EXPERIMENTAL AND MODEL-BASED INVESTIGATION OF INDIVIDUAL SOUND PERCEPTION AND LISTENING PREFERENCES

Von der Fakultät für Mathematik und Naturwissenschaften

der Carl von Ossietzky Universität Oldenburg

zur Erlangung des Grades und Titels einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

angenommene Dissertation von

Aleksandra Maria Kubiak

Gutachter:	Prof. Dr. Dr. Birger Kollmeier
Zweitgutachter:	Prof. Dr. Steven van de Par
Tag der Disputation:	2. Dezember 2019

ABSTRACT

Playback personalization is a powerful tool used for increasing consumers' satisfaction. However, origins, as well as stability of listening preferences, are topics we still lack understanding of. The main goals of this thesis were to investigate how individual factors (such as age and hearing abilities) may influence listening preferences and to systematically investigate the stability of such preferences across different audio processing schemes and listening conditions.

In the first experiment a group of 30 subjects - highly heterogeneous with respect to their age, pure-tone average thresholds (PTA) and speech intelligibility in standardized, simple listening conditions - underwent speech reception threshold (SRT) measurements under complex listening conditions including energetic and informational masking (IM), spatial unmasking and/or possibility to exploit masker envelope fluctuations. Additionally, a binaural speech intelligibility model (BSIM, Beutelmann et al., 2010) was used to predict SRTs in those conditions. Results showed that despite the highly diversified subjects' group – two individual factors (PTA and speech intelligibility performance) were relatively good predictors of SRTs in all conditions except the one highest in IM. For this condition, age explained most of the variance in experimental SRT results indicating a special role of that factor in susceptibility to IM. BSIM simulations brought reliable outcomes and proved its applicability for clinical purposes (except for conditions high in IM).

The second experiment investigated listening preferences of the same group of subjects and in the same listening conditions regarding four signal modification schemes: i) linear gain, gain at the cost of ii) clipping distortions or iii) compression distortions, and iv) frequency-shaping. Subjects showed high test/retest reliability in their preferences for target speech modifications in all conditions and for all signal modifications but frequency shaping. Moreover, results showed that subjects can be consistently categorized in scenarios ii) and iii) either as "noise haters" – allowing distortions in the target speech signal to avoid noise, "distortion haters" – allowing noise to avoid distortions, or "indifferent" – showing no such visible preference (as first proposed by Völker et al., 2018). Such categorization remained stable across all maskers, spatial conditions, and types of distortions. Additionally, a quick and reliable test method was proposed to differentiate subjects upon the abovementioned trait. Correlation

of listening preferences with individual factors and SRT data (form the previous experiment) showed that preferences in complex listening conditions can be relatively well predicted by SRTs obtain in same listening conditions.

The third experiment investigated the influence of physical fatigue on listening preferences in music and addressed the possible negative impact on hearing thresholds that listening to loud music during sport may have. Ten normal-hearing subjects took part in the study and their results were compared to the online survey data (N=138). The outcomes of both indicated a common trend of increased sensibility to high-level and/or high-frequency sounds with growing fatigue. Similarly, experimental data suggested a possible existence of common preference shift (in line with abovementioned increased sensibility) that could be observed among 70% of subjects. Potentially dangerous impact of sport exercises accompanied with loud music on hearing thresholds was not confirmed – possibly due to personalized and, as a result, much lower signal levels employed here than in the study reporting such effects (Vittitow et al., 1994).

Taken together, the results presented in this thesis indicated several individual factors as good predictors for SRT performance in complex listening conditions and tested the applicability of BSIM model in a great majority of such conditions obtaining reliable predictions. Results collected within the second experiment proved general stability of listening preferences under the abovementioned conditions and confirmed the existence of stable preference profiles in listening scenarios involving distortions - that let for subjects' classification regarding noise/distortion tolerance (proposing efficient method to perform such classification). A novel experiment investigated an issue of listening preferences' change under physical fatigue and indicated a consistent change of such preferences. Results of this thesis can be used to improve fitting of hearing support devices and playback personalization algorithms for increased comfort of their users.

ZUSAMMENFASSUNG

Die Personalisierung der Wiedergabe ist ein leistungsstarkes Instrument zur Steigerung der Verbraucherzufriedenheit. Die Herkunft und die Stabilität der Hörpräferenzen sind jedoch Themen, für die wir noch wenig Verständnis haben. Die Hauptziele dieser Arbeit waren die Untersuchung, wie individuelle Faktoren (wie Alter und Hörfähigkeiten) die Hörpräferenzen beeinflussen können, und die systematische Untersuchung der Stabilität solcher Präferenzen über verschiedene Audioverarbeitungsbedingungen und Hörbedingungen hinweg.

Im ersten Experiment wurde eine Gruppe von 30 Probanden - heterogen in Bezug auf Alter, Reinton-Durchschnittsschwellen (PTA) und Sprachverständlichkeit unter standardisierten, einfachen Hörbedingungen - unter komplexen Hörbedingungen, einschließlich energetischer und informationeller Maskierung, räumlicher Demaskierung und / oder der Möglichkeit die Schwankung der maskierten Hüllkurve auszunutzen einer Sprachempfangsschwellenmessung (SRT) unterzogen. Zusätzlich wurde ein binaurales Sprachverständlichkeitsmodell (BSIM, Beutelmann et al., 2010) verwendet, um SRTs unter diesen Bedingungen vorherzusagen. Die Ergebnisse zeigten, dass trotz der stark diversifizierten Probandengruppe - zwei einzelne Faktoren (PTA und Sprachverständlichkeit) waren relativ gute Prädiktoren für SRTs unter allen Bedingungen außer dem höchsten in der Informationsmaskierung (IM). Für diese Bedingung erklärt das Alter die meisten der Varianz in den experimentellen SRT-Ergebnissen, was auf eine besondere Rolle dieses Faktors bei der Anfälligkeit für IM hinweist. BSIM-Simulationen brachten zuverlässige Prognosen und bewiesen ihre Anwendbarkeit für klinische Zwecke, außer für Bedingungen mit hohem IM - Wert.

Das zweite Experiment bestand aus einer Studie, die die Hörpräferenzen derselben Probandengruppe unter gleichen Hörbedingungen in Bezug auf vier Signalmodifikationsschemata: i) lineare Verstärkung, Verstärkung auf Kosten von ii) Clipping-Verzerrungen oder iii) Kompressionsverzerrungen und iv) Entzerrung. Die Probanden zeigten eine hohe Test-/Retestzuverlässigkeit in ihren Präferenzen für Zielsprachmodifikationen unter allen Bedingungen und für alle Signalmodifikationen außer Entzerrung. Darüber hinaus die Ergebnisse zeigten, dass die Probanden in den Szenarien ii) und iii) entweder als "Rauschhasser" - also Verzerrungen gegenüber Rauschen bevorzugten -, " Verzerrungshasser "

- Rauschen gegenüber Verzerrungen bevorzugten - oder "gleichgültig" - ohne eindeutige Präferenz kategorisiert werden können (wie erstmals von Völker et al., 2018 vorgeschlagen). Eine solche Kategorisierung blieb über alle Masken, räumlichen Bedingungen und Arten von Verzerrungen hinweg stabil. Zusätzlich wurde eine schnelle und zuverlässige Testmethode vorgeschlagen, um die Probanden nach einem der oben genannten Präferenzen zu unterscheiden. Die Korrelation der Hörpräferenzen mit einzelnen Faktoren und SRT-Daten (aus dem vorherigen Experiment) zeigte, dass Präferenzen unter komplexen Hörbedingungen mit den SRTs von gleichen Hörbedingungen relativ gut vorhergesagt werden können.

Das dritte Experiment untersuchte den Einfluss von körperlicher Müdigkeit auf die Hörpräferenzen in der Musik und befasste sich mit den möglichen negativen Auswirkungen auf die Hörschwellen, die das Hören von lauter Musik beim Sport haben kann. Zehn normalhörende Probanden nahmen an der Studie teil und ihre Ergebnisse wurden mit den Daten der Online-Umfrage verglichen (N=138). Die Ergebnisse beider Studien/Umfragen zeigten einen gemeinsamen Trend zu einer erhöhten Sensibilität gegenüber hochfrequenten Klängen und/oder Klängen, die einen hohen Pegel aufweisen, mit zunehmender Müdigkeit. Ebenso deuteten experimentelle Daten auf eine mögliche Existenz einer gemeinsamen Präferenzverschiebung (im Einklang mit der oben genannten erhöhten Sensibilität) hin, die bei 70% der Probanden beobachtet werden konnte. Möglicherweise gefährliche Auswirkungen von Sportübungen mit lauter Musik auf die Hörschwellen wurden nicht bestätigt (möglicherweise aufgrund personalisierter und damit deutlich niedrigerer Signalpegel als in der Studie, die über solche Effekte berichtet (Vittitow et al., 1994).

Insgesamt zeigten die in dieser Arbeit vorgestellten Ergebnisse mehrere individuelle Faktoren als gute Prädiktoren für die SRT-Leistung unter komplexen Hörbedingungen und bestätigten die Anwendbarkeit des BSIM-Modells bei einer großen Mehrheit dieser Bedingungen, um zuverlässige Vorhersagen zu erhalten. Experimentelle Ergebnisse bewiesen die allgemeine Stabilität der Hörpräferenzen unter den oben genannten Bedingungen und bestätigten das Vorhandensein eines stabilen Präferenzprofils in Hörszenarien mit Verzerrungen (das eine Klassifizierung der Probanden in Bezug auf Rausch- und Verzerrungstoleranz ermöglicht). Ein neuartiges Experiment untersuchte die Veränderung der Hörpräferenzen unter körperlicher Müdigkeit und zeigte eine konsistente Veränderung dieser Präferenzen an. Die Ergebnisse dieser Arbeit können eine bessere Anpassung der Hörgeräte

und Algorithmen zur Personalisierung der Wiedergabe für mehr Komfort für ihre Benutzer unterstützen.

GLOSSARY

2D	two-dimensional
ACR	absolute category rating
adjS	subjects could introduce changes only to the target speech signal while the maskers remained unchanged
adjSN	changes introduced by the subjects influenced both target and masker signals in the same way
AFC	alternative-forced choice
AI	articulation index
ANOVA	analysis of variance
ANSI	American National Standards Institute
ATM	asynchronous transfer mode
BPM	beats per minute
BSIM	binaural speech intelligibility model
cl	co-located
DIN	Deutsches Institut für Normung
DRC	dynamic range compressor
DSCQS	double-stimulus continuous quality scale
DSIS	double-stimulus impairment scale
DTT	digit triplet test
EC	equalization-cancellation
EMA	ecological momentary assessment
EPSM	envelope power spectrum model

EQ	equalization
ESII	extended speech intelligibility index
EU	European Union
FS	full scale
Goesa	Göttingen sentence test
GUI	graphical user interface
HRIR	head-related room impulse response
IEC	International Electrotechnical Commission
IM	informational masking
IP	internet protocol
ISO	International Organization for Standardization
ITU	International Telecommunication Union
ITU-R	ITU radiocommunication sector
ITU-T	ITU telecommunication standardization sector
MAE	mean absolute error
MOS	mean opinion score
mr-sEPSM	multi-resolution speech-based envelope power spectrum model
MT	multi-talker
MTI	masker type impact
MUSHRA	multiple stimuli with hidden reference and anchor
NAL	The National Acoustic Laboratories
NAL-NL2	NAL prescription procedure (second generation) for fitting wide dynamic range compression hearing aids
NAL-R	prescription procedure for fitting linear hearing aids

NR	noise reduction
PEAQ	perceptual evaluation of audio quality
PEMO-Q	perception model with quality assessment
PESQ	perceptual speech quality measure
PTA	pure-tone-average
R	linear correlation
R^2	squared linear correlation
RMS	root mean square
RMSE	root-mean-square error
RPE	rate of perceived exertion
S1	first session
S1AU1	first audiogram measurement during the first session
S1AU2	second audiogram measurement during the first session
S1PM	preferences' measurement during the first session
S2	second session
S2AU	audiogram measurement during the second session
S3	third session
S3AU	audiogram measurement during the third session
S3PM1	first preferences' measurement during the third session
S3PM2	second preferences' measurement during the third session
S3PM3	third preferences' measurement during the third session

sEPSM	speech-based envelope power spectrum model
SII	speech intelligibility index
SNR	signal-to-noise ratio
sp	spatially separated
SRM	spatial release from masking
SRT	speech reception threshold
SSN	speech-shaped noise
STI	speech transmission index
STOI	short-time objective intelligibility measure
TT	two-talker
TTS	temporary threshold shift
UI	user interface
WL	workload
ϱ	Spearman’s rank correlation

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. MOTIVATION

Inter-personal differences in abilities and preferences motivate the development of personalization techniques allowing for better adjustment of the end product to customer's needs. Examples are easy to find in every aspect of our lives - especially in the field of technology and personal devices. From improving the mobile listening experience by utilizing information on environmental noise (Walton et al., 2018) to personalized digital resources helping school children get more involved into reading (Kucirkova and Flewitt, 2018) - the aim is always to enhance comfort and access potential of individuals by accurately meeting their needs.

This work investigates personalization in the field of audio, involving adjustment of audio signal processing to specific abilities and preferences of an individual. Such adjustment can be based on different sets of factors. It can rely purely on environmental ones, employing variables from the current listening situation, e.g., target signal (music or speech), different maskers and their spatial constellation, etc. In addition, it may also incorporate data on personal hearing abilities and listening preferences – user specific profile - like it is the case in hearing aids. Rehabilitative audiology employs, however, detailed, developed-over-years clinical measures as well as hearing aids' fitting procedures enabling best, individualized support (Keidser et al., 2011; Kollmeier et al., 2015).

One of the research goals of this work is to find an efficient way to personalize audio processing based on both: current listening conditions and user's specific profile but without the use of hearing aids. The idea relies on making the best use of consumer's audio devices and on addressing a very specific group of users – those without (or with not yet treated) hearing loss. Another constraint is the limited amount of data available to profile each such consumer (based on personal factors such as hearing abilities, for instance) and the data acquisition process itself – that has to be both swift and playful if it is supposed to be implemented on e.g., mobile phones and done without or with minimal supervision. Perspectives are, nonetheless, promising – offering hearing support and much higher listening comfort without a need of hearing aid fitting. Employing devices already used by the consumer (e.g., mobile phone, car

audio system, TV) can significantly improve lives of millions. **Gaining more knowledge about what set of personal factors could play a decisive role in shaping user preferences under given listening conditions and how this set of factors may predefine audio processing needed is, therefore, one of the goals of this thesis.**

The idea of utilizing individual factors to create more accurate personalization of an audio signal is not new and has been investigated for decades, especially in the field of hearing aids research. However, this research almost exclusively targeted hearing-impaired people, while the issue of playback personalization based on individual factors of normal-hearing persons, or those with (untreated) mild to moderate hearing loss that do not use hearing aids (also called "subclinical population"), remains largely unaddressed. The term "normal" or "impaired" hearing refers to an individual level of hearing pure tones compared to international standards (ISO 389-1: 2017). Such pure tones are measured at the better hearing ear and are reported as absolute hearing thresholds (in dB HL) averaged across several audiometric frequencies (e.g., 500, 1000, 2000 and 4000 Hz). To put this problem into perspective, it is estimated that about 10.1 million people in Germany alone suffer from hearing loss (Euro Track, 2018), so their average hearing threshold is higher than what the norm describes as normal, yet only 37% of them (approx. 3.7 million) have hearing aids (among those aided individuals 6% do not use their device(s) and further 12 % wear them less than an hour a day). This vast untreated part of a population could potentially benefit from some type of individualized audio processing provided by their current audio devices. This would, in turn, prolong their pre-hearing aids state and still significantly improve their quality of life. Exemplary scenarios could include dedicated gain, equalization or compression algorithms implemented on devices such as TVs, phones or car infotainment systems considering the personal needs of a user, e.g. to overcome hearing loss by improving speech reception. **There is a visible need to understand the link between hearing abilities and listening preferences of normal-hearing individuals and those with not yet treated hearing loss. This constitutes another goal of the thesis.**

One of the first challenges when investigating personal preferences is a proper way to capture them. Measuring a sensation is a first step in measuring preferences and conducting an experimental investigation on audio personalization requires a careful selection of a data collection procedure due to its crucial influence on results obtained (e.g., learning effect or sensitivity increase). Additionally, the choice of a method is often constrained by the time

available for measurements. Here, to the quickest methods belong those in which the value of interest can be elicited within a single trial - such as methods of adjustments, methods of tracking and magnitude estimation. In the first method, subject has a full control over the adjustment of the stimulus within the range given. In the second, the subject controls only the direction in which the stimulus varies but not the size of such changes. In magnitude estimation method each magnitude of a stimulus under test has a specific number assigned to it creating a scale in some defined dimension (e.g., loudness, sharpness, etc.). The abovementioned methods, however, though quick, assume that the subject has access to her or his internal representation of the respective stimulus magnitude and can rate this representation to produce a value of interest (similar to a psychometric function).

Another group of methods consists of forced-choice procedures, such as "yes-no", where the subject has to decide whether the signal occurs in the current trial or not (hence the procedure is also referred to as "one interval two alternative forced-choice"). Similarly, in "two-interval forced choice" procedure not one but two intervals are presented to the subject who decides e.g., whether the signal was present in the first or second one. This procedure can be modified to involve more intervals if needed – being referred to as "multiple alternative forced-choice". Similarly, a procedure of comparison of stimulus pairs requires subjects to compare the differences between two pairs of stimuli (e.g., AB and CD) where each pair represents difference along one of two dimensions (e.g. pair AB represents a difference in loudness while pair CD – in sharpness). All the above-mentioned methods do not involve any adaptation based on the user's responses. This contrasts to "adaptive procedures" where presented stimuli depend on the responses already provided by the subject in the preceding trials. In such methods, a given number of correct responses would lead to an increase in task difficulty while incorrect ones would cause the difficulty to decrease. The value of interest is elicited by averaging some amount of last reversals (Fastl and Zwicker, 2007). The majority of the abovementioned methods aims at measuring sensation across a given dimension or a set of dimensions (like loudness, sharpness, pitch).

