 62(3):99–117, 2014.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., and Kollmeier, B. Database of multichannel in-ear and behind-the-ear head-related and binaural

room impulse responses. *EURASIP Journal on Advances in Signal Processing*, 2009:1–10, 2009. doi: 10.1155/2009/298605.

Keidser, G. and Dillon, H. What's new in prescriptive fittings down under? In Palmer, C. and Seewald, R., editors, *Hearing Care for Adults 2006: Proceedings of the First International Adult Conference*, pages 133–142, Chicago, 2006. Phonak AG.

Kiessling, J., Müller, M., and Latzel, M. Fitting strategies and candidature criteria for unilateral and bilateral hearing aid fittings. *International Journal of Audiology*, 45(1):53–62, 2006.

Kochkin, S. Customer satisfaction with hearing instruments in the digital age. *Hearing Journal*, 58:30–37, 2005.

Kondo, K. *Subjective Quality Measurement of Speech, Signals and Communication Technology*. Springer-Verlag, Berlin, Heidelberg, 2012. doi: 10.1007/978-3-642-27506-7_2.

Kuk, F. K. *Strategies for selecting and verifying hearing aid fittings*, chapter Paired comparisons as a fine-tuning tool in hearing aid fittings, pages 125–150. Thieme, New York, 2nd edition, 2002.

Künemund, H. and Tanschus, N. M. The technology acceptance puzzle — results of a representative survey in lower saxony. *Zeitschrift für Gerontologie und Geriatrie*, 47:641–647, 2014. doi: 10.1007/s00391-014-0830-7.

Leijon, A., Eriksson-Mangold, M., and Bech-Karlsen, A. Preferred hearing aid gain and bass-cut in relation to prescriptive fitting. *Scandinavian Audiology*, 13(3): 157–161, 1984.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences of the United States of America*, 103(49):18866–18869, 2006.

Lorho, G., Le Ray, G., and Zacharov, N. eGauge — a measure of assessor expertise in audio quality evaluations. In *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, pages 191–200. Audio Engineering Society, Jun 2010. URL `http://www.aes.org/e-lib/browse.cfm?elib=15471`.

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Buechler, M., Dillier, N., Houben, R., Dreschler, W. A., Froehlich, M., Puder, H., Grimm, G., Hohmann, V., Leijon, A., Lombard, A., Mauler, D., and Spriet, A. Multicenter evaluation of signal enhancement algorithms for hearing aids. *Journal of the Acoustical Society of America*, 127:1491–1505, 2010. doi: 10.1121/1.3299168.

Maj, J.-B., Royackers, L., Wouters, J., and Moonen, M. Comparison of adaptive noise reduction algorithms in dual microphone hearing aids. *Speech Communication*, 48 (8):957–970, 2006.

Marzinzik, M. and Kollmeier, B. *Psychophysics, Physiology and Models of Hearing*, chapter Development and evaluation of single-microphone noise reduction algorithms for digital hearing aids, pages 279–282. World Scientific Publishing, 1999.

Marzinzik, M. *Noise reduction schemes for digital hearing aids and their use for the hearing impaired*. Shaker-Verlag, Aachen, Germany, 2000.

Mick, P., Kawachi, I., and Lin, F. R. The association between hearing loss and social isolation in older adults. *Otolaryngology — Head and Neck Surgery*, 150(3): 378–384, 2014.

Muralimanohar, R. K., Kronen, C., Arehart, K., Kates, J., and Pichora-Fuller, M. K. Quality of voices processed by hearing aids: Intra-talker differences. In *Proceedings of Meetings on Acoustics*, volume 19, 2013. doi: 10.1121/1.4800397.

Neher, T., Laugesen, S., Jensen, N. S., and Kragelund, L. Can basic auditory and cognitive measures predict hearing-impaired listeners' localization and spatial speech recognition abilities? *Journal of the Acoustical Society of America*, 130(3): 1542–1558, 2011.

Neyer, F. J., Felber, J., and Gebhardt, C. Entwicklung und Validierung einer Kurzskala zur Erfassung von Technikbereitschaft. *Diagnostica*, 58(2):87–99, 2012.