Another group of methods investigates the overall quality of experience, hence measures a delight or annoyance of a customer concerning the entire service or the processing provided. This holistic concept originates from the field of telecommunication as a consequence of introducing new technologies (e.g., voice over IP, voice over ATM) that no longer fulfil the principles of (nearly) linear time-invariant systems and introduce new types of distortions (e.g.,

packet loss or variable delay). Perceptual measurements of such systems have to rely on a quality evaluation of the output of the system under test and employ scales for quality rating. The most important example of these subjective procedures recommended by the International Telecommunication Union (ITU) for speech material testing is the Mean Opinion Score (MOS, ITU-T Rec. P.800, 1996 – speech non-conversational; ITU-T Rec. P.800, 2008 – speech conversational) using predefined standard rating scales such as absolute category rating (ACR), double-stimulus impairment scale (DSIS) or double-stimulus continuous quality scale (DSCQS); Tominaga et al. (2010) found very high correlations between MOS obtained by these different methods. Another method recommended by ITU for conducting codec listening tests is a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA, ITU-R BS.1534-3 - employed i.e., for output quality assessment from lossy audio compression algorithms). The MUSHRA method has the advantage of letting the subject compare between multiple stimuli directly as they are displayed. Also, the time taken to perform the test is shorter than in the case of experiments employing MOS methodology. (ITU-R BS.1116-3).

While a way to measure preferences was a first decision undertaken, the remaining question was how to most efficiently use the data obtained? Since a comparatively large and diversified group of subjects (both normal-hearing and hearing-impaired) was planned to be invited to take part in the listening tests which involved complex listening conditions - the data collected could serve an additional purpose, i.e., be compared to an objective estimate using a speech intelligibility prediction model that aims at reproducing functions of the human auditory system. Such an evaluation of model predictions against subjective data allows for testing or improving the respective model accuracy and should bear the future potential of replacing lengthy and costly measurements of human subjects with a quick estimation. Here, another challenge was to decide which model to use in the current study that would capture all (or a majority of) the parameters present in the scenarios.

Modelling of speech perception has a relatively long history reaching back to monaural speech intelligibility predictions of the Articulation Index (AI, ANSI, 1969; French and Steinberg, 1947; Kryter, 1962) later replaced by the Speech Intelligibility Index (SII, ANSI, 1997) that shares a similar working principle. The AI can be described as a function of intensities of speech and unwanted sounds received by the ear, while the SII considers the signal-to-noise ratio (SNR) in narrow frequency bands with different importance (weighting function). Nevertheless, the SII model was originally limited to the stationary noises' scenarios

due to its use of the long-term speech and noise spectra. That renders it inaccurate for predictions of speech intelligibility in non-stationary noises that are commonly present in every-day life. A direct improvement of SII model was its successor, an Extended Speech Intelligibility Index – ESII (Rhebergen and Versfeld, 2005; Rhebergen et al., 2006) - capable of predicting SRTs for non-stationary noise conditions. These conditions usually lead to a much better speech intelligibility performance for normal-hearing subjects due to noise fluctuations – occurrences of "dips" when masking is reduced. This is often not the case for hearing-impaired subjects whose ability to perform "dip listening" may be reduced (Festen and Plomp, 1990; Versfeld and Dreschler, 2002). Another model that was considered due to its design to predict the influence of spatial unmasking and the abovementioned "dip listening", and its capability of "effective" modeling of speech processing stages of binaural hearing, is the BSIM. This model employs a gammatone filterbank to analyze binaural speech and noise signals followed by a short-term binaural equalization-cancelation stage (EC, Durlach, 1963; Wan et al., 2015). Binaural processing errors are accounted for to adapt to the imperfect processing of the human auditory system. The model outcome is a (binaurally improved) SNR measure used as an input for the SII model (working as a short-term speech intelligibility index, similar to ESII). The SRT is calculated by setting a fixed SII value and varying the SNR measure until the resulting SII equals the reference SII. The model can account for the individual hearing loss by adding to its processing an additional external noise derived from the individual audiogram and was extensively tested in multiple complex listening conditions on groups of both normal-hearing and hearing-impaired participants showing high correlation with experimental data for the majority of conditions (Völker et al., 2015; Ewert et al., 2017; Rennies et al., 2011, 2014).

Since the broad scope of this thesis involves topics of audio quality within the field of personal audio devices – other groups of models were considered were those explicitly designed to predict the effect of non-linear processing of noisy speech - e.g., an outcome of hearing aid processing. One of the first models proposed was the speech transmission index (STI, Steeneken and Houtgast, 1980; IEC, 2003) which offers predictions of speech undergoing a simple non-linear degradation such as clipping. The STI successfully predicts intelligibility in noisy and reverberant conditions (e.g., Houtgast et al., 1980; Steeneken and Houtgast, 1980). Its main principle is to estimate the reduction in amplitude modulation of a processed signal in comparison to the clean speech signal across those frequency bands contributing mostly to the speech intelligibility. However, multiple studies reported shortcomings of the model if the noisy

speech is non-linearly processed, e.g. using spectral subtraction (e.g., Ludvigsen et al., 1993; Dubbelboer and Houtgast, 2008), envelope compression (e.g., Drullman, 1995; Hohmann and Kollmeier, 1995; Rhebergen et al., 2009) and deterministic envelope reduction (Noordhoek and Drullman, 1997). A further improvement in the field of noisy speech processing was offered by Taal et al. (2010) with a Short-Time Objective Intelligibility measure (STOI). This model employs an intermediate intelligibility measure on a short time window of time-frequency weighted regions followed by the cross-correlation of the temporal envelopes of clean and noisy speech. During evaluations, the STOI outperformed the abovementioned STI procedure and showed a high correlation with subjective speech intelligibility data (Kjems et al., 2009). Another group of models offering further improvement of speech intelligibility predictions of processed noisy speech are envelope power spectrum models (EPSM, sEPSM and mr-sEPSM), sharing the same principle structure with the classical power-spectrum models of masking (Fletcher, 1940; Fastl and Zwicker, 2007, Patterson and Moore, 1986).The first representative from this group is an envelope power spectrum model (EPSM, Dau et al., 1999; Ewert and Dau, 2000) designed to predict modulation detection and masking. The speech-based envelope power spectrum model (sEPSM) presented by Jørgensen and Dau (2011) is an extended EPSM model prepared to predict speech intelligibility by estimating the envelope power signal-to-noise ratio after the modulation-frequency-selective processing stage of the EPSM model. sEPSM was successful in predicting speech intelligibility for conditions including additive noise, reverberation and non-linear processing including spectral subtraction. However, the main limitation of the model is its limited functionality restricted to stationary maskers (due to long-term integration of the power envelope). This resulted in yet another extension – the multi-resolution speech-based envelope power spectrum model (mr-sEPSM) - where the estimation of the envelope power signal-to-noise ratios takes place in temporal segments with a modulation-filter dependent duration. This resulted in a high versatility of the model being capable of estimating speech intelligibility in both stationary and fluctuating interferers for conditions of additive noise, reverberation, and spectral subtraction.

The development and standardization of new telephony technologies like Voice over Internet Protocol (IP), Voice over Asynchronous Transfer Mode (ATM) or voice over mobile required yet another new type of model for predicting speech intelligibility. One reason was that these technologies introduced new types of distortions such as packet loss, variable delay, cell loss, frame repeat, front-end clipping or comfort noise generation. As a consequence, the

objective quality measurement techniques described above (relying grossly on principles of nearly linear time-invariant systems) had to be replaced with a perception-based approach that consisted of feeding the system under test with real-world signals (e.g., speech with various maskers) and measuring the perceptual quality at its output. The Perceptual Speech Quality Measure (PESQ) was the first perceptual quality measurement of narrow-band telephony (300-3400 Hz) that became an international standard (ITU-T P.861 (from 1996) for the assessment of speech codecs only, currently suspended). The corresponding procedure for wide-band audio signals (20-20000 Hz) is the Perceptual Evaluation of Audio Quality (PEAQ) measure that became an ITU-R recommendation (ITU-R BS.1387-1) two years later. The general working principle of abovementioned models relies on mapping both original and degraded signal onto an internal representation (i.e., the assumed output of the peripheral auditory system) with the use of a perceptual model. Subsequently, the obtained difference between those signals is further assessed by a cognitive model to predict the perceived speech quality of the signal that was degraded. Nevertheless, several years after the introduction of the PEAQ model, a new method for the objective audio quality assessment using a model of auditory perception (PEMO-Q) was proposed (Huber and Kollmeier, 2006) aiming at accurately predicting the perceived audio quality for any type of distortion and audio signal type. It employs a single, broadly validated auditory model (e.g., Dau et al., 1996 part I & II; Derleth et al., 2001) instead of averaging across multiple models, (as in the case of PEAQ) and is capable of covering a wide range of tasks (i.e., predicting both very small and severe quality degradations for various types of audio signals). The method - sharing similar working principle with PEAQ model - employs a much simpler cognitive model. Due to its lower degree of specialization, the PEMO-Q model is assumed to be broader applicable than PEAQ.

With an increasing number of complex signal processing algorithms enclosed in a single device (e.g., a hearing aid) it became apparent that such a device could be capable of personalizing its parameters, tailoring processing to the particular needs of an end-user. This requires yet another type of models, namely such that would be capable of connecting the outcome of an objective measure (e.g. like PESQ, or SII) with a subset of personal factors (e.g., age, or audiogram) and/or subjective "traits" referring to an individual need for extra amplification, or willingness to accept distortions in the signal in return for suppressing unwanted noise. Such concept requires a complex set of "weights" being an individual´s description of how a combination of objective measures is able to predict subjective needs.

Such an individualized set of parameters could indicate how complex signal processing algorithms might be tuned or combined to produce the optimized outcome for the particular end-user. This conceptual approach was recently proposed by Völker et al. (2018) to predict and adjust for inter-individual differences in hearing aid fitting. This approach relies on the assumption that the above mentioned "personal traits" remain stable over time within given conditions and let for classification of subjects along the dimensions of noise-versus-distortion tolerance. A previous study (Marzinzik, 2000) also suggests that some subjects may show higher tolerance to distortions caused by noise reduction algorithms, hence could be referred to as "noise haters", while others could willingly accept noise present in the signal only to avoid artifacts ("distortion haters"). This may also be related to works of Luts et al. (2010) who investigated the relation between basic audiological and cognitive factors and noise/distortion tolerance and (surprisingly) found a negligible influence of PTA or working memory capacity on subjective preferences for binaural noise reduction algorithms' processing. Similarly, Neher et al. (2014) investigated the relation of several acoustical, audiological, and cognitive measures with speech recognition performance and noise reduction preferences finding reasonably high test/retest accuracy, yet no measure was a strong predictor for noise-reduction strength expected on a personal basis. Reports from Hawkins and Naidoo (1993) and Dawson et al. (1990) also suggest an unclear relation between speech recognition performance or overall quality ratings and type of distortions – namely those caused by compression versus caused by peak clipping – among hearing-impaired subjects. Additionally, there is an extensive body of research showing how significantly one factor, as basic as loudness perception, can differ not only between hearing-impaired and normal-hearing subjects but also within the normal-hearing group (Brand and Hohmann, 2001; Oetting et al., 2014), indicating that the factors underlying individual loudness perception are not yet fully understood. Despite multiple reports showing the difficulty to link personal factors to speech recognition performance or preference for speech processing in various conditions, it has also been proven that subjects strongly prefer personalized settings to general ones (Walden et al., 2004; Walden et al., 2005; Cord et al., 2007). Yet another important limitation for creating "personal traits" driven models or algorithms is the fact that there is still a limited amount of studies systematically investigating which personal factors may influence user preferences and how that relationship between subjective abilities and preferences exactly looks like (especially within the subclinical population). Fulfilling this gap was a strong motivation for this thesis that led to the development of specific research questions as reported in the following.

1.2. FORMULATION OF RESEARCH QUESTIONS

There is an extensive body of research targeting listening preferences alone - that often did not tackle the issue of personal abilities but focused on user satisfaction ratings instead (Chung et al., 2009; Chen et al., 2017). Other studies measured individual preferences but were limited to normal-hearing listeners and/or specific age groups (Torcoli et al., 2018; Reddy et al., 2017). Hearing-aid-related studies often employed detailed methods of personalization and investigated the relationship between subjects' hearing abilities and preferences under different listening conditions (Wu et al., 2015; Nielsen et al., 2015) - which is in line with the goals of this thesis. However, the current research project addresses a different group of subjects (normal hearing or those with not yet treated hearing loss) and targets different devices and usability scenarios, so the results or approaches from the field of rehabilitative audiology have only limited use. For this group of audio consumers, there is still a significant lack of data on which this current research could directly build on. Hence, there is a need to develop a systematic framework to approach this research problem - namely to address the topic of playback personalization based on hearing state and listening preferences (establish and test the possible link between these two) among un-aided audio consumers.

1.2.1. ESTABLISHING A LINK BETWEEN HEARING ABILITIES IN SIMPLE AND COMPLEX LISTENING CONDITIONS

In order to successfully track personal abilities and playback expectations under various listening conditions, like an audio personalization algorithm would do, it is necessary (for experimental purposes) to narrow down the scope of possible scenarios to some simple and specific ones. One of the promising possibilities would be to use the outcomes of one (or more) listening tests held in simple, artificial listening conditions (e.g., diotic tests) and relate them , together with a set of personal factors (e.g., age), to the actual hearing abilities of a person measured in complex listening scenarios, closely matching life-like conditions. Building such a link could involve a swift, clinically applicable listening test followed by tests in more complex listening setups. This approach was followed in the second chapter of this thesis that aimed at assessing individual performance both in simple and complex listening conditions to

link various individual factors (age, spatial hearing) to abilities (actual speech reception). The next step towards quantifying the results was to apply them in the modeling field. The goal was to test how successfully individual performance could be predicted by a perceptual model for various listening scenarios and to gain additional knowledge on how the model accuracy could be improved considering individual differences.

Early psychoacoustic research on speech perception showed that the main external factor affecting its recognition (or reception) is noise (Licklider, 1948; Miller and Licklider, 1950; Pollack, 1948). Further experiments showed that the spatial constellation of speech and noise sources determines interaural time, level and phase differences (ITD, ILD and IPD respectively) at the ear canal entrance at both ears and significantly affects speech intelligibility (e.g., Peissig and Kollmeier, 1997; Bronkhorst and Plomp, 1988; Carhart, 1968; Culling, 1995; Levitt, 1967; Kidd et al., 1998). On the other hand, the most profound and common personal factors influencing proper speech understanding are age and hearing state (Gelfand, 1988; Dubno et al., 1984; Hopkins, Moore, 2011, Stuckenberg et al., 2018; Völker et al., 2015). Hence, testing a wide variety of subjects of different age and PTA with respect to their performance in speech perception under simple (e.g., single noise source) and complex listening conditions (multiple, spatially separated masker sources) could lead to a better understanding of how our hearing abilities are used when the difficulty of the task changes and how age and hearing loss affect the performance. This approach led to defining the following research questions:

- How do individual factors such as age and hearing state influence subjective performance in speech-in-noise recognition tests?
- Can speech intelligibility performance under complex listening conditions involving multiple, spatially separated noise sources be predicted from outcomes of a standard diotic speech-in-noise recognition test?
- How accurately can individual performance in complex listening conditions be modeled by existing binaural speech intelligibility models using test signals and hearing threshold information?

The outcome from this part of the thesis was expected to be a better understanding of the impact that individual factors may have on speech recognition performance in adverse listening conditions and whether this performance could be predicted from standard diotic speech-in-

noise recognition tests supplemented by basic individual factors such as age and audiogram. A heterogeneous group of subjects was involved in the study (from normal hearing to moderately hearing impaired) and tested under different maskers and their spatial constellations. To link the empirical findings to auditory theory, the second part of the chapter focuses on the verification of how accurately a binaural speech intelligibility model predicts the experimental findings. To better interpret the data based on the known properties of binaural hearing, the BSIM model was chosen due to its capability of predicting a benefit of masker amplitude modulations ("dip listening") and a combined benefit of masker amplitude modulations and spatial unmasking - which covered the main study conditions. BSIM requires binaural speech and masker recordings as an input. Yet, it allows simulating individual hearing state (and hearing loss) by supplementing hearing thresholds and adding uncorrelated noise to the two-ear signals if hearing loss is present. Additionally, predictions were calculated utilizing several approaches of accounting for individual speech recognition performance e.g., various ways of deriving reference speech intelligibility performance and/or audiogram data. The goal was not only to find which individual factors would be the most influential in improving the model predictions but also to account for limitations in the framework.

1.2.2. Relation between hearing abilities and listening preferences

Apart from investigating a connection between personal factors (e.g., hearing abilities, age) and performance in speech intelligibility tasks, listening preferences needed to be considered since they play an important role in how subjects perceive an audio scene and benefit from signal modifications (Brons et al., 2012; Neher et al., 2016; Brons et al., 2014). There is a growing volume of studies collecting data on subjective hearing abilities and investigating their relation to preference judgments in complex, everyday listening scenarios. Techniques such as the Ecological Momentary Assessment (EMA) (Wu et al., 2015; Kowalk et al., 2017) became a valid resource of data both for hearing aids' tuning as well as audio processing algorithms' development. EMA requires collecting data on current or very recent subjective experiences directly in the environments in which they occur, which involves multiple assessments made by participants (e.g., in a form of surveys) asking for their hearing performance and detailed context description. Numerous studies (Walden et al., 2004; Walden et al., 2005; Cord et al.,

2007; Palmer et al., 2006; Humes, 2002) connecting personal abilities and preferences show that subjects (in general) prefer individualized processing and that such treatment leads to higher comfort for the users. Another interesting outcome is that listeners can indeed well estimate their speech understanding and describe their listening context in both laboratory settings and semi-controlled real-world environments. For example, aggregated data (EMA) suggested a high degree of validity of that methodology showing high correlations between surveys' outcomes and established audiological knowledge on listening experience and its relationship with listening context in real-life scenarios. Nevertheless, it is important to notice that a great majority of the studies researching preference judgment concerning hearing abilities targeted hearing aid users, while the normal-hearing and subclinical part of the population often remained unaddressed. Therefore, the goal of the next part of this thesis was to develop an approach and collect experimental data to provide a personal preferences' profiles for such subclinical subjects. That step of profiling audio consumers involved adding individual listening preferences to the already existing set of abilities (from chapter 2). Hence, chapter 3 presents an investigation on such connection between individual preferences for speech processing in adverse listening conditions (the same conditions as in the chapter 2, with the same group of participants) and speech reception data together with individual factors (age, hearing state - that were collected within the first experiment, in chapter 2).

In the experiments described in this thesis, methods of adjustments were used to collect data due to their high time efficiency and the nature of the experiments that involved simple linear or close to linear signal modifications, hence no audio quality rating methods were used.

The research questions underlying the second experiment were:

- Are hearing abilities or other personal factors (e.g. age) related to listening preferences?
- Are listening preferences dependent on the listening scenario?
- Are listening preferences stable over time and/or across listening scenarios?
- Is the relation between hearing abilities and listening preferences stable across conditions and/or over time?
- What are personal factors or abilities that play a decisive role in preference judgement?

The outcome of this part of the thesis was expected to be a better understanding of how personal features underlie listening preferences and how stable such preferences are both over time and throughout different listening scenarios. Subjects adjusted speech stimuli to their

preferences using four signal modification schemes: i) linear gain, gain at the cost of ii) clipping distortions or iii) compression distortions, and iv) frequency-shaping. The motivation for the study design were data obtained by Völker et al., (2018), suggesting that subjects could be consistently categorized into "noise haters" versus "distortion haters". In the current investigation, one of the aims was to test whether this trait remains stable through different maskers, spatial conditions, and types of distortions. Since the same group of subjects was tested in two experiments – a comparison of listening preference with individual factors (data from chapter 2) was performed to elicit which preferences in complex listening conditions correlate best with personal factors (age, PTA) and individual speech intelligibility results.

1.2.3. INTRODUCING THE IDEA OF PLAYBACK PERSONALIZATION IN SPORT

The first experiment focused on finding a set of personal factors (e.g., age, PTA) that could potentially influence speech recognition performance as well as help to model such performance under complex listening conditions. In the second, the investigation on how listening preferences for supra-threshold processing could be linked to those personal factors and performance was conducted. While the third experiment used physical exercise and fatigue under sport condition as an experimental variable, hence replacing hearing loss and age as an experimental parameter. The main motivation behind it was that thousands of people exercise to music every day, yet very little to nothing is known on how their growing physical fatigue affects both hearing abilities and listening preferences. There are just a few studies of normal-hearing subjects (Hutchinson et al., 1991; Lindgren, F., Axelsson, 1988; Vittitow et al., 1994) showing that exercising in silence can lower hearing thresholds, hence improve hearing abilities, yet when such exercise is performed in noise or in music the same studies' results are contradictory. One study found that listening to music while exposed to physical fatigue can increase the hearing threshold even further than being exposed to loud music alone (Vittitow et al., 1994), while another study suggested the opposite (Hutchinson et al., 1991). This research problem may be crucial to address to correctly assess risks associated with doing sport while listening to (loud) music as well as to develop hearing protection against increased hearing thresholds and other hearing inconveniences possibly associated. It is also important to realize that a phenomenon of possible listening preferences' shift due to the growing fatigue has - to

the author's best knowledge - never been investigated before. This motivated the study described in the fourth chapter of the thesis. The main research questions were:

- Does hearing perception (as expressed by preferred volume and frequency shaping) change while doing sports?
- Are hearing thresholds stable under growing fatigue?

The experiment described in chapter 4 investigates the impact of sport exercises (cycling on an ergometer) on listening preferences (regarding gain and frequency shaping). The study consists of both experimental data (10 normal-hearing participants) as well as online survey data (N = 138). One of the following research questions was to determine whether there is any consistent shift in listening preference when moving from resting condition through growing fatigue. Another: how comparable listening preferences are between the states of individual maximum fatigue and the onset of exercises? Finally, could the potentially dangerous impact of sport exercises (when accompanied with loud music) on hearing thresholds reported in previous studies also be confirmed in the current one? Are there any long-term hearing inconveniences present even after the end of the exercises? To address these questions, a series of experiments with 10 normal-hearing subjects were conducted.

1.3. CONCLUDING REMARKS

The main goal of the thesis was to find a set of personal factors (or "personal traits") responsible for preference judgment of normal hearing and those with not yet treated hearing loss. Listening preferences were investigated in various conditions, but specific stress was put on speech intelligibility in adverse listening conditions as being essential for everyday life. Experiments reported in chapters 2 to 4 were designed to spread a wide range of realistic scenarios (from speech perception in noise to sport with music), collect data from a set of diversified subjects, employ novel research methods and a state-of-art modelling approach to provide useful insights into the relation between personal factors (e.g., age, PTA), auditory perception (e.g., SRT) and listening preferences. Obtained results are applicable both for further clinical research as well as industrial applications.

2. PREDICTION OF INDIVIDUAL SPEECH RECOGNITION PERFORMANCE IN COMPLEX LISTENING CONDITIONS[1]

This study examined the relation between age, average hearing loss and performance in clinical standard diotic speech-in-noise recognition tests with individual speech recognition performance in complex listening conditions. The SRT data were obtained both with the Digit Triplet test (DTT, Zokoll et al., 2012) and the Goettingen sentence test (Goesa, Kollmeier and Wesselkamp, 1997) for 7 normal-hearing and 23 hearing-impaired subjects. The listening conditions included two masker types (multi-talker and two-talker), and two spatial conditions – with target speech being presented always from the front, and maskers being either co-located or spatially separated (60 degrees to the left and right from the target source in same horizontal plane). The BSIM was used to predict the experimental SRT data for the different conditions. The results indicate that prediction accuracy could be increased if the outcome of the clinical sentence-in-noise test is used in addition to the individual audiogram (or SII derived from it), thus demonstrating the applicability of this approach for clinical purposes. However, larger discrepancies remained between predictions and data in conditions characterized by informational masking for all hearing-loss or age groups, indicating that the current model and factors do not yet characterize this phenomenon in a sufficiently precise way.

2.1. INTRODUCTION

Listening to speech in noisy backgrounds is a challenge for everyone but affects us very differently. Although different contributing factors, e.g., hearing loss and masker type, have been thoroughly investigated, the reasons for this inter-individual variability are yet not fully understood and cannot be well predicted by current speech intelligibility models. One goal of this study was to investigate how individual factors such as age, PTA and speech intelligibility in standardized, simple listening conditions relate to speech recognition performance in more

[1] A slightly modified version of this chapter has been published in the Journal of the Acoustic Society of America (Kubiak et al., 2020).

complex listening scenarios involving spatial hearing and IM. A second goal was to test to what extent the individual speech recognition performance in these scenarios can be predicted by models employing individual hearing thresholds, which have been shown to well predict group data of binaural unmasking in both stationary and modulated maskers, especially for normal-hearing listeners (Beutelmann and Brand, 2006; Beutelmann et al., 2010; Rennies et al., 2011). Finally, a third goal was to provide an accessible set of individual speech perception data in complex auditory scenes, along with a set of individual factors (age, audiograms, speech performance in simpler conditions), which can be used by speech perception scientists to develop and validate models and their applicability to individual performance prediction.

Speech perception in adverse conditions has been investigated extensively, including early works applying the AI as a function of intensities of speech and unwanted sounds received by the ear, different noise types (Hawkins Jr., 1950), noise levels as well as speech processing techniques and their impact on intelligibility (Licklider, 1948; Miller and Licklider, 1950; Pollack, 1948). In many of these early studies, it became clear that SRT, i.e., SNRs at which listeners can understand 50% of a speech signal, may vary depending on many factors such as, for instance, background noise, a spatial constellation of the speech signal and masker, and SNR. Also, individual features of the listener play an important role especially concerning the ability to efficiently segregate speech signals in complex auditory scenes. This ability, often referred to as the cocktail party effect (Cherry E., 1953) has been researched ever since. Multiple factors can influence speech perception in complex listening conditions ranging from the signals used to the listener's personal features including hearing and cognitive abilities per se, or the current listening environment.

Two of the most distinctive personal factors that may affect speech perception are age and hearing loss (Gelfand, 1988). These two factors often co-vary together (Dubno et al., 1984; Hopkins, Moore, 2011) as hearing abilities decline with age, and there are numerous studies concerning age and hearing loss in the light of different masker type characteristics as well as SNR, (e.g. Duquesnoy, 1983; Brons et al., 2012; Brons et al., 2014; Stuckenberg et al., 2018; Völker et al., 2015). In general, people with cochlear hearing loss require higher SNRs than normal-hearing people to obtain similar speech recognition performance. This has been found in conditions with so-called energetic masking, when noise and target signal components simultaneously excite the same auditory filters. Higher SRTs for hearing-impaired listeners have also been found for conditions involving IM, which occurs due to a high degree of

similarity between useful signal and the masker along any other stimulus dimensions (except for those related to energetic masking, see Kidd and Colburn, 2017, for a recent review). However, not all studies agree as to the role of hearing loss in solving the cocktail party problem. For example, the study by Micheyl et al. (2000) failed to find significant differences in IM between normal-hearing and hearing-impaired subjects with sensorineural hearing loss of diverse etiologies.

Other sources of acoustic factors influencing speech recognition in noise are those related to the acoustic scene, such as ILD, ITD and IPD (e.g., Peissig and Kollmeier, 1997, Bronkhorst and Plomp, 1988; Carhart, 1968; Culling, 1995, Levitt, 1967; Kidd et al., 1998). For example, Bronkhorst and Plomp (1992) found that hearing-impaired listeners with symmetrical sensorineural hearing loss (PTA between 16 and 56 dB HL) needed between about 4 to 10 dB better SNR to achieve speech intelligibility scores equal to those of normal-hearing subjects in a setup of multiple speech-like maskers. With different masker modulations present in addition to spatial setups, Beutelmann et al. (2010) found up to 15.5 dB SRT benefit due to a masker change from stationary speech-shaped noise to single-talker modulated noise for normal-hearing subjects in binaural listening conditions. This benefit decreased or even vanished for listeners with hearing impairment. The same study also found that spatial release from masking (SRM), being up to 9 dB for normal-hearing listeners (anechoic conditions, speech-like maskers) was considerably decreased in listeners with hearing impairment. Research data on spatial hearing, produced for years in numerous setups (e.g., Peissig and Kollmeier, 1997; Bronkhorst and Plomp, 1992; Helfer and Freyman, 2008, 2014; also reviews of Drullman and Bronkhorst, 2000, and Bronkhorst, 2015), generally agree that spatial separation of target speech and maskers improves speech intelligibility, but that strongly depends on the individual subject and factors related to the signals employed such as the degree of IM (e.g., Best et al., 2012). The inter-individual differences in the ability to use spatial unmasking have been linked to hearing impairment (e.g., Goverts and Houtgast, 2010) or acoustic scene width (e.g., Peissig and Kollmeier, 1997), while there is controversy as to the role of age. Füllgrabe et al. (2015) did not find a systematic effect of age if the hearing loss does not co-vary with his subgroup of aged normal listeners, while other authors showed a clear age-dependency (e.g., van Esch, Dreschler, 2015).

Many of the previous studies focused on rather small variations of individual factors (e.g., by using rather homogenous groups of subjects with similar age and/or hearing loss) or

focused on particular listening conditions without varying the complexity of acoustic scenarios. Similarly, studies testing current binaural speech intelligibility prediction models have mostly focused on homogeneous listener groups with similar hearing loss, (e.g., Beutelmann and Brand, 2006; Beutelmann et al., 2010; Brand et al., 2017) or have been limited to normal-hearing listeners (e.g., Lavandier and Culling, 2010; Lavandier et al., 2012; Leclère et al., 2015; Wan et al., 2014; Chabot-Leclerc et al., 2016; Rennies et al., 2011, 2014). Although the models have been shown to achieve good prediction accuracy in relatively simple listening conditions (when accounting for individual audibility for hearing-impaired listeners), it is unclear to what degree they can predict individual speech recognition in more complex scenarios. The partly contradictory results and the limited consideration of individual factors in addition to average hearing loss or age indicate that it is currently not well understood which individual properties determine speech perception in complex listening scenarios and how this could be modeled. One possible indicator of individual acuity in communication situations of different complexity may be the speech recognition performance in highly standardized, comparatively simple speech-in-noise recognition tests (DTT), which have been used as a screening tool for hearing impairment (Lyzenga and Smits, 2011). Another kind of tests, often employed in conjunction with stationary speech-shaped noise and monaural or diotic listening, are sentence tests like Goesa or the Leuven intelligibility sentence test (Jansen et al., 2014). Even though the outcome from these tests in listeners with hearing impairment is highly correlated with age and PTA (van Esch and Dreschler, 2015; Luts et al., 2010), there is still considerable unexplained variability of the test outcomes that may be due to sensory processing, cognitive processes or individual, nonauditory factors (see Kollmeier and Kiessling (2018) for a review).

It is currently not well understood, however, how these individual factors could be integrated in advanced speech intelligibility models to improve the prediction accuracy with respect to interindividual performance differences. One possible way for improved model individualization could be to derive an individual reference when mapping the model output (often an index) to the perceptual metric (e.g., the SRT). Such an approach was proposed by George et al. (2010) for noisy and reverberant speech. They employed the STI, which had been developed for normal-hearing listeners (and hence does not account for reduced audibility) and determined an individual STI value in a reference condition from which they could predict SRTs in other listening conditions. A similar approach was employed by Brand et al., (2017) in the framework of the BSIM model. In contrast to George et al. (2010), they used individual

audiograms in addition to deriving individual references and showed that BSIM was applicable for predicting individual SRTs of aided speech (individual NAL-R fitting) in speech-on-speech masking conditions and spatial noise reduction schemes, except for conditions with a high degree of IM, which is not accounted for by the model. Brand et al., (2017) focused on aided and noise-reduced speech and did not investigate differences in prediction accuracy for different ways of deriving the individual reference values (e.g., including the individual audiograms or not). Similarly, neither of these studies investigated the relation between individual reference values and other individual factors such as SRTs in more standardized speech intelligibility tests, so these remain open questions. Another approach for predicting SRT of individual listeners without requiring speech-material-specific reference values was recently provided by Schädler et al. (2018) and Kollmeier et al. (2016) who achieved, by automatic speech recognition, a high prediction accuracy using the individual precisely determined audiogram, and one additional supra-threshold measure estimating the distortion. However, their approach did not include a specific binaural interaction model (like BSIM) and it was limited to the Matrix sentence test (Kollmeier et al., 2015), thus not considering the clinically more common speech tests (like DTT and Goesa) and complex auditory scenes.

The general goal of the current study therefore is to investigate the relation between the individual performance in complex listening scenarios with the factors age, hearing loss, and additional speech recognition-related factors that are captured by speech recognition performance in two standard diotic speech tests (DTT and Goesa). The performance in complex scenes is characterized here by the individual ability to efficiently segregate speech signals in complex auditory scenes involving energetic masking, IM as well as potential benefits due to spatial unmasking or masker amplitude modulations. Both single- and multi-talker maskers were considered (varying the similarity between masker and target) and SRM was measured in different conditions. Subjects were widely spread with respect to their individual factors to facilitate the analyses if these factors were correlated to individual SRTs in the tested listening conditions. Predictions of BSIM were calculated with different ways of accounting for individual recognition performance in an attempt to find the most suitable way to account for individual limitations in the framework.

2.2. METHODS

2.2.1. SUBJECTS

Thirty subjects (18 female, 12 male) aged 23-85 years (mean 54.5 years, standard deviation of 19.7 years) participated in the study. Subjects were selected from the database of the Hörzentrum Oldenburg, Germany, which comprises several hundred subjects, with information on, among others, age, audiogram, and speech recognition performance. For this study subjects were selected to span a wide range along four dimensions: their age, their PTA, and their speech recognition performance in formal diagnostic intelligibility tests (DTT and Goesa, both performed diotically via headphones).

When assessed according to grades of hearing impairment as proposed by Martini (1996), seven subjects were normal-hearing with PTAs ≤20dB HL at their better ear (ranging from -2.5 to 5 dB HL, mean 0.5 dB). Ten of the listeners were slightly hearing-impaired, with PTAs ranging from 21.25 to 30 dB HL (mean 24.0 dB HL). Thirteen subjects were moderately hearing-impaired with PTAs in the range from 42.5 to 70 dB HL (mean 53.5 dB HL). All subjects were native German speakers. None of the subjects had asymmetric hearing loss, determined as more than 20 dB inter-aural threshold difference at the six octave frequencies from 250 to 8000 Hz (Pittman and Stelmachowicz, 2003), although for the subject with the most severe high-frequency hearing loss no audiogram data were available above 4 kHz because the output limit of 105 dB HL was reached. Nineteen subjects had a sloping shape of the audiogram and eight subjects had a flat audiogram. Additionally, two subjects had a V-shaped audiogram, and one subject a tent-shaped audiogram. Four subjects had self-reported Tinnitus and had performed audiogram measurements using frequency-modulated tones instead of sinusoids. For the DTT, the reference SRT range for native, normal-hearing listeners is -9.3±0.2 dB SNR, and all seven normal-hearing listeners were better than or within this range (SRTs between -10.0 and -9.2 dB SNR, mean -9.5 dB SNR). For Goesa, the reference range is -6.1±0.2 dB SNR, and not all seven subjects were better than or equal to this reference (SRTs between -7.6 to -4.6 dB SNR, mean SRTs of -5.8 dB SNR). The slightly hearing-impaired subjects had mean SRTs of -7.1 and -2.8 dB for DTT and Goesa, respectively; and the moderately hearing-impaired subjects had mean SRTs of +0.2 dB and +1.4 dB SNR,

respectively. Linear correlations between the individual factors used in this study were checked for the whole probe of subjects. Among the best correlated were PTA with Goesa test results (R^2=0.79), PTA with DTT (R^2=0.75), and both speech intelligibility tests with each other (R^2=0.78). A much lower correlation was observed between age and the other individual factors, with R^2 not exceeding 0.40 (age-DTT: R^2=0.31, age-Goesa: R^2=0.40, age-PTA: R^2=0.35). Fifteen subjects were hearing-aid users but performed the measurements unaided. Subjects received an hourly compensation for their participation in the study.

2.2.2. APPARATUS AND PROCEDURE

The study was conducted using a Matlab software environment installed on a personal computer, an RME Fireface UC USB High - Speed Audio Interface soundcard, and a Tucker-Davis Technologies HB7 headphone driver. The stimuli were presented to the subjects in a sound-attenuated booth via Sennheiser HD 650 headphones that were calibrated to dB sound pressure level (SPL) using a Bruel&Kjær (B&K) 4153 artificial ear, a B&K 4134 microphone, a B&K 2669 preamplifier, and a B&K 2610 measuring amplifier. The impact of the headphones was free-field equalized using a finite impulse response filter with 118 coefficients.

For each measurement condition (see below), SRTs were measured using the Oldenburg Sentence Test (Wagener et al., 1999) in an adaptive procedure (Brand, Kollmeier, 2002) aiming to determine the SNR at which 50% of the presented words were understood correctly. After each sentence the subjects repeated what they had understood to the experimenter, no visual representation of the response alternatives (open response format) and no feedback was provided. The level of the target sentences varied during the measurement (starting from an SNR of 10 dB) while the masker level was fixed at 71 dB SPL. For each condition SRTs were measured using a list of 20 sentences. A different, randomly selected list was used for each condition. Before the measurement session subjects performed a training, session consisting of two lists presented on the background of the standard test procedure noise (stationary speech-shaped noise, SSN) and aiming at determining SRTs for 80% of correctly understood words. The entire experimental procedure lasted approximately 30-40 minutes per subject.