Ngo, K., Spriet, A., Moonen, M., Wouters, J., and Jensen, S. H. A combined multi-channel wiener filter-based noise reduction and dynamic range compression in hearing aids. *Signal Processing*, 92(2):417–426, 2012.

Norton, M. I., Mochon, D., and Ariely, D. The IKEA effect: When labor leads to love. *Journal of Consumer Psychology*, 22:453–460, 2012.

Parsa, V., Scollie, S., Glista, D., and Seelisch, A. Nonlinear frequency compression: Effects on sound quality ratings of speech and music. *Trends in Amplification*, 17 (1):54–68, 2013. doi: 10.1177/1084713813480856.

Peeters, H., Kuk, F., Lau, C., and Keenan, D. Subjective and objective evaluation of noise management algorithms. *Journal of the American Academy of Audiology*, 20(2):89–98, 2009. doi: 10.3766/jaaa.20.2.2.

Peters, R. W., Moore, B. C. J., and Baer, T. Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *Journal of the Acoustical Society of America*, 103(1):577–587, 1998.

Pfitzinger, H. R. Local speech rate as a combination of syllable and phone rate. In *5th International Conference on Spoken Language Processing*, volume 3, pages 1087–1090, Sydney, 1998.

Pfitzinger, H. R. Acoustic correlates of the IPA vowel diagram. In *Proceedings of the International Congress of Phonetic Sciences 2003*, volume 2, pages 1441–1444, Barcelona, 2003.

Plomp, R. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *Journal of the Acoustical Society of America*, 63(2):533–549, 1978.

Plomp, R. A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *Journal of Speech, Language, and Hearing Research*, 29:146–154, 1986.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *Journal of the Acoustical Society of America*, 120(6):3988–3997, 2006.

Rohdenburg, T., Hohmann, V., and Kollmeier, B. Subband-based parameter optimization in noise reduction schemes by means of objective perceptual quality measures. In *2006 International Workshop on Acoustic Echo and Noise Control*, Télécom Paris, France, 2006.

Roy, A. T., Jiradejvong, P., Carver, C., and Limb, C. J. Assessment of sound quality perception in cochlear implant users during music listening. *Otology & Neurotology*, 33(3):319–327, 2012a. doi: 10.1097/MAO.0b013e31824296a9.

Roy, A. T., Jiradejvong, P., Carver, C., and Limb, C. J. Musical sound quality impairments in cochlear implant (ci) users as a function of limited high-frequency perception. *Trends in Amplification*, 16(4):191–200, 2012b. doi: 10.1177/1084713812465493.

Schlich, P. GRAPES: A method and a SAS® program for graphical respresentations of assessor performances. *Journal of Sensory Science*, 9:157–169, 1994.

Schädler, M. R., Warzybok, A., Hochmuth, S., and Kollmeier, B. Matrix sentence intelligibility prediction using an automatic speech recognition system. *International Journal of Audiology*, 54:1–8, 2015. doi: 10.3109/14992027.2015.1061708.

Simmer, K. U., Bitzer, J., and Marro, C. *Microphone Arrays: Signal Processing Techniques and Applications*, chapter Post-Filtering Techniques, pages 39–60. Springer-Verlag, 2001.

Simonsen, C. S. and Legarth, S. V. A procedure for sound quality evaluation of hearing aids. *Hearing Review*, 17(13):32–37, 2010.

Tversky, A. and Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

Van den Bogaert, T., Doclo, S., Wouters, J., and Moonen, M. Speech enhancement with multichannel wiener filter techniques in multimicrophone binaural hearing aids. *Journal of the Acoustical Society of America*, 125(1):360–371, 2009.

Van Veen, B. and Buckley, K. Beamforming: a versatile approach to spatial filtering. *ASSP Magazine, IEEE*, 5(2):4–24, April 1988.

Völker, C., Grimm, G., Vormann, M., Hohmann, V., Kollmeier, B., and Ernst, S. M. A. How to derive the best parameter combination for complex hearing aid algorithms? — prototypic evaluation of a binaural coherence-based noise reduction algorithm. *International Journal of Audiology*, submitted, 2016a.