2.2.3. STIMULI AND MEASUREMENT CONDITIONS

Target speech stimuli consisted of sentences of the Oldenburg sentence test that are built with a fixed five-word order (name-verb-numeral-adjective-object), are grammatically correct, but semantically unpredictable. The sentences were spoken by a male talker. The target speech was convolved with head-related room impulse responses (HRIRs) corresponding to frontal incidence (0 degrees) in the horizontal plane. HRIRs were taken from the database of Kayser et al. (2009), recorded with a B&K head and torso simulator (HATS, 1904128C) in an anechoic environment, with a distance to the speaker of 80 cm. The target speech material was embedded in the maskers, so that the length of the masker signals for each sentence presentation was 5 seconds, while the length of the sentences varied from 1.9 to 2.9 seconds. Each sentence was placed symmetrically in the middle of the masker segment, which resulted in pre- and post-sentence masker presentation lasting about 1 to 1.5 s.

Two different maskers were used:

- "two-talker" (TT) – two streams of competing sentences from the Oldenburg sentence test spoken by another male talker (recordings from Hochmuth et al., 2015). The two masking streams consisted of different sentences with randomly selected starting positions so that the sentence rhythm randomly differed between the two masker streams and between the masker streams and the target sentences;
- "multi-talker" (MT) - multi-talker babble consisting of 10 male and female talkers, the material was cut and remixed so that the meaning as well as language of each talker were not understandable.

Target speech and maskers were presented in two different spatial conditions:

- "co-located" (indicated as "cl" in figures and tables): target speech and maskers were convolved with the HRIR of 0 degrees; hence all signals were presented from the front;
- "spatial" (indicated as "sp" in figures and tables): target speech was presented from the front and the maskers were symmetrically located away from the target source on the horizontal plane 60° in both directions, so that the left masker was convolved with HRIRs for -60°, and the right masker with HRIRs for +60°. Different segments of the multi-talker babble were used for both sides.

All maskers were equalized in 1/3-octave bands after the convolutions to match the long-term spectrum of the frontal target speech. Hence, there was no long-term spectral difference (due to the equalization), and no long-term better ear listening (due to the masker symmetry) in any of the conditions.

2.2.4. SPEECH INTELLIGIBILITY PREDICTION

To assess to what degree the interindividual variability could be predicted by a current speech intelligibility model mimicking the effective speech processing stages involved in binaural hearing, the BSIM model was used. This model was chosen because it had been explicitly designed to predict the influence of two of the main factors involved in the acoustic conditions of the present study, i.e., the benefit of masker amplitude modulations, also referred to as “dip listening”, and the combined benefit of masker amplitude modulations and spatial unmasking. BSIM receives the binaural speech and interferer recordings as input. Individual hearing thresholds can be provided as input to simulate hearing loss, which is implemented by adding uncorrelated noise to the two ear signals. The input signals are processed in 30 gammatone frequency channels with center frequencies between about 140 Hz and 8.3 kHz. In each filter, an independent Equalization Cancellation (EC, Durlach, 1963, 1972) process is applied to model binaural processing. In the EC-stage, the two ear signals are amplified and delayed relative to each other and gain and delay parameters are chosen such that an optimum SNR is available after subtracting the ear signals. This means that if speech and noise have different interaural level and/or phase differences, the SNR can be improved relative to the monaural or diotic SNRs. The output of the binaural stage consists of time- and frequency-dependent SNRs, which are then analyzed by the SII in a short-term version (similar to the concept of the ESII). The SII of each frame is then converted to an SRT, and the final SRT prediction is obtained by averaging across frames. The mapping from the output SII to SRTs is achieved by first computing the index for a reference condition at the SNR corresponding to the experimentally measured SRT. For all other conditions, this reference value is then kept constant, and SRT predictions are obtained by varying the SNR until the reference value is reached.

Five different ways of individualizing BSIM were tested and compared in the present study:

A. Normal reference SII, individual audiograms: in this model version, the reference SII was determined by the mean SRT of the seven normal-hearing subjects for frontal speech and co-located, MT maskers (reference condition). This reference SII was then fixed for all subjects, and individual audiograms were used to predict individual SRTs in all conditions. This is conceptually equivalent to several previous studies using BSIM (e.g., Beutelmann and Brand, 2006; Beutelmann et al., 2010; Rennies et al., 2011).
B. Individual reference SII, normal audiogram: in this version, normal audibility (i.e., a flat 0 dB HL audiogram) was assumed in the model and individualization was achieved by deriving individual SII reference values from the individual SRTs in the same reference condition as used in A. This corresponds to the concept proposed by George et al. (2010), who used the "normal-hearing" STI in conjunction with individual reference values.
C. Individual reference SII, individual audiograms: in this version both individual reference SII values and individual audiograms were used, again using the same reference condition as in A and B. This version was also employed by Brand et al., (2017) for their set of stimuli and conditions.
D. Individual reference SII derived from Goesa, normal audiogram: this version was the same as B, but the individual reference SII values were derived from the available Goesa-SRTs rather than the co-located MT masker condition of the present study.
E. Individual reference SII derived from Goesa, individual audiograms: this version was the same as D, but in addition to the individual reference SII value derived from the Goesa, individual audiograms were used.

For each of these versions, model predictions were calculated using stationary SSN (with the same long-term spectrum and direction as the target sentence in the experiment) to mimic the target speech, and the same masking signals as in the experiment. This corresponds to the procedure of the ESII as proposed by Rhebergen and Versfeld, (2005). For each prediction in the present study, ten repetitions with randomly created maskers and randomly selected 3-s portion of the SSN were calculated. The mean across these ten repetitions was used as SRT prediction for each subject and condition.

2.3. RESULTS

2.3.1. SPEECH RECEPTION THRESHOLDS

Individual, as well as mean, SRTs for the whole subject group are shown in Fig. 2.1, sorted according to individual performance for the spatially separated TT masker. The top and bottom panel show results for TT and MT maskers, respectively. Regarding the mean performance of the whole group (right panels) it could be observed that, in general, subjects performed better (i.e., scored lower SRTs) in spatially separated than in co-located conditions for each masker type. For the TT masker, all 30 subjects scored better when target speech and masker source were spatially separated (mean SRT of -9.5 dB compared to -0.5 dB), but the standard deviation associated with this condition (6.5 dB) was almost four times greater than for the co-located listening condition (1.6 dB). Similar results were observed for the MT masker, where the mean subjective results for spatially separated maskers were about 3.0 dB lower than for co-located maskers (-6.9 dB vs. -4.0 dB SNR). The inter-individual standard deviation (3.2 dB) for this case was about twice as large as for the co-located condition (1.8 dB). In general, SRTs in the co-located conditions were lower for the MT masker than for the TT masker (average -4.0 vs. -0.5 dB SNR), while SRTs for spatially separated maskers were lower for TT than for MT (-9.5 vs. -6.9 dB SNR).

In order to assess the group effects of masker type (MT vs. TT) and HRIR conditions (co-located vs. spatially separated) on observed SRTs, a two-way repeated measure analysis of variance (ANOVA) was performed with a significance level of $\alpha = 0.05$ and Greenhouse-Geisser corrections of the degrees of freedom. The ANOVA showed no significant main effect of masker type [$F(1,29)=1.863$, $p=0.183$], but a significant effect of HRIR condition [$F(1,29)=91.93$, $p<0.01$]. The interaction between both factors was also significant [$F(1,29)=72.203$, $p<0.01$], indicating that the observed opposing trend in how the masking properties of TT and MT maskers were affected by spatial separation was significant, and led to a compensation of the overall effect of masker type in the present study.

Regarding individual results (left panels in Fig. 2.1), it could be observed that the performance in spatially separated conditions for the TT masker (according to which subjects are ordered in Fig. 2.1) could well predict the rank order of SRTs for the MT masker under the

same spatial condition, i.e., subjects with low SRTs for one masker generally had low SRTs also for the other masker (compare top and bottom panels). SRTs for the TT masker in spatially separated conditions had the widest spread within the study (ranging from -20.3 to 1.8 dB SNR), while the MT masker type in this spatial condition produced only about a half of that variability ranging from -12.5 to 0.1 dB SNR. In co-located conditions the spread for the TT (-4.5 to 2.4 dB SNR) and for the MT (-7.6 to 0.1 dB SNR) maskers was more similar. Squared linear correlations of SRTs between both maskers in the same spatial conditions (see Table 2.1) were relatively high and significant (R^2=0.91 between spatially separated TT and MT maskers, and R^2=0.74 between co-located TT and MT maskers). Also, the correlations between TT spatial and MT co-located (R^2=0.73), and between TT co-located and MT spatial (R^2=0.70) were significant, as were correlations between SRTs for the same masker type in the two spatial conditions (TT: R^2=0.62, MT: R^2=0.84).

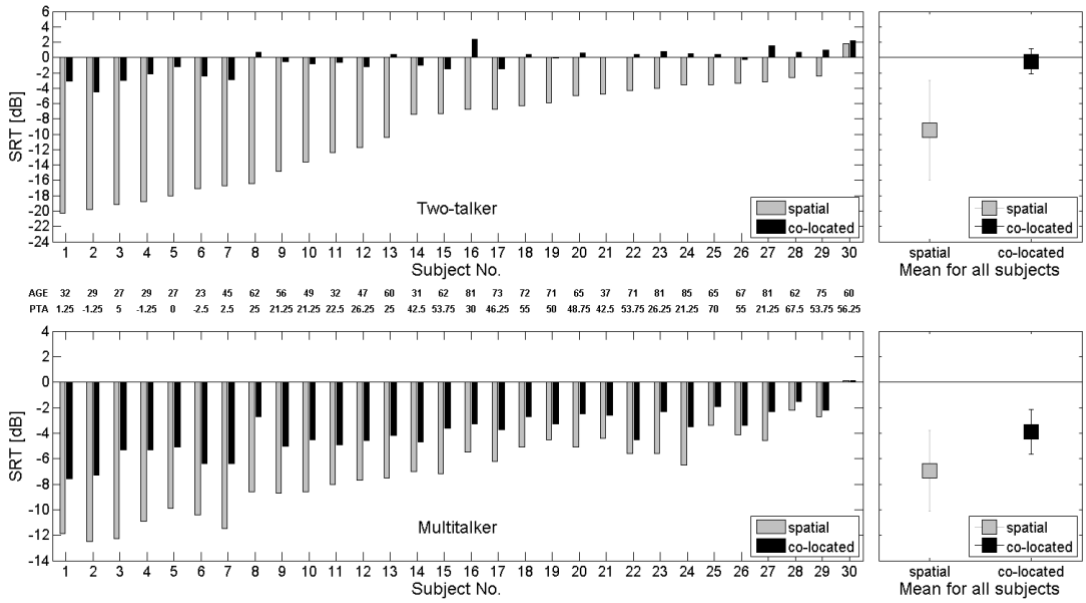

FIG. 2.1: INDIVIDUAL (LEFT) AND MEAN SRTS (RIGHT) FOR BOTH MASKER TYPES AND SPATIAL CONDITIONS. THE SUBJECTS' AGE AND PTA ARE INDICATED BETWEEN THE TWO PANELS

2.3.2. SPATIAL RELEASE FROM MASKING

The SRM is further illustrated in the top panels of Fig. 2.2. SRM was calculated as the difference in SRTs between co-located and spatially separated conditions for the same masker type (subjects sorted, as previously, according to their SRT performance in the spatially separated, TT masker condition). As shown, SRM was always positive, i.e., for all subjects, spatial separation was beneficial, but the amount of SRM varied considerably from 0.4 to 17.2 dB for TT, and from 0.0 to 7.0 dB for MT maskers. The inter-individual standard deviation of SRM was more than three times greater for the TT masker (5.3 dB) than for MT (1.7 dB). Regarding individual SRM results it could be observed that SRM was greater for subjects who scored lower (better) SRTs. In other words, subjects with relatively better speech recognition abilities within this study tended to experience a higher benefit from spatial separation of target and masker sources than subjects with relatively worse SRTs. This effect was especially visible in SRM results for the TT masker. The correlation between individual SRM for the two noise types was highly significant (R^2=0.78, rank correlation 0.88).

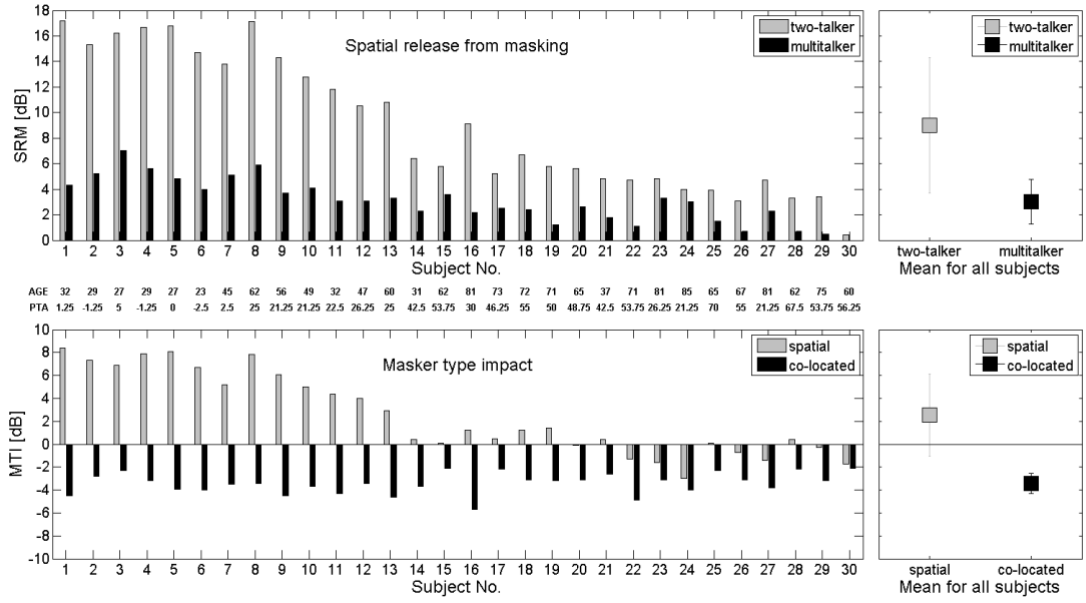

FIG. 2.2: INDIVIDUAL (LEFT) AND MEAN (RIGHT) RESULTS OF SPATIAL RELEASE FROM MASKING (SRM, TOP) AND MASKER TYPE IMPACT (MTI, BOTTOM) IN BOTH MASKER TYPES AND SPATIAL CONDITIONS.

2.3.3. IMPACT OF MASKER TYPE

To analyze the masker type impact (MTI) on SRTs, the differences between SRTs in MT and TT conditions for the same spatial conditions were calculated. MTI thus expresses the potential benefit of changing the masker from MT to TT in a given spatial constellation, so exchanging a more continuous masker by one that may enable dip listening due to its stronger temporal modulations (but that presumably also conveys more IM). Individual, as well as mean, results for the whole group are shown in the bottom panels of Fig. 2.2. The mean MTI for the co-located maskers (black square) was -3.4 dB, indicating that the MT masker type was the one for which subjects tended to perform better than for TT maskers. In contrast, for spatially separated maskers (gray square) the mean MTI was positive (2.5 dB), indicating that subjects scored lower SRTs for TT maskers than for MT maskers.

As observed for SRM, MTI differed across subjects with standard deviations of 0.9 dB for co-located and 3.6 dB for spatially separated maskers. For spatially separated maskers, subjects who scored better SRTs generally experienced a higher benefit from masker type change from MT to TT. However, a considerable number of the subjects (generally those with the poorest SRTs), did not benefit from masker type change, i.e., showed very small or even negative MTIs. For co-located maskers, all subjects showed negative MTIs, and the amount of MTI was similar across subjects (range between -5.7 and -2.1 dB, standard deviation 0.9 dB). Overall, the individual MTI data for the two spatial conditions were not significantly correlated (R^2=0.04, rank correlation -0.26).

TAB. 2.1: SQUARED LINEAR CORRELATIONS (R^2) AND CORRESPONDING RANK CORRELATIONS (IN PARENTHESES) BETWEEN INDIVIDUAL SRTS. BOLD VALUES INDICATE SIGNIFICANT CORRELATIONS ($p<0.05$).

		SRT		
		TT sp	MT cl	MT sp
SRT	TT cl	**0.62 (0.79)**	**0.74 (0.87)**	**0.70 (0.77)**
	TT sp		**0.73 (0.88)**	**0.91 (0.95)**
	MT cl			**0.84 (0.91)**
	MT sp			

2.3.4. RELATIONS BETWEEN SRTS AND INDIVIDUAL FACTORS

To investigate possible relations between SRTs and individual factors, linear correlation coefficients, coefficients of determination (R^2) - as a measure of how much of the observed variance in the data could be explained by the individual factors - as well as rank correlations were calculated and are reported in the following. The left part of Table 2.2. summarizes the correlation measures for relating SRTs measured in the different conditions with age, PTA, and SRTs measured by DTT and Goesa. For all conditions and factors, the observed linear correlations with SRTs were significant, i.e., SRTs generally decreased with decreasing age, decreasing PTA, and decreasing SRTs as measured in the standard intelligibility tests. However, the predictive power varied substantially with R^2 values ranging from 0.28 to 0.79. In general, R^2 values were always larger for spatially separated maskers than for co-located maskers. Except for the factor age, all factors showed the lowest R^2 values for the co-located TT masker, i.e., the condition with the smallest inter-individual SRT variation. Largest R^2 values (exceeding 0.70) were observed for SRTs measured with spatially separated maskers and PTA, as well as Goesa. The same trends could be observed based on rank correlations. The corresponding analyses for the SRT-difference measures (SRM and MTI) are also shown in Table 2.2. For SRM, all correlations were significant, and PTA and Goesa were the best predictors for explaining individual SRM variability for both masker types. Age showed considerably lower predictive performance. For MTI in the spatially separated masker conditions, all factors except DTT showed similar predictive performance ($R^2 \approx 0.55$) and were all significant. MTI for the co-located condition only correlated weakly with PTA, DTT, and Goesa, but not with age. Again, rank correlations produced very similar trends.

TAB. 2.2: SQUARED LINEAR CORRELATIONS (R^2) AND SPEARMAN'S RANK CORRELATION (IN PARENTHESES) BETWEEN INDIVIDUAL FACTORS AND SRT, SRM, AND MTI. VALUES FOR SIGNIFICANT CORRELATIONS ($p<0.05$) ARE BOLDED.

	SRT				SRM		MTI	
	TT cl	TT sp	MT cl	MT sp	TT	MT	cl	sp
Age	**0.55 (0.73)**	**0.56 (0.75)**	**0.45 (0.70)**	**0.46 (0.67)**	**0.47 (0.66)**	**0.31 (0.57)**	0.00 (0.04)	**0.57 (-0.72)**
PTA	**0.38 (0.56)**	**0.72 (0.79)**	**0.59 (0.77)**	**0.76 (0.87)**	**0.72 (0.79)**	**0.68 (0.83)**	**0.15 (0.46)**	**0.57 (-0.67)**
DTT	**0.28 (0.62)**	**0.58 (0.87)**	**0.49 (0.81)**	**0.66 (0.90)**	**0.59 (0.85)**	**0.61 (0.84)**	**0.16 (0.43)**	**0.44 (-0.76)**
Goesa	**0.43 (0.64)**	**0.71 (0.84)**	**0.69 (0.85)**	**0.79 (0.90)**	**0.70 (0.82)**	**0.63 (0.79)**	**0.19 (0.47)**	**0.55 (-0.73)**

2.3.5. MODEL PREDICTIONS

Predictions of BSIM were compared to experimental data as scatter plots in Fig. 2.3. Each row shows predictions of one of the model's individualization versions (A to E - as indicated in the panels). The left column shows experimentally measured SRTs plotted against predicted SRTs. In each panel, different symbols represent the four measurement conditions, and for each condition, each subject is shown as one symbol. The solid gray line illustrates the main diagonal, i.e., a perfect agreement between data and predictions, while the dashed and dotted lines indicate a horizontal or vertical shift of ±5 dB and ±10 dB from the diagonal, respectively. From the predicted SRTs, the individual SRM and MTI were computed in the same way as for the experimental data (see middle and right column, respectively). Black solid lines represent first-order polynomial fits to all data points in each panel. Various measures for the models' prediction accuracy were computed and are summarized in Table 2.3. These included the measures assessing if the model predicted the trend of the data, i.e., the linear correlation (R), Spearman's rank correlation (ϱ), and the slope of the linear fit (the better the predictions the closer these values should approach unity), as well as measures indicating the absolute deviation of predicted SRTs from experimental data, i.e., the root-mean-square error (RMSE), the mean absolute error (MAE), and the prediction bias (the smaller the better). The bias was calculated as the y-intercept of a linear fit with unity slope and indicates a general leftward or rightward shift of the predictions compared to the experimental data in the scatter plots. Table 2.3 includes the analyses for all data points in each panel of Fig. 2.3 (*overall*) as well as the error measures for the different masker types (SRT and SRM data) and spatial constellations (SRT and MTI data) separately.

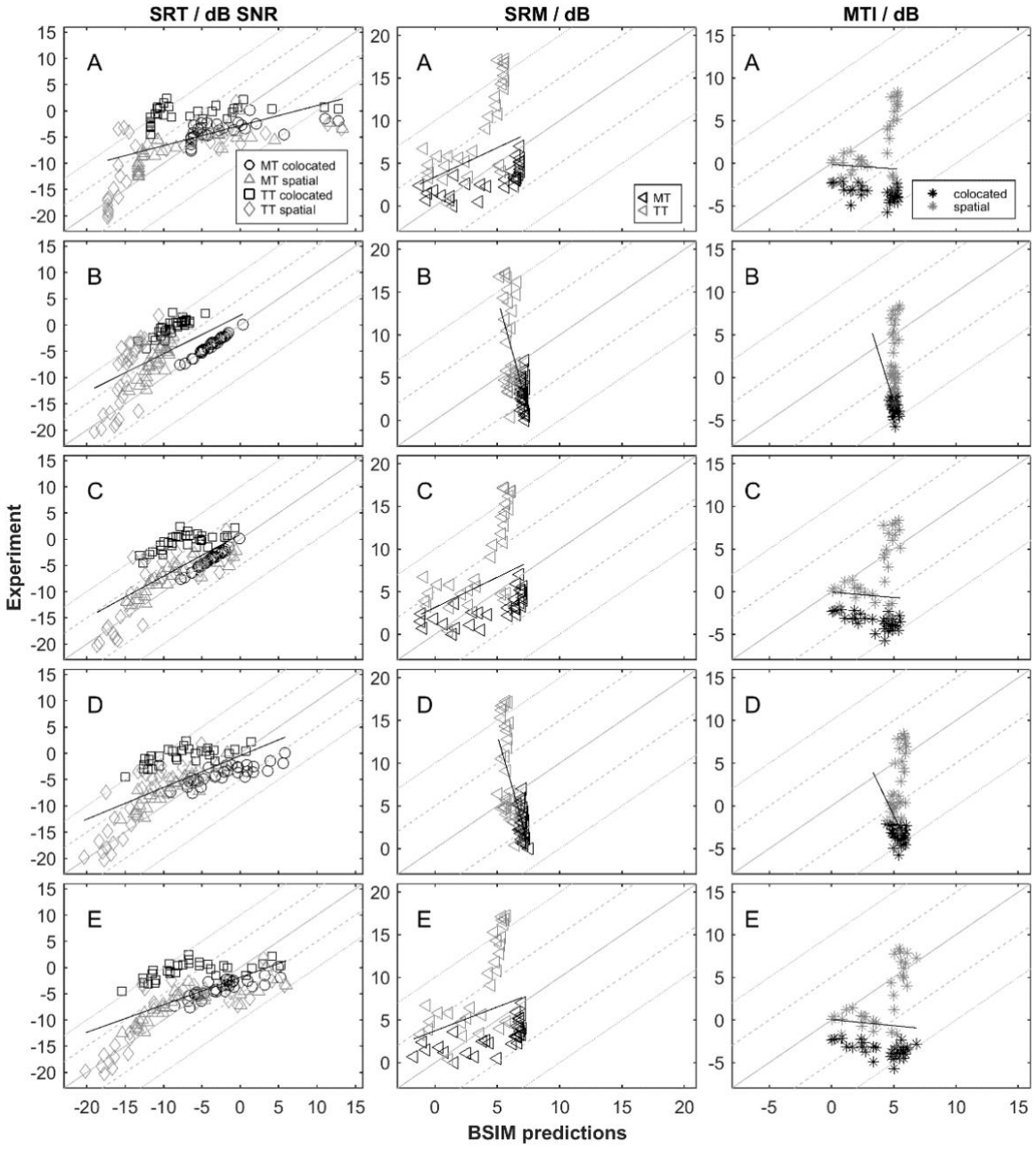

FIG. 2.3: COMPARISON OF EXPERIMENTAL DATA (ORDINATES) TO PREDICTIONS OF BSIM (ABSCISSAE). SRTS, SRM, AND MTI ARE ILLUSTRATED FOR THE DIFFERENT MODEL VERSIONS A TO E (ROWS). IN EACH PANEL DIFFERENT MASKING CONDITIONS ARE INDICATED BY GRAY SCALES AND/OR SYMBOLS. GRAY LINES SERVE AS VISUAL GUIDES AND REPRESENT PERFECT AGREEMENT BETWEEN PREDICTIONS AND DATA (SOLID) AS WELL AS DEVIATIONS BY ±5 DB (DASHED LINES) AND ±10 DB (DOTTED LINES). BLACK LINES REPRESENT LINEAR FITS BASED ON ALL DATA POINTS IN EACH PANEL.

SRT predictions of version A of BSIM based on normal-hearing reference value and individual audiograms (top left panel of Fig. 2.3, top part of Table 2.3) indicated that there were, in general, significant (rank) correlations between data and predictions. The lowest correlation ($R=0.42$, $\varrho=0.56$) was observed for SRTs in co-located TT maskers, while higher correlations were achieved in the other condition ($R\geq0.58$). However, some notable outliers were also observed: For some subjects SRTs were overestimated by up to about 17 dB (see rightmost data points), while other data points were above the diagonal by up to 12 dB, indicating that BSIM predicted thresholds much better than observed in the experiment. These deviations resulted in rather poor error measures (RMSE between 4.4 and 8.3 dB, MAE between 2.7 and 7.4 dB) and a rather shallow slope of the linear relation (≤0.51). Both of these measures were generally worst for the co-located TT masker. Concerning SRM predictions of version A (top mid panel of Fig. 2.3), it was observed that the individual rank order within each masker type was reasonably well predicted ($R\geq0.72$, $\varrho\geq0.71$), but that absolute errors of about 6 dB occurred for the TT masker. For the MT masker, error measures were acceptably small (about 2 dB). MTI could only be predicted for spatially separated maskers (significant correlations, error measures about 2 dB), but not for co-located maskers (no /negative correlations, error values of more than 6 dB).

TAB. 2.3: PREDICTION ACCURACY MEASURES FOR THE MODEL VERSIONS A TO E: LINEAR CORRELATION (R), RANK CORRELATION (ϱ), SLOPE OF A LINEAR FIT, BIAS (IN DB), ROOT-MEAN-SQUARE ERROR (RMSE, IN DB), AND MEAN ABSOLUTE ERROR (MAE, IN DB).

			R	ϱ	slope	bias	RMSE	MAE
A	SRT	MT cl	0.58	0.77	0.20	-1.3	4.4	2.7
		MT sp	0.73	0.86	0.31	0.0	5.6	4.2
		TT cl	0.42	0.56	0.10	5.6	8.3	7.4
		TT sp	0.70	0.80	0.51	-0.1	6.2	4.4
		overall	0.57	0.56	0.38	1.1	6.3	4.6
	SRM	MT	0.72	0.81	0.44	-1.3	2.3	2.1
		TT	0.73	0.71	1.64	5.7	6.8	5.8
		overall	0.37	0.31	0.69	2.2	5.1	3.9
	MTI	cl	-0.51	-0.45	-0.25	-6.9	7.3	6.9
		sp	0.64	0.62	1.65	0.0	2.8	2.2
		overall	-0.04	-0.33	-0.10	-3.4	5.5	4.6
B	SRT	MT cl	0.99	0.99	0.98	-0.1	0.2	0.2
		MT sp	0.90	0.90	1.64	4.0	4.4	4.0
		TT cl	0.83	0.83	0.73	8.4	8.5	8.4
		TT sp	0.82	0.86	2.93	5.4	7.4	6.3
		overall	0.63	0.60	0.73	4.4	6.0	4.7
	SRM	MT	-0.18	-0.24	-2.51	-4.1	4.4	4.1
		TT	-0.32	-0.29	-4.98	3.0	6.1	4.7
		overall	-0.66	-0.63	-5.28	-0.6	5.3	4.4
	MTI	cl	-0.30	-0.23	-1.41	-8.5	8.5	8.5
		sp	-0.08	0.04	-0.96	-1.4	3.8	3.5
		overall	-0.72	-0.76	-4.70	-4.9	6.6	6.0
C	SRT	MT cl	1.00	0.99	0.99	0.0	0.2	0.1
		MT sp	0.88	0.92	0.66	1.4	2.5	2.2
		TT cl	0.68	0.64	0.35	6.9	7.3	6.9
		TT sp	0.81	0.83	0.97	1.5	4.0	3.1
		overall	0.73	0.59	0.80	2.5	4.3	3.1
	SRM	MT	0.71	0.78	0.42	-1.4	2.5	2.2
		TT	0.72	0.70	1.55	5.3	6.6	5.6
		overall	0.39	0.33	0.71	2.0	5.0	3.9
	MTI	cl	-0.60	-0.56	-0.32	-6.8	7.2	6.8
		sp	0.58	0.55	1.58	-0.1	3.0	2.3
		overall	-0.05	-0.36	-0.13	-3.5	5.5	4.6

			R	ϱ	slope	bias	RMSE	MAE
D	SRT	MT cl	0.84	0.84	0.37	-2.0	3.3	2.5
		MT sp	0.89	0.89	0.74	2.1	2.7	2.4
		TT cl	0.67	0.62	0.25	6.8	7.6	6.9
		TT sp	0.85	0.85	1.36	3.7	5.2	4.2
		overall	0.68	0.61	0.60	2.7	5.1	4.0
	SRM	MT	-0.40	-0.23	-3.97	-4.0	4.4	4.0
		TT	-0.48	-0.50	-6.89	3.2	6.3	4.8
		overall	-0.71	-0.71	-5.04	-0.4	5.4	4.4
	MTI	cl	-0.34	-0.32	-0.71	-8.8	8.9	8.8
		sp	0.53	0.59	5.86	-1.6	3.7	3.3
		overall	-0.57	-0.70	-3.06	-5.2	6.8	6.1
E	SRT	MT cl	0.84	0.85	0.41	-1.5	2.7	2.0
		MT sp	0.83	0.91	0.43	0.0	3.8	3.2
		TT cl	0.59	0.59	0.18	5.8	7.4	6.7
		TT sp	0.78	0.82	0.67	0.4	4.7	3.7
		overall	0.66	0.56	0.53	1.2	5.0	3.9
	SRM	MT	0.66	0.74	0.40	-1.6	2.6	2.4
		TT	0.65	0.61	1.50	5.5	6.8	5.8
		overall	0.30	0.26	0.56	1.9	5.2	4.1
	MTI	cl	-0.53	-0.45	-0.24	-7.4	7.8	7.4
		sp	0.63	0.61	1.50	-0.3	2.8	2.2
		overall	-0.07	-0.34	-0.14	-3.9	5.9	4.8

Using model version B (individual reference SII, normal audiograms; second row in Fig. 2.3), the prediction patterns for the SRTs, SRM, and MTI changed in comparison to version A. In general, correlations between measured and predicted SRTs (left panel) were higher ($R \geq 0.82$, $\varrho \geq 0.83$) for all conditions. Note that the close-to-perfect agreement for co-located MT maskers is a result of fitting the individual SII values and not an indication for model accuracy (the fact the agreement is not perfect is due to the variability across randomly chosen stimuli for the ten repetitions of each condition). RMSE and MAE were similar to predictions of model version A, although SRTs were generally underestimated by version B (most data points were above the diagonal). The maximum deviation was about 13 dB. In contrast, interindividual variability of SRM was not predicted by version B: all predicted SRM

values were similar (within about 2 dB) and correlations between data and predictions were negative. The same was true for MTI, which was not predicted for either spatial condition.

Predictions of model version C are shown in the third row of Fig. 2.3. As for version A, SRT predictions were poorest for the co-located TT masker (R=0.68, ϱ=0.64, error measures about 7 dB), but all correlations were better than for version A and comparable to version B. Error measures were always better than for version B. RMSE and MAE for spatial maskers were also about 1-2 dB better than for version A, and similar for co-located TT maskers. With a few exceptions, all predicted SRTs except for MT maskers and spatially separated TT maskers were within ±5 dB of the experimental data (the maximum deviation of the exceptions was about 10 dB). In contrast, SRTs for co-located TT maskers were systematically underestimated by about 7 dB. With respect to SRM (mid panel), predictions were very similar to version A, i.e., clearly better than predictions of version B. The same was observed for MTI (right panel), i.e., MTI was predicted reasonably well for spatial maskers, but not for co-located maskers.

Versions D and E were included to test how much worse predictions become when the individual reference SII values were derived from the Goesa. Comparing the second and fourth row (version B vs. D, normal audiograms), it was found that the prediction patterns were qualitatively very similar. The largest deviations were observed for co-located MT maskers, which had been used as an individual reference in version B and hence produced an almost perfect agreement. In version D, SRT predictions for this condition were no longer perfect, but the correlations with the data were reasonably high (R=0.84, ϱ=0.84), while the error measures were about 2-3 dB. For the co-located TT maskers, correlations were also lower by 0.16 (R) and 0.20 (ϱ) compared to version B, while correlations were very similar for spatial maskers. The error measures of the SRT predictions were slightly lower (about 1-2 dB) for version D than for version B. With respect to SRM and MTI, there was a slightly increased variability in the predictions but, as observed for version B, version D failed to predict the considerable differences between subjects observed in the experiment. Similarly, the comparison of model versions C and E (third vs. fifth row of Fig. 2.3) revealed that both model versions produced very similar prediction patterns. The main difference in the SRT predictions was again the reference condition, where correlations were still reasonably good (R=0.85, ϱ=0.84, RMSE=2.7 dB). For all other conditions, correlations were slightly better for version D than for version E, but this difference was small (about 0.04 on average), while error measures were very similar for both model versions. With respect to SRM, the same limitations as for

version D were also observed for version E, i.e., SRM predictions clustered around the diagonal for MT maskers, but were overestimated by about 6 dB for TT maskers. Likewise, the variability in MTI was only predicted in part for spatial maskers, but not for co-located maskers (as for version D).

2.3.6. SUB-GROUP ANALYSES

As described above, the individual SRT data revealed some notable individual trends. In particular, some subjects performed considerably worse or better than would be expected from the performance of other subjects with very similar hearing loss or age. Similarly, there were some extraordinary deviations between the data and predictions of BSIM, although the overall trends seemed to be predicted reasonably well (at least by model version C). The group of young subjects in the present study, which was also normal-hearing or close to normal-hearing, showed the best performance both in terms of SRT and the SRT-difference measures (SRM and MTI). In contrast, the sub-group of older subjects with close-to-normal hearing (subjects #16, 23, 24, and 27; aged ≥81; PTA 21.25-30 dB HL, see Fig. 2.1) showed clearly elevated SRTs compared to young normal-hearing listeners in all conditions, a clearly reduced SRM (see Fig. 2.2), and also an absent or negative MTI even for spatial maskers, for which young normal-hearing subjects had a strong positive MTI. In comparison, another sub-group of older listeners with significant hearing loss (#17, 18, 19, 22, and 29; aged 71-75; PTA 46.25-55 dB HL, see Fig. 2.1) was qualitatively undistinguishable from this group, i.e., their performance was not markedly worse despite their hearing loss. Another interesting trend emerged for another sub-group with normal and close-to-normal hearing and medium age (subjects #8, 9, and 13; aged 56-62; PTA 21.25-25 dB HL, see Fig. 2.1). These subjects had somewhat elevated overall SRTs compared to the young normal-hearing group, but otherwise "normal" SRM and MTI results. A closer investigation of the individual data therefore seemed justified. To systematically assess the predictability of the experimental data within the framework of BSIM, two sets of about equally large sub-groups within the highly heterogeneous subject pool of the present study were built, i.e., one set based on age and another set based on hearing loss. Experimental data and model predictions were then compared for each subgroup. The grouping according to *better-ear PTA* was performed based on grades of hearing impairment as proposed by Martini (1996), which resulted in groups of seven normal-hearing subjects (aged 27 to

45 years), ten slightly hearing-impaired subjects (aged 45 to 85 years), and thirteen moderately hearing-impaired subjects (aged 31 to 75 years; see section II.A). To group the whole set of 30 subjects according to their *age*, a k-means clustering algorithm (Lloyd, 1982) was applied, which comprises an iterative procedure that assigns every observation to exactly one of a pre-defined number of clusters. The number of clusters was set to three to be comparable with the grouping according to hearing loss and to avoid too small sub-group sizes. The resulting groups consisted of nine, twelve, and nine subjects, respectively. The first (youngest) group varied in age from 23 to 37 years (mean 29.7 years), and in PTA from -2.5 to 42.5 dB HL (mean 12.1 dB HL). The second group varied in age from 45 to 67 years (mean 58.3 years), and in PTA from 2.5 to 70.0 dB HL (mean 39.4 dB HL). The third (oldest) group consisted of subjects from 71 to 85 years of age (mean 76.7 years), with PTAs spread from 21.3 to 55.0 dB HL (mean 39.7 dB).