Völker, C., Grimm, G., and Ernst, S. M. A. Objective evaluation of binaural noise reduction schemes. In *Proceedings of AIA-DAGA 2013 Merano*, pages 1110–1113, 2013.

Völker, C., Bisitz, T., Huber, R., Kollmeier, B., and Ernst, S. M. A. Modifications of the multi stimulus test with hidden reference and anchor (MUSHRA) for use in audiology. *International Journal of Audiology*, submitted, 2016b.

Wagener, K., Brand, T., and Kollmeier, B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests. *Zeitschrift für Audiologie*, 38(2):44–56, 1999a.

Wagener, K., Brand, T., and Kollmeier, B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests. *Zeitschrift für Audiologie*, 38(3):86–95, 1999b.

Wagener, K., Kühnel, V., and Kollmeier, B. Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. *Zeitschrift für Audiologie*, 38(1):4–15, 1999c.

Walden, B. E., Surr, R. K., Cord, M. T., Edwards, B., and Olson, L. Comparison of benefits provided by different hearing aid technologies. *J Am Acad Audiol*, 11(10): 540–560, 2000.

Warzybok, A., Rennies, J., Brand, T., and Kollmeier, B. Prediction of binaural speech intelligibility in normal-hearing and hearing-impaired listeners: a psychoacoustically motivated extension. In *DAGA - 40. Jahrestagung für Akustik*, pages 351–352, Oldenburg, 2014.

Wittkop, T. and Hohmann, V. Strategy-selective noise reduction for binaural digital hearing aids. *Speech Communication*, 39(1-2):111–138, 1 2003.

Wong, L. L. N. Evidence on self-fitting hearing aids. *Trends in Amplification*, 15(4): 215–225, 2011.

World Health Organization. Deafness and hearing loss — fact sheet 300, 2015. URL `http://www.who.int/mediacentre/factsheets/fs300/en/`.

Danksagung

Ich möchte mich ganz herzlich bei allen bedanken, die dazu beigetragen haben, dass diese Promotionsschrift letztendlich zustande gekommen ist.

Dabei ist zunächst Birger zu nennen, bei dem ich mich bedanken möchte, dass ich innerhalb der Medizinischen Physik arbeiten und promovieren konnte.

Ich bedanke mich besonders bei Stephan Ernst für seine hervorragende und ermutigende Betreuung und Begleitung über die Jahre. Danke für Deine Zeit, die Du investiert hast.

Den Ko-Autoren der einzelnen Kapitel gilt mein aufrichtiger Dank: Ania, Thomas und Giso. Großer Dank gilt Matthias Vormann, Graham, Tobias und Felix für Hilfe bei Auswertungen, beim MHA oder Kalibrierungen.

Für hilfreiche Kommentare an den Skripten oder Bemerkungen bei den Medi-Seminaren bin ich Rainer Huber, Steven van de Par, Volker Hohmann und Tom Brand dankbar.

Mein Dank gilt zudem Steffen, Arne und Regina — für die mittäglichen Ausflüge zur Mensa.

Ich danke Beate, meinen Eltern, der sonstigen Familie und meinen Freunden für jegliche Unterstützung in diesem Prozess.

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe.

Diese Arbeit wurde weder in ihrer Gesamtheit noch in Teilen einer anderen wissenschaftlichen Hochschule zur Begutachtung in einem Promotionsverfahren vorgelegt.

Das Kapitel 2 wurde bereits in einer Fachzeitschrift veröffentlicht, wie es an der entsprechenden Stelle vermerkt ist. Die Kapitel 3 und 4 sind zur Veröffentlichung im *International Journal of Audiology* eingereicht worden. Die Modellierung der empirischen Daten mit dem binauralen Sprachverständlichkeitsmodell BSIM in Kapitel 2 wurde von Dr. Anna Warzybok durchgeführt und beschrieben. Die Modifikation MUSHRA simple in Kapitel 3 wurde von Thomas Bisitz entwickelt. Im Übrigen wurden die Methoden und Experimente, soweit im Text nicht anders angegeben, von mir entwickelt bzw. durchgeführt.

Christoph Völker