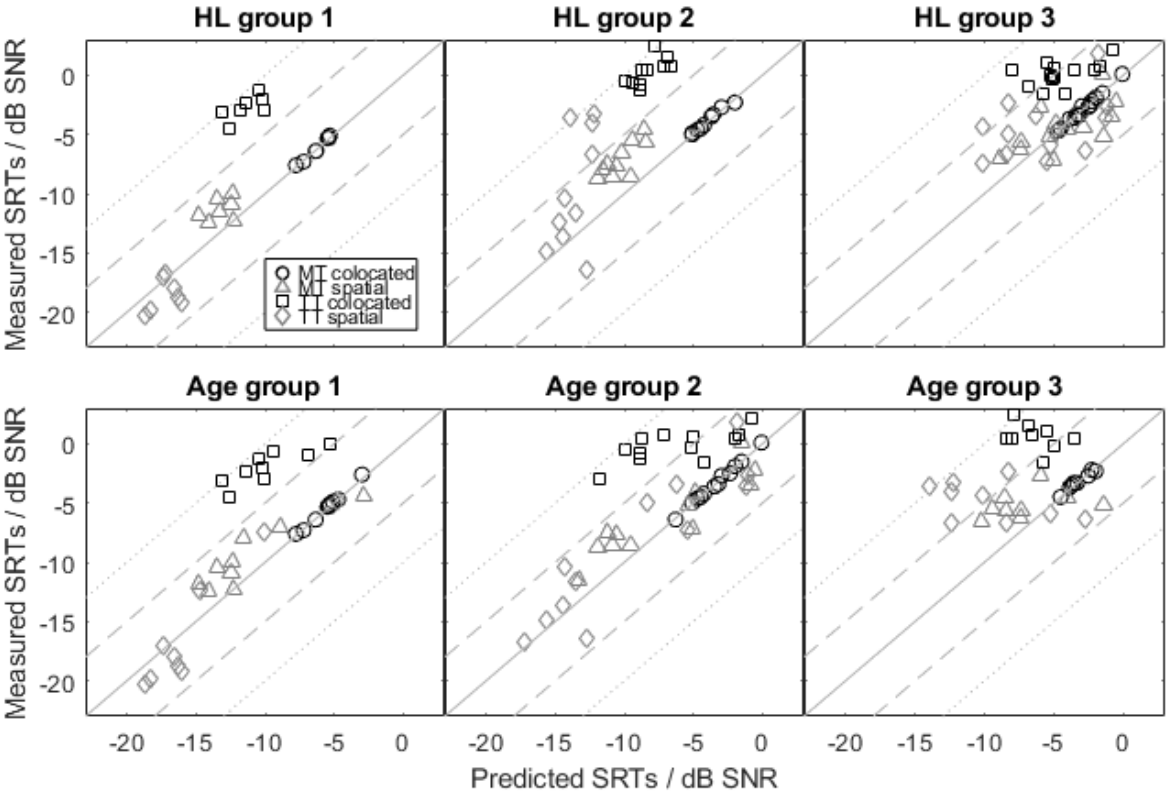

FIG. 2.4: COMPARISON OF EXPERIMENTAL SRTS (ORDINATES) TO PREDICTIONS OF BSIM VERSION C (ABSCISSAE) FOR SUBGROUPS BASED ON HEARING LOSS (TOP ROW) AND BASED ON AGE (BOTTOM ROW). IN EACH PANEL DIFFERENT MASKING CONDITIONS ARE INDICATED BY GRAY SCALES AND/OR SYMBOLS. GRAY LINES SERVE AS VISUAL GUIDES AND REPRESENT PERFECT AGREEMENT BETWEEN PREDICTIONS AND DATA (SOLID) AS WELL AS DEVIATIONS BY ±5 DB (DASHED LINES) AND ±10 DB (DOTTED LINES)

The analyses presented in the following were limited to model version C, because this was the only model version that predicted at least parts of the observed trends both in SRTs as well as in SRM/MTI with reasonable accuracy. The comparison of experimental and predicted SRTs is depicted in Fig. 2.4 for the three subgroups based on hearing loss (top row) and based on age (bottom row). Conditions are encoded using the same symbols and gray scales as in Fig. 2.3. As before, the reference condition (co-located MT maskers, shown as circles) should not be included when assessing the prediction accuracy, because the individualization of the model was based on this condition. For the normal-hearing subgroup (top left panel), data points of both spatial TT and spatial MT masker clustered around the diagonal at values below those of the reference condition, i.e., the model correctly predicted the lower SRTs as well as the differences between the two spatial maskers. In contrast, SRTs for the co-located TT masker were underestimated by between 7 and 10 dB. For the slightly hearing-impaired subjects (top mid panel), this pattern was partly preserved, although SRTs for spatial MT maskers were generally underestimated (i.e., SRM is overestimated) by about 3 dB. Notable deviations were observed for spatial TT maskers, where predicted SRTs were similar for all subjects (range <3 dB), while measured SRTs ranged from -16.4 to -3.2 dB SNR. The deviation for co-located TT maskers were similar to the normal-hearing group. For the subgroup with highest hearing loss, data points clustered along the diagonal at the upper towards higher SRT values. The spread was larger than for the individual clusters observed for the normal-hearing subgroup, although most of the predictions were within ±5 dB of the measured SRTs.

With respect to the subgroups based on age, predicted SRTs were in good agreement with experimental data for the youngest group (bottom left panel) except for the co-located TT maskers. The same was observed for the second age group (bottom mid panel), where all individual predictions (except for co-located TT maskers) were within ±4 dB of the experimental data. For the oldest subgroup (bottom right panel), data points clustered well around the diagonal for spatial MT maskers, but predictions for spatial TT maskers did not agree well with the data. In particular, SRTs were underestimated by about 9-10 dB for three subjects. These were subjects #23, 25, and 27, i.e., the subjects aged 81 years and older but with PTAs close to the normal range. For the same subjects, the difference between data and model predictions did not exceed 3-4 dB for the spatial MT masker.

2.4. DISCUSSION

2.4.1. SRT, SRM, AND MTI DATA FOR NORMAL-HEARING LISTENERS

About one third of the present subjects had normal hearing, and their data can be compared to previous studies focusing on normal-hearing listeners in similar conditions. As expected, all subjects scored better (lower) SRTs when maskers were spatially separated than when they were co-located. This is not surprising, and was reported by many studies (e.g., Cherry, 1953; Bronkhorst and Plomp, 1992; Culling et al., 2004; Kidd et al., 2016; Ewert et al., 2017). The present data can quantitatively be compared to data of Ewert et al. (2017), who employed the same procedure, target signals, TT maskers as well as spatial source constellations to measure SRTs in normal-hearing listeners. For the subgroup of normal-hearing subjects, SRT results of the current study for the TT maskers in co-located and spatially separated conditions coincided well with data of Ewert et al. (2017) with mean SRTs difference between studies of only 1 and 3 dB for abovementioned scenarios, respectively. The observed SRM was also similar for normal-hearing listeners in both studies.

When maskers were co-located with the target talker, subjects scored on average lower SRTs when the masker was MT babble. In contrast, when makers were separated from the target talker the opposite occurred and participants performed better in the presence of a TT masker. This dependence of MTI on spatial configuration was also in line with data of Ewert et al. (2017), who used a stationary SSN instead of MT babble. They found that SRTs reduced by about 5 dB when a co-located TT masker was replaced by a co-located SSN masker, which is similar to the 3.5-dB reduction observed for normal-hearing subjects in the present study. Similarly, Ewert et al. (2017) observed a positive MTI for spatial maskers (i.e., SRTs increased in SNN compared to TT maskers) of about 4 dB. This is somewhat lower than the average MTI of the seven normal-hearing listeners of the present study, which may be due to different listener groups (the standard deviation for spatial TT maskers was >5 dB in Ewert et al. (2017), or due to differences in masker type (MT vs. SSN).

At least two effects may have driven the observed interaction of MTI and spatial separation of target and maskers. On the one hand, the TT masker had much stronger temporal

fluctuations, including envelope minima that potentially enabled the listeners to exploit dip-listening at times when the masker energy was low. On the other hand, the TT masker consisted of two intelligible talkers of the same gender as the target talker uttering sentences of the same build (name-verb-numeral-adjective-object) with similar speed, while the MT masker consisted of unintelligible speech babble. This means that the TT masker potentially induced a much larger degree of IM than the MT masker. Kidd et al. (2016) showed that both the gender relation of target and interferers as well as their spatial separation have a large impact on SRTs of normal-hearing listeners. They found that both a change of masker gender (for co-located maskers) and a spatial separation of same-gender maskers to ±90° produced approximately the same SRT reduction of about 20 dB relative to the co-located, same-gender reference condition (for which the SRT was about -3 dB SNR in their study, i.e., similar to the present normal-hearing data). Kidd et al. (2016) also included a control condition in which the influence of the respective maskers was reduced to energetic masking by using ideal time-frequency segregation that retained only those time-frequency bins that were available after accounting for energetic masking. Effectively, this kind of processing performs the segregation of the target from the interferers for the listener and renders the masker unidentifiable, thereby eliminating its IM character. Kidd et al. (2016) found that SRTs were very similar in this control condition for both the gender cue and the spatial cue and that, hence, energetic masking was similar for both cues. As a result, they argued that the SRT decrease in the spatial condition was due to a release from IM rather than due to an increased availability of speech "glimpses" due to envelope fluctuations of the maskers. In this light, the large spatial benefit observed for TT maskers in the present study could be interpreted as a large release from IM. For the MT masker, the influence of IM was much smaller and, hence, so was the benefit due to spatial separation.

Other studies also investigated the impact of MT babble maskers on speech intelligibility of normal-hearing listeners in conditions without spatial unmasking, i.e., in conditions comparable to the co-located MT masker in the present study. Simpson and Cooke (2005) measured consonant identification as a function of the number N of talkers in a MT babble (N = {1, 2, 3, 4, 6, 8, 16, 32, 64, 128, 512}). They found that the babble consisting of 8 talkers had the greatest masking impact and argued that this was due to the most detrimental combination of IM (largest for small N) and absence of temporal masker modulation (largest for large N). Comparing their results for 2- and 8- talker masker condition (which is closest to the TT and MT masker employed in the present study) the consonant identification score

dropped rapidly by over 30% for the 8-talker babble. This indicates that a 2-talker masker should produce lower SRTs than 8-talker maskers for co-located masker conditions. This was not observed in the current study, where the reported positive MTI (i.e. SRT decrease) of 7.2 dB was observed. One possible reason for the discrepancy is that the interaction of IM (which is stronger for TT than for MT maskers) and the potential to benefit from masker amplitude modulations (which is also larger for TT than for MT) is different for the stimuli and procedure applied here than for consonant identification tasks as, e.g., applied by Simpson and Cooke (2005). In contrast, Freyman et al., (2004) measured speech recognition using nonsense sentences and found that SRTs decreased for co-located 10-talker babble maskers compared to TT maskers, while SRTs increased when a "perceived" spatial separation was introduced between target and maskers (see also Freyman et al., 2001). Similarly, SRM was much larger for TT maskers than for MT maskers and Freyman et al. (2004) argued that this was due to a larger release from IM. Thus, the MTI and SRM data for the subgroup of normal-hearing subjects of the present study are in line with previous employing target sentences to measure speech recognition.

2.4.2. Inter-individual differences in SRT, SRM, and MTI data

In addition to normal-hearing subjects, the current study employed a wide variety of subjects both concerning their age (23 to 85 years) and hearing loss (PTA from -2 to 70 dB HL) with the goal to explore their speech recognition performance in complex listening conditions involving spatial hearing, energetic and informational masking, and exploiting masker envelope minima ("dip listening"). As expected, large inter-individual performance variations were observed in the present data both for the SRTs as well as for the SRT-difference measures (SRM and MTI), for which an individual baseline performance was subtracted. In general, SRTs were higher for hearing-impaired listeners than for normal-hearing listeners as expected from previous studies using similar speech-on-speech masking paradigms (e.g., Xia et al., 2015; Ellinger et al., 2017). Likewise, SRM was generally smaller than for normal-hearing subjects, which is also in line with previous studies (e.g., Beutelmann and Brand, 2006; Ellinger et al., 2017; Xia et al., 2015). The differences between subjects were especially pronounced in conditions involving both spatial unmasking and strong masker envelope modulations and a

presumed high impact of IM (i.e., for SRM in TT maskers, and for MTI in spatial maskers), so the listeners' ability to benefit from either factor or to suppress IM in spatially separated conditions varied considerably. The observed benefit was generally higher for subjects with a good baseline performance, i.e., lower SRTs. This could indicate that some subjects tended to be "generally better", regardless of the nature of available cues, while other subjects tended to be "generally worse". A similar observation was made by Swaminathan et al. (2015), who found that the individual benefit from masker time reversal was highly correlated to the benefit from spatial separation (i.e., for two segregation cues which presumably rely on very different stimulus properties). The same observation was made by Kidd et al. (2016) who, in addition, also found highly correlated individual benefits from masker gender difference and both spatial separation and masker time reversal.

Despite wide inter-individual variability all the subjects benefited from the spatial separation of maskers from the target speech source. This was not the case for the MTI measure. Subjects that did not benefit from the increase in temporal fluctuation of the masker or the release from IM in spatial masker constellations were those with the poorest SRTs, but also either high PTA and middle age (60-71 years) or those with advanced age (81-86 years), but close-to-normal PTA. This trend brings to attention the aspect of age and a role of possible cognitive factors, which is discussed further below.

2.4.3. Relation between speech recognition performance and individual factors

One purpose of this study was to investigate how individual performance in complex listening conditions could be predicted. In one approach, the relation between individual factors collected in clinical diagnostics (age, PTA, and subjective scores in the simpler and clinically applicable speech intelligibility tests - DTT and Goesa) and the measured speech recognition performance was investigated. These analyses showed that all of these factors were capable of predicting between 28% and 79% of the variance observed in the SRT data for the whole group of 30 subjects. This is not surprising since the group was selected to vary widely across all factors, and a large range of values more readily leads to high correlations. The factors with the highest predictive power across the great majority of conditions were SRTs (as measured by the Goesa)

and PTA. SRTs as measured by the DTT always had lower predictive power than Goesa SRTs. This may indicate that speech intelligibility tests employing more complex speech material (everyday sentences) may be more meaningful for predicting speech perception in complex scenarios than digit triplet tests, even if both are measured monaurally.

Of all individual factors investigated here, age had the lowest predictive power for SRTs measured with spatial and co-located MT masker, or with spatial TT maskers. The same was true for SRM measured with TT maskers. In contrast, age showed the highest correlations (R^2=0.55 compared to $R^2 \leq 0.43$) with SRTs measured for the co-located TT masker, i.e., the masker with the presumably highest degree of IM and the highest demand for cognitive processing in segregating the target from the maskers. No individual factor could predict a larger portion of the variance observed in the MTI data for co-located maskers (although some correlations were significant). The effect of age on SRM was also investigated by Ellinger et al. (2017), who tested both older normal-hearing as well as older hearing-impaired subjects with respect to their SRM and the relationships with aging and hearing loss. They found that age was a better predictor of SRT than hearing loss for both groups, which they explained by the correlations of age and high-frequency hearing loss observed. Nevertheless, no significant correlations between age and SRM were observed suggesting that SRM was less affected by age than by hearing loss, which was not found in the current study results. Furthermore, the authors found that there was only a weak impact of hearing loss itself on SRM abilities as well and suggested the possible existence of an unknown factor (or set of factors) to be even more important than age and hearing loss in the exploitation of binaural cues in complex listening conditions, such as working memory, cognitive ability, and temporal processing ability. For some subjects of the present study, age appeared to be more strongly related to speech recognition than their hearing loss as measured in terms of PTA. These subjects were close-to-normal hearing, but considerably older than the seven young normal-hearing subjects. Their SRTs were considerably higher (and their SRM poorer) than for young normal-hearing subjects. In fact, they were not better in performance than other subjects of similar age but with considerable hearing loss. This reduction in SRT performance might thus be hypothetically linked to, e.g., age-related decline in cognitive skills. Such findings, especially relating older subject with normal or close-to-normal hearing with very poor SRT performance have also been discussed in previous studies (e.g., Humes et al., 2006, 2009; Lee, Humes, 2012), which concluded that audibility does not ensure understanding when a target signal is embedded

within a masker signal consisting of one or more competing talkers leading to both energetic masking and IM. The results of SRM being affected by mainly age among normal- or close-to-normal-hearing subjects are also in line with the findings of Füllgrabe et al. (2015), where only normal-hearing individuals were tested, but widely spread in age. This study showed that, despite similar PTA, subjects' speech identification in noise declined with increasing age and authors expressed the need to take age into account while examining the effects of hearing loss. These findings, pointing to age as one of the main performance-related factors, are in line with the current study results, where both mean as well as individual performance of the oldest subjects hinted at a decreasing performance with increasing age, which was not consistently observed for increasing hearing loss (PTA).

In conclusion, the present data are in line with several previous studies that age is a relevant factor in speech recognition performance. For conditions with the highest degree of IM, age explained a larger degree of the variance than the other individual factors, which was particularly apparent for older subjects with close-to-normal audiogram but considerably reduced performance.

2.4.4. Predictability of individual performance using a quantitative binaural prediction model

If a quantitative model could reliably predict individual speech recognition performance based on a limited set of individual factors, it could provide valuable information, e.g., for diagnostic purposes (how well should a patient be able to perform given a set of diagnostic measures?) or the fitting or design of individualized speech enhancement strategies (which parameters will produce the largest benefit for this individual?). BSIM was shown to provide good prediction accuracy both for normal-hearing and hearing-impaired subjects in simpler listening conditions (e.g., using a single stationary noise source, see Beutelmann and Brand, 2006; Beutelmann et al., 2010). The present study investigated how well BSIM could predict the data of a highly heterogeneous subject group in more complex listening conditions. To this end, different methods for individualizing the model predictions were tested. The comparison of them showed that using individual SII references and normal audiograms (version B) improved SRT predictions compared to using a normal-hearing SII reference and individual audiograms

(version A). However, this destroyed the capability of BSIM to predict SRM, which was reasonably predicted at least for MT maskers by version A. In contrast, using both individual SII references and individual audiograms maintained the improved SRT predictions as well as the predicted trends in SRM. Larger deviations between SRT predictions of this model version and experimental data were observed only for co-located TT maskers, which were considerably underestimated. This reflects the fact that BSIM can account for energetic masking and spatial unmasking, but not for IM. Accordingly, SRM for TT maskers was considerably underestimated by the model, reflecting the fact that subjects experienced a large spatial release from IM not accounted for by the model. Similarly, interindividual differences in MTI were predicted only for spatial maskers, but not for co-located maskers. In summary, model version C achieved reasonable prediction accuracy for all conditions and SRT-differences not involving the condition high in IM. This makes BSIM applicable in principal to speech-on-speech masking conditions except for conditions very high in IM (specifically, co-located, same-sex maskers consisting of few intelligible talkers uttering similar sentences as the target talker). As mentioned above, it was shown, e.g., by Kidd et al. (2016), that IM was strongly reduced when the similarity between target and masker talkers were reduced even for co-located maskers (e.g., by using sex differences). Similarly, it was found that even for symmetric maskers a considerable portion of the maximum amount of spatial unmasking was reached already at masker azimuths of ±15° (the full amount was reached at ±45°, see Marrone et al., 2008). Therefore, it seems likely that the considerable differences between model predictions and data are specific to reference conditions explicitly designed to induce a high degree of IM but are less pronounced in more realistic conditions with at least a small degree of spatial separation and a larger degree of dissimilarity (such as different sex, different speech material).

The division of the subjects into subgroups showed that the model's failure to predict speech recognition in the high-IM conditions was not specific to any hearing loss or age group but was observed regardless of how the subjects were grouped. For the other conditions, the prediction accuracy was comparable to previous studies evaluating BSIM with normal-hearing and hearing-impaired subjects (Beutelmann and Brand, 2006; Beutelmann et al., 2010; Brand et al., 2017). Individual predictions were reasonably accurate for almost all subjects, except those with close-to-normal audiograms but considerably advanced age, for which SRTs were considerably underestimated for spatial TT maskers. Compared to the other subjects with (close-to-) normal hearing, these subjects had considerably reduced SRM for both masker types

(see Fig. 2.2). It appears that their reduced ability to exploit spatial separation (for any masker) went along with an increased distraction by TT maskers even when they were spatially separated. Since model predictions and data agreed reasonably well for spatial MT maskers for these subjects, but not for spatial TT maskers, this may indicate that these older "normal-hearing" subjects were more strongly affected by IM than their younger counterparts in spatial masking conditions. As summarized above, it was shown in several studies that spatial separation can strongly reduce IM, and the present data suggest that the spatial benefit can be considerably reduced in older subjects with approximately normal hearing. One possible explanation could be that the reduced general ability to use spatial cues reduces the "perceived" spatial separation between target and maskers, thereby reducing the release from IM (cf. Freyman et al., 2001). Not surprisingly, these effects were not predicted by BSIM since IM cannot be accounted for. While it seems generally possible to introduce an age-related correction to improve prediction accuracy, it is not straightforward how to extend BSIM (or in fact any current speech intelligibility model) to predict IM and release from IM. The present data suggest that whatever functional model extension will be developed in the future, it should not be specific to normal-hearing- or hearing-impaired subjects or subjects of any age group.

Model versions D and E were included in the present study to test how much worse predictions became when the individual SII reference was not derived from a condition similar in complexity and target material to the remaining test conditions, but from a more standardized speech intelligibility test. This may be important for clinical applications of BSIM, because tests such as the Goesa may be readily available in practice and it is important to know if improved predictions can only be achieved when tailored to specific experimental conditions. The comparison of model version C (SII references from experimental reference condition) and version E (SII references from Goesa) was promising in that correlations were only slightly worse (0.04 on average) while predictions errors were similar. Together with the data of Brand et al., (2017), who used individualized SII references within BSIM to predict SRTs of aided and enhanced speech, these results suggest that BSIM can potentially be used for applications with hearing-impaired listeners.

2.5. CONCLUSIONS

The following conclusions can be drawn from this study:

- In line with previous studies, SRTs in complex scenarios including energetic and informational masking, spatial unmasking and/or exploiting masker envelope fluctuations differed widely in the present group of subjects, which was highly heterogeneous with respect to four individual factors, i.e., their age, hearing loss, and performance in two standardized and clinically applicable speech intelligibility tests (DTT and Goesa).
- All four factors could explain a significant portion of the observed interindividual SRT variance. PTA and Goesa were the best predictors for all conditions except the conditions highest in IM (co-located two-talker masker). For this condition, age showed the largest determination coefficients, indicating that age may play a special role in susceptibility to IM.
- Predictions of the BSIM could be improved by using both individualized reference SII values (matched to one of the experimental conditions for each subject) and individual audiograms.
- Larger discrepancies between predictions and data remained in conditions high in IM for all hearing-loss or age groups.
- As a tool to support future model evaluation and development, the experimental data collected in this study along with the individual factors are accessible as supplementary material – see Appendix.

3. RELATION BETWEEN HEARING ABILITIES AND PREFERRED PLAYBACK SETTINGS FOR SPEECH PERCEPTION IN COMPLEX LISTENING CONDITIONS

This study investigated individual preferences for speech processing in different adverse listening conditions including speech maskers with two degrees of informational masking both in a co-located and a spatially separated constellation. Thirty subjects (the same that took part in the study described in chapter 2) differing widely in hearing status (normal-hearing to moderately impaired) and age (23 to 85 years) adjusted stimuli to their preferences using four signal modification schemes: i) linear gain, gain at the cost of ii) clipping distortions or iii) compression distortions and iv) frequency-shaping. The first set of preference adjustments was conducted for the speech signal only in fixed-masker conditions to investigate the preferred trade-off between distortions and noise disturbance. The second set of adjustments was made for speech and maskers simultaneously, i.e., with a constant signal-to-noise ratio. High test-retest reliability was found for all modification schemes except for frequency-shaping. The preference adjustments suggested that subjects could be consistently categorized along a scale from "noise haters" to "distortion haters", and that this preference trait remained stable through all maskers, spatial conditions, and types of distortions. Comparing listening preferences with individual factors like hearing loss, age, and speech intelligibility performance suggested that preferences in complex listening conditions correlated best with individual speech intelligibility data measured in the same conditions.

3.1. INTRODUCTION

It is a well-known phenomenon that individual preferences for playback settings, e.g., with regards to sound level or frequency shaping, can differ markedly between listeners. Likewise, the benefit of different speech enhancement algorithms often differs dramatically even within groups of listeners which are rather homogenous with respect to, e.g., their hearing status or age. The origin of this variability is not yet fully understood. The main goal of this study was to investigate individual listening preferences for speech playback in adverse listening

scenarios, and how they relate to personal factors such as age, hearing loss, or speech recognition performance.

Characterizing individual listeners is a very common concept to adapt audio playback to their individual needs and preferences. One very typical measure is the PTA, as a first indicator of a listener's hearing abilities, and many commonly used fitting rules for hearing aids are based on information contained in the audiogram (e.g., NAL-NL2; Keidser et al., 2011). It is, however, well known that amplification settings derived from the audiogram are not sufficient to provide individually optimized hearing device settings and that often an intensive fine-tuning process is required. Such fine-tuning often involves an assessment of individual loudness perception, i.e., supra-threshold perception of audio stimuli, which can provide insights into a listener's most comfortable listening level, uncomfortably loud levels, or loudness growths functions. This is often used to adjust desired input/output characteristics of hearing aids with automated gain control (e.g., Moore et al., 1992; Kießling et al., 1996). Although loudness is a rather basic psychophysical quantity and various methods exist to reliably measure loudness perception, its prediction is still a challenge and the underlying factors are not fully understood yet. For example, Brand and Hohmann (2001) reported that subjects with similar audiograms may differ significantly regarding their loudness functions. Similarly, Oetting et al. (2014) found that subjects with similar hearing thresholds could have profoundly different perception at supra-threshold levels, e.g., with respect to comfortable loudness or levels at which stimuli became uncomfortably loud, especially in binaural listening scenarios. In other words, despite being rather homogenous, as indicated by the audiogram, listeners can differ markedly in the level range in which they prefer to listen to auditory stimuli. Völker et al. (2018) presented a conceptual approach to using these interindividual differences in the hearing-aid fitting. The approach is based on the assumption that these individual listening preferences or listening needs constitute "personal traits", which are stable over time at least in similar listening conditions. For the different preferences in loudness, the corresponding trait, as described by Völker et al. (2018), would be the classification of a subject as "power junkie" or "amplification hater".

Another dimension along which individual preferences can differ substantially is the tradeoff between having a better SNR and the degree of distortions introduced by the signal processing to remove noise from a signal. This tradeoff is a typical problem in hearing-aid fitting, where the degree of noise reduction has to be adjusted to meet individual needs. Some

subjects prefer moderate noise reduction (few artifacts, but also more residual noise and a lower SNR) and could be described by the personal trait "distortion hater". In contrast, other subjects ("noise haters") prefer aggressive noise reduction with less residual noise, but also more artifacts. A set of predictors for such listeners' needs is still being investigated. For instance, Luts et al. (2010) investigated how individual noise tolerance and distortion tolerance was related to basic audiological and cognitive factors. They found negligible influence of factors such as PTA and working memory capacity on response to binaural noise reduction algorithms. Neher et al. (2014) aimed at relating several acoustical, audiological, and cognitive measures to noise reduction preferences and speech recognition performance. They found a reasonably high test / re-test accuracy of individual preferences, supporting the concept of stable personal traits. However, while their data suggested that larger (worse) PTA may be linked with preferences for stronger noise reduction, none of the investigated measures was a strong predictor for preference in the noise-reduction strength.

The tradeoff between SNR and distortions is also present for other nonlinear signal manipulation schemes, such as dynamic range compression or peak clipping as methods employed during signal amplification. Hawkins and Naidoo (1993) measured the sound quality and clarity of speech in silence, speech in noise as well as music in subjects with mild to moderate hearing loss. Stimuli were processed either by compression or peak clipping. They reported a significant preference for output limiting compression rather than for peak-clipping for all three types of stimuli. The preference became stronger with increasing degree of distortions. Dawson et al. (1990), on the other hand, investigated preferences of profoundly hearing-impaired subjects performing speech recognition tasks via a master hearing aid incorporating a peak clipping and a compression limiting method left for the choice of the subjects. They found that subjects with better speech recognition abilities preferred compression limiting, while those with worse speech recognition opted for peak-clipping. Yet another dimension along which audio manipulation is typically applied is frequency-dependent gain to adjust the degree of bass, mid-frequencies, or treble. It is a common daily observation that listeners can have widely different preferences regarding how they set the EQ in their audio devices. It has also been reported true for listeners with very similar age-related hearing loss (Rennies et al., 2016), so it is probably oversimplified to assume that hearing-impaired listeners tend to counteract their hearing loss characteristics by only amplifying signals in the frequency range where their hearing is most impaired.

In summary, interindividual differences regarding listening preferences can be large, and they occur in several dimensions (e.g., preferred loudness, SNR, noise vs. distortion tolerance, frequency shaping, etc.). It is currently unknown what the underlying mechanisms are or how individual preferences vary in complex listening conditions involving, for example, binaural unmasking, energetic masking, informational masking, or dip listening. This study aimed at shedding more light on the factors contributing to individual listening preferences. One goal was to determine if (and to what degree) individual listening preferences are stable over time in complex listening scenarios, which was assessed by conducting a retest session about one week after the first session for the same listening conditions and tasks. Another goal was to investigate how individual preferences of the same subjects vary depending on the listening conditions involving different maskers and spatial constellations of sound sources. A further goal was to investigate the relation between individual listening preferences and basic individual factors such as age, hearing loss, and speech intelligibility performance in standardized speech tests (such as diotically measured tests with simple digits or established sentence test – here DTT and Goesa). Finally, this study addressed the question of how individual preferences relate to individual speech recognition performance, which had been measured in the same subjects and listening conditions in the experiment described in chapter 2. The key approach of this study was to measure individually preferred playback settings along the dimensions such as loudness, noise-vs.-distortion tradeoff, and frequency shaping in subjects with strongly varying hearing-loss, age, and speech recognition performance. Knowledge gained from this study is intended to serve as a basis for creating profiles of "personal traits" that may help to improve the personalization of hearing devices and hearing support technologies.

3.2. METHODS

3.2.1. SUBJECTS

Thirty subjects (18 female, 12 male) aged 23-85 years (mean 55.2 years with a standard deviation of 19.6 years) participated in this study (the same group that had previously participated in the experiment described in chapter 2 – please refer to section 2.2.1). They were

selected from the database of the Hörzentrum Oldenburg, Germany, to span a wide range along the following individual factors: their age, their PTA, and their speech intelligibility performance in formal diagnostic intelligibility tests (DTT and Goesa). In addition to these basic auditory performance measures, SRTs of these subjects were known for the same listening conditions as investigated in this study (see section 2.3). Fifteen subjects were hearing-aid users but performed the measurements of the present study unaided. Subjects received an hourly compensation for their participation.

3.2.2. APPARATUS

The study was conducted using the same apparatus as in the study from chapter 2 (please refer to section 2.2.2): a Matlab software environment installed on a personal computer, an RME Fireface UC USB High - Speed Audio Interface soundcard, and a Tucker-Davis Technologies HB7 headphone driver. The stimuli were presented to the subjects in a sound-attenuated booth via Sennheiser HD 650 headphones that were calibrated to dB SPL using a Bruel&Kjær (B&K) 4153 artificial ear, a B&K 4134 microphone, a B&K 2669 preamplifier, and a B&K 2610 measuring amplifier. The impact of the headphones was free-field equalized using a finite impulse response filter with 118 coefficients.

3.2.3. STIMULI AND MEASUREMENT SCENARIOS

Both stimuli and measurement scenarios were the same as the ones used in the study described in chapter 2, section 2.2.3. Target speech stimuli consisted of sentences from the Oldenburg sentence test (Wagener et al., 1999), which are built with the fixed order (name-verb-numeral-adjective-object), are grammatically correct, but semantically unpredictable. The sentences were spoken by a male talker. The target speech was convolved with head-related room impulse responses (HRIRs) corresponding to frontal incidence (0 degrees) in the horizontal plane. HRIRs were taken from the database of Kayser et al. (2009), recorded with a B&K head and torso simulator (1904128C) in an anechoic environment, with a distance to the speaker of 80 cm. The target speech material was temporally centered in the 5-s long masker signals. The

duration of the sentences varied from 1.9 to 2.9 s, which resulted in pre- and post-sentence masker presentation of about 1-1.5 s.

In addition to speech in silence, two different masking conditions were used:

- Competing sentences from the Oldenburg sentence test spoken by two other male talkers than the target talker (recordings from Hochmuth et al., 2015) - named in the following "two-talker (TT) condition". The two masking talkers uttered different sentences with randomly selected starting positions so that the sentence rhythm randomly differed between the two maskers as well as between the maskers and the target sentences;
- Multi-talker babble consisting of 10 male and female talkers, cut and remixed so that the meaning as well as language of each talker were not recognizable - named in the following "multi-talker (MT) condition";

Target speech and maskers were presented in two different spatial conditions, motivated by the study of Schubotz et al. (2017):

- co-located (indicated as "cl" in figures and tables): target speech and maskers were convolved with HRIR of 0 degrees; hence all signals were presented from the front;
- spatially separated (indicated as "sp" in figures and tables): target speech was presented from the front and the maskers were symmetrically spread out from the target source on the horizontal plane 60° in both directions. Different segments of the multi-talker babble or competing sentences were used for both sides.

All maskers were equalized in 1/3-octave bands to match the long-term spectrum of the target speech (adjustments made during the experiments introduced differences between speech and noise, see below). During the experiments subjects could modify either the speech signal only (while the maskers remained fixed), or both speech and maskers simultaneously (for details of adjustments see next section). The starting level of the speech stimuli was individually adjusted relative to the SRTs of unprocessed speech for each masker and spatial condition. In conditions for which subjects adjusted the speech signal alone, the masker level was fixed at 65 dB SPL A-weighted (dB A). The speech level varied depending on individual SRT and the adjustment performed by the subject, i.e., the lowest possible speech level corresponded to the speech level at SRT for the given spatial and masker conditions. In conditions for which subjects adjusted both speech and noise simultaneously, the level of the maskers was set to

65 dB A and the speech level was adjusted to the individual SRT plus 9 dB to ensure very good speech intelligibility. The starting levels of speech stimuli presented in silence were set to the level corresponding to individual SRTs obtained in a diffuse masker with the same target material for the study part when only modification of the speech signal was possible, and the same level incremented by 9 dB when both target and maskers signals were adjustable simultaneously.

3.2.4. PROCEDURE

The experiment consisted of two main parts. In the first part (called adjS), subjects could introduce changes only to the target speech signal, while the maskers remained unchanged. In other words, by modifying the speech in constant background noise, subjects could vary the similarity / dissimilarity between target and maskers (e.g., with respect to level or frequency shaping). In the second part (called adjSN), each change introduced by the subjects influenced both target and masker signals in the same way (i.e., target and masker signals were pre-mixed before processing). The task of the subjects was to perform signal adjustments according to their individual listening preference. The signal modifications occurred in real time while subjects utilized the user interfaces (UIs) described below to find their preferred setting by experimenting with the available parameter range and immediately experiencing the resulting perceptual change. The adjustments consisted of varying the following four sound parameters for every masker and spatial constellation condition:

3.2.4.1. LINEAR BROADBAND GAIN

Subjects were asked to adjust the volume setting to their preferred level (linear broadband gain adjustment). The UI consisted of a horizontal slider without any description and with an original gain range of 30 dB. This range could be limited depending on the presentation level, which was set relative to the individual SRTs as described above. If the individual starting level was higher than 65 dB A, the range of the slider (allowing 30 dB gain by default) was limited so that the maximum gain accessible from the UI did not lead to an overall level of the output signal exceeding 95 dB A. The gain range accessible from the slider (in dB) was then divided

into 100 parts, which were equally spaced along the length of the slider, making each value accessible by drag-dropping the slider or by using arrow handles at the edges of the slider. In order to reduce the possible tendency of subjects to use the slider position as a visual anchor from one adjustment to the next, the slider direction was randomized for each trial where this UI element was used, i.e., the gain increased from left to right or from right to left. Additionally, the initial slider position at the beginning of stimulus playback was also randomized but was limited to be within the lower 10% of the slider range to avoid startlingly high initial playback levels. In the adjS condition, the subjective adjustment introduced changes to the speech signal only (hence modifying the SNR) while, in the adjSN condition, both speech and masker were processed together (hence the SNR remained unchanged). These adjustments aimed at determining the preferred SNR (adjS) and listening level (adjSN), respectively.

3.2.4.2. GAIN AT THE COST OF CLIPPING DISTORTIONS

In this condition subjects adjusted the broadband gain to their preferred settings, but here peak clipping was performed at the same time. The user interface was the same horizontal slider without any descriptions. The default range of adjustable gain was 30 dB, and the range was limited (if needed) as described above to keep the output level of the signal below 95 dBA. The signals were processed such that increasing gain corresponded to increasing clipping distortions. The peak-clipping ranged from 0 to 80 percent of the samples clipped. This was achieved by dividing the gain range (in dB) as well as the 80% clipping range into 100 equal parts. These parts were then equally spaced on the length of the slider resulting in its range varying in equal steps from 0 to 30 dB (by default) and from 0% to 80% of the samples being peak-clipped, respectively. When the gain range of the slider had to be limited, only the accessible gain underwent such a limitation, while the accessible clipping percentage range was always constant. Both adjS and adjSN tests for this scenario were conducted with the same UI and processing scheme but differed in the signals processed. This condition aimed at determining the preferred trade-off between higher SNR and a higher degree of distortions (adjS), and the overall tolerance of distortions (adjSN), respectively. The direction of gain/distortion increase was randomized, and the initial gain was randomized but limited to the lower gain part as described above.

3.2.4.3. GAIN AT THE COST OF DISTORTIONS BY DYNAMIC RANGE COMPRESSION

In this condition, similar to the clipping scenario, subjects adjusted the broadband gain which was accompanied by a varying degree of signal distortions. Here, distortions were introduced by a broadband dynamic range compressor (DRC). The default range of the horizontal slider was again 30 dB, and an increase in gain was accompanied with an increase in compression ratio from 1:1 to 1:8. The range of compression ratio denominators (from 1 to 8) was equally spaced along the slider length in 100 steps. The DRC compressed the signal employing release and attack time constants of 5 ms and 400 ms, respectively. As for clipping, this condition aimed at determining the preferred trade-off between higher SNR and a higher degree of distortions (adjS), and overall tolerance of distortions (adjSN). The direction of gain/distortion increase was randomized, and the initial gain was randomized but limited - as described above.

3.2.4.4. EQUALIZATION

In the equalization (EQ) scenario, also referred to as frequency shaping scenario, subjects could choose their favorite spectral shaping of the signal using a two-dimensional UI. Two dimensions rather than one as used for the other adjustments were selected here because frequency shaping can be realized in various ways which cannot be ordered along a single dimension. Even two dimensions only allow a limited range of equalization options, but the accessible parameter changes as described below were considered a reasonable compromise between parameter freedom and usability for the subjects. The task of the subject was to move a point on the two-dimensional (2D) surface using the cursor of a touchscreen. The 2D matrix consisted of 19x19 presets. One axis ranged from low-frequency boost and high-frequency attenuation to high-frequency boost and low-frequency attenuation. The other axis ranged from mid-frequency boost to mid-frequency attenuation. The resulting equalization adjustment was a summation of presets from the x- and y- axes. As for the previous UI elements, the orientation of the matrix' axes were randomized to avoid subjects using visual anchors for their settings. This was achieved by switching x- and y-axes as well as the axes' directions (presets from 1 to 19 being assigned from left to right or vice versa). The accessible dynamic range of the presets' configuration was 30 dB. Each x-y-preset constellation had its own correction factor adjusted

with respect to the stimuli's long-term spectrum so that there was no change in broadband level when moving the cursor across the 2D surface. The broadband level of the signal(s) played in this scenario was set to the preferred level chosen by each subject in scenario 1) "linear broadband gain" for each masker and spatial condition to ensure a comfortable listening level and that subjects could focus on preferences for spectral content. The UI design and processing were the same for both adjS and adjSN. This condition aimed at determining the preferred spectral dissimilarity between speech and maskers (adjS) or a general preference for frequency content for speech in noise (adjSN). The initial position of the cursor was set to a random position anywhere on the 2D surface for each adjustment.

The order of these four adjustments was quasi-randomized for each subject in the adjS the adjSN part of the experiment, which were held separately (first adjS, then adjSN). In order to measure the subjectively preferred presentation level that was used in EQ scenario, this scenario had to be measured after subjects had performed the gain scenario for a particular masker and spatial condition. Apart from this constraint, all adjustments and masker conditions were randomized for each subject and session. Prior to the measurement session subjects performed a training session where they could test how each of the user interfaces (slider and 2D matrix) worked to get familiar with the tasks. Despite this familiarization procedures, one subject, who was not familiar with a computer mouse or touch screen equipment, required support with handling the UI. This subject's adjustments were done by a trained hearing aid acoustician who had no bias (no knowledge about the analyzed preferences data of current or previous sessions) and was orally instructed by the participant to move the slider or cursor on the participant's behalf. Each session lasted approximately 30-40 minutes. The retest session took place about one week after the test session.

3.3. RESULTS

3.3.1. TEST-RETEST COMPARISON FOR PREFERENCE JUDGEMENTS

3.3.1.1. PROCESSING OF THE TARGET SPEECH SIGNAL ONLY (ADJS)

Correlations between test and retest adjustments for the three scenarios in which subjects could adjust the level of the frontal target speech in constant-noise conditions (as described above) showed that for all comparisons the linear correlations were positive and significant, ranging from R=0.49 for the TT co-located condition in the gain scenario up to R=0.98 in silence for the compression scenario. For each noise condition, the correlation was lowest for the gain scenario, and similar for the compression and clipping scenarios. Comparing the different noise conditions, it could be observed that correlations between test and retest results were higher in silence for each adjustment than in the presence of maskers. Within masker type (TT or MT), correlations for the separated setup were higher than for co-located maskers in every scenario. In summary, subjects were reasonably consistent in their individual preferences in test and retest in all of the measured conditions and gain adjustment methods. It was hence decided to conduct the subsequent analyses based on the individual data averaged between test and retest sessions for all three gain adjustment scenarios (gain, clipping and compression).

It was also determined how robust subjective choices were with respect to preferred frequency shaping in the EQ scenario. Only two out of the ten correlations were significant (for collocated MT maskers for y-presets, and in silence for x-presets) indicating that the subjects' choices regarding frequency shaping differed considerably between the two sessions. The spectral modifications resulting from the frequency shaping adjustments were also assessed by two measures reflecting the possible settings the subjects could choose. The center of gravity of the long-term spectrum was computed as a measure for low- vs. high-frequency balance. The mid-frequency amplification was measured as ratio of the average third-octave band power of the frequency range from 500 Hz to 4 kHz to the average power in the third-octave bands below and above this range. All of these measures confirmed limited reliability of subjective choices between test and retest sessions with not more than four of the ten correlation reaching significance. It was therefore decided not to proceed with further analyses of possible relations

between individual factors and the preference data obtained for frequency shaping in the sections that follow.

3.3.1.2. JOINT PROCESSING OF TARGET SPEECH AND MASKERS (ADJSN)

In analogy to the previous analyses, it was also tested how consistently subjects adjusted the signals in the adjSN part of the study, i.e., the part in which speech and maskers were always adjusted in the same manner and no SNR improvements could be chosen. In general, the correlations were higher than those obtained in the adjS part of the study (except for separated TT maskers in the compression scenario), ranging from 0.58 to 0.96. Across adjustment methods, correlations were more similar, i.e., the trend for correlations to be smaller in the gain scenario compared to compression and clipping was no longer as pronounced. In almost all cases correlations for co-located maskers were higher than for separated ones for a given masker type (with one exception for the gain scenario and the MT masker). As in the adjS analysis, due to the high test-retest reliability expressed in all conditions the data were averaged across test and retest for all three scenarios (gain, clipping and compression) for the subsequent analyses. The corresponding correlation analysis for the subjective choices in the EQ scenario was also performed. Only one computed correlation (in separated MT maskers for x-presets) was significant and no further analyses as to possible factors underlying subjective preferences with respect to frequency shaping were conducted in the following.

3.3.2. INTRODUCING SIGNAL MODIFICATIONS TO TARGET SPEECH ONLY

This part of the study investigated preferences for so-called “near-end” signal modifications, i.e., conditions in which the noise is not modifiable, but the target (speech) signal is. A comparable everyday scenario would be an announcement system in a train station.

3.3.2.1. PREFERRED LINEAR BROADBAND GAIN

Figure 3.1 shows the individually adjusted listening levels when subjects could adjust the speech level without introducing distortions in the presence of fixed TT masker (top) and MT masker (bottom). Gray and black bars represent data for spatially separated and co-located maskers, respectively. White bars (plotted in both panels for comparison) represent data for silence (no masker). Subjects are ordered (1-30) according to their SRTs in the spatially separated condition for the TT masker. The subjects' age and PTA are indicated between the two panels. For each subject the lowest possible playback level, which had been set to equal individual SRTs in each condition, is indicated by an asterisk at each bar. Thus, if the bars extended above the corresponding asterisks, this indicates that the subject preferred a level higher than that level (which was always the case). The right panels show the corresponding average data across all subjects.

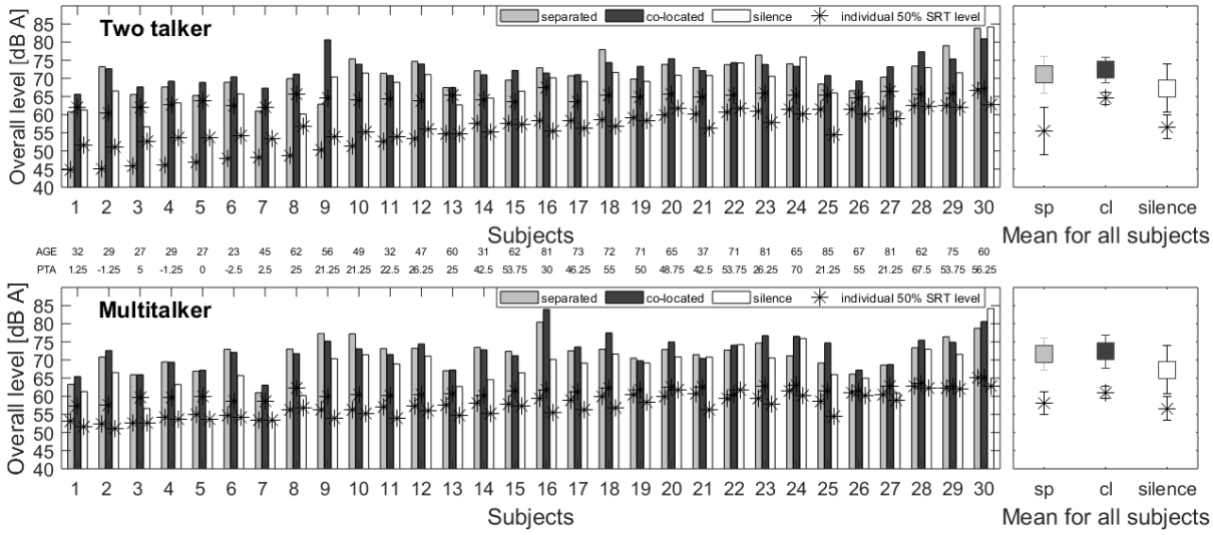

FIG. 3.1: INDIVIDUALLY PREFERRED SPEECH LEVELS (LEFT PANELS) FOR LINEAR GAIN ADJUSTMENTS IN THE PRESENCE OF TT MASKERS (TOP) AND MT MASKERS (BOTTOM). BARS REPRESENT DATA FOR CO-LOCATED MASKERS (BLACK), SPATIALLY SEPARATED MASKERS (GRAY), AND SILENCE (WHITE), RESPECTIVELY. THE LEVEL AT SRT FOR EACH CONDITION IS MARKED BY AN ASTERISK ON THE BAR SURFACE. MEAN RESULTS ACROSS ALL SUBJECTS ARE SHOWN IN THE RIGHT PANELS. ERROR BARS INDICATE STANDARD DEVIATIONS.

First, the adjusted speech levels were analyzed on a group level. The mean data for TT maskers (Fig. 3.1, upper-right panel) showed that despite a difference of 9 dB between mean SRTs for co-located (65 dB A) and spatially separated (56 dB A) maskers (see asterisks), the subjective

preference judgements led to a mean overall level difference of only 1 dB between these conditions (72 vs. 71 dB A). The same level difference between adjusted levels in spatial and co-located maskers was observed for the MT maskers (lower-right panel), where the mean SRT level difference was 3 dB (58 vs. 61 dB A). Considering the silence condition, the mean adjusted level was 67 dB A, being 4 and 5 dB lower than mean levels adjusted for TT maskers in separated and co-located conditions, respectively, and 5 dB lower than mean adjustments for MT maskers in both spatial set-ups. The effect of masker type and spatial constellation on adjusted speech levels was analyzed by means of a two-way repeated ANOVA measure, with a significance level of $\alpha = 0.05$. The degrees of freedom were Greenhouse-Geisser corrected. The two factors were masker type (TT or MT) and spatial constellation of sources (co-located or separated). The ANOVA showed a significant main effect of spatial constellation [$F(1, 29) = 7.409$, $p=0.011$]. The main effect of masker type was not significant [$F(1, 29) = 0.637$, $p=0.431$], neither was the interaction between both factors [$F(1, 29) = 0.349$, $p=0.559$].

Next, the individual data were inspected more closely since the subject group was highly inhomogeneous and large interindividual differences were observed for both masker types as well as in silence. For spatially separated TT maskers (top panel of Fig. 3.1, gray bars), the overall spread in preferred listening level was larger than 23 dB. For the subgroup of normal-hearing subjects (subjects #1-7), the spread was more than 12 dB, for the hearing-impaired subjects (#8-30) - it was about 21 dB. With SRT levels varying from 45 to 67 dB A (56 dB A on average), the minimum preferred level for this condition was just above 60 dB A (with a masker level being fixed at 65 dB A). The maximum output level chosen was 84 dB A and belonged to subject #30. The averaged overall level chosen by normal-hearing subjects in this condition was 66 dB A (ranging from 61 to 73 dB A), and 73 dB A for hearing-impaired listeners (from 63 to 84 dB). The overall mean level across all subjects was 71 dB A, i.e., subjects applied an average gain of 15 dB above SRT to frontal speech to meet their preferences, but this adjustment varied strongly from 5 to 28 dB. Normal-hearing subjects applied between 13 dB (#7) and 28 dB (#2) amplification (20 dB on average) to their 50% SRT level, while hearing-impaired subjects added between 5 dB (#26) and 24 dB (#10) – 14 dB on average. For the fixed co-located TT masker (Fig. 3.1, top panel, black bars) the overall spread of preferred speech levels varied from 66 (#1) to 81 dB A (#9), being on average 69 dB A for normal-hearing, and 73 dB A for hearing-impaired subjects. The SRT levels for this condition varied from 61 to 67 dB A, which was only a 6-dB spread in comparison to the 22-dB spread

in the spatially separated condition. Subjects applied on average 8 dB amplification to their SRT level in that condition (adjustments ranged from 2 dB for subject #13 to 16 dB for subject #9). For normal-hearing subjects this adjustment varied from over 4 to 12 dB (7 dB on average), and for hearing-impaired subjects from 2 to 16 dB (8 dB on average). Results for the spatially separated MT masker (Fig. 3.1, bottom panels, gray bars) spread in overall preferred level from 61 (#7) to 80 dB A (#16), with an average preferred adjustment of 72 dB A. Normal-hearing subjects preferred listening levels from 61 to 73 dB A (#7 and #6, mean of 67 dB), while hearing-impaired subjects adjusted levels to between 66 (#26) and 80 dB A (#16), with a mean preferred level of 73 dB A. The SRT level for this condition varied from 53 to 65 dB A across subjects and the amplification added to the target speech ranged from 5 (#26) to 22 dB (#9 and #10), with a mean of 14 dB. Both normal-hearing and hearing-impaired subjects preferred on average 14 dB added to their SRTs, with individual adjustments ranging from 7 to 18 dB and from 5 to 21 dB for each group, respectively. Both the co-located MT masker's minimum and maximum preference adjustments were very similar to the adjustments in the spatial case (ranging from 63 to 84 dB and belonging to the same subjects #7 and #16). SRT levels varied from 57 to 65 dB A, i.e. within a 5-dB smaller spread than in the separated condition, and a 3-dB higher average level across subjects. Participants decided to add on average 12 dB amplification to their SRT level (ranging from 4 (#7) to 22 dB (#16)). With a few exceptions, subjects adjusted the speech to lower levels in the silence condition (white bars) compared to conditions with a co-located masker, while the co-located condition adjustments were often similar to those adjusted for spatially separated maskers (mean difference 1.3 and 0.8 dB for the TT and MT masker, respectively). This also means that the average self-adjusted levels resulted in very similar SNRs, which were about +6 dB (TT spatial), +8 dB (TT co-located), +7 dB (MT spatial), and +8 dB (MT co-located).

3.3.2.2. PREFERRED PROCESSING IN SCENARIOS INVOLVING SIGNAL DEGRADATION

The purpose of the two other scenarios involving overall gain adjustments (compression and clipping scenarios) was to test how the acceptance of processing artifacts would affect preferred level adjustments of the participants. Specifically, the comparison to the level adjustments without artifacts aimed at investigating the personal trait of "noise hater" vs. "distortion hater"

(Völker et al., 2018). Results of 29 subjects are included in the following, because participant #5 quitted the study and did not take part in the compression scenario. Each scenario involving signal degradation (compression and clipping) was first analyzed separately to assess the effects of masker type and spatial constellations of sources on preferred speech levels. Two-way repeated-measures ANOVAs with a significance level of $\alpha = 0.05$ were conducted with factors masker type (TT or MT) and spatial constellation (co-located or separated). The degrees of freedom were Greenhouse-Geisser corrected. For both the clipping and the compression scenario, the ANOVA showed a significant main effect of spatial constellation [clipping: $F(1, 29) = 21.680$, $p<0.001$; compression: $F(1, 28) = 11.952$, $p=0.002$], a non-significant main effect of masker type [clipping: $F(1, 29) = 1.204$, $p=0.282$; compression: $F(1, 28) = 2.064$, $p=0.162$], and a significant interaction between both factors [clipping: $F(1, 29) = 7.646$, $p=0.01$; compression: $F(1, 28) = 20.563$, $p<0.001$], indicating that the effect of spatial separation depended on masker type.

The comparison of the preferred listening levels in conditions with signal degradations to preferred listening levels without distortions is illustrated as scatter plots in Fig. 3.2. Each panel illustrates the correlation between the speech levels adjusted in two different scenarios. Symbols correspond to individual data (symbol types are discussed below). The first two rows represent relations between the overall adjusted levels in the linear gain scenario (ordinates) and the clipping scenario (abscissae, first row) and compression scenario (abscissae, second row), respectively. The third row compares the level adjustments in the compression and clipping scenario. Linear correlations are indicated in the left top corner of each panel, together with the bias (in dB). The bias, illustrated as dashed line in each panel, was calculated as the y-intercept of a linear regression function with unity slope and can be interpreted as a measure of offset between the two variables in a scatter plot. For instance, a bias of -3.3 dB between adjustments in the clipping and gain scenarios for co-located TT maskers (top left panel) indicates that subjects adjusted the speech to a 3.3-dB lower level in the clipping scenario than in the gain scenario (on average).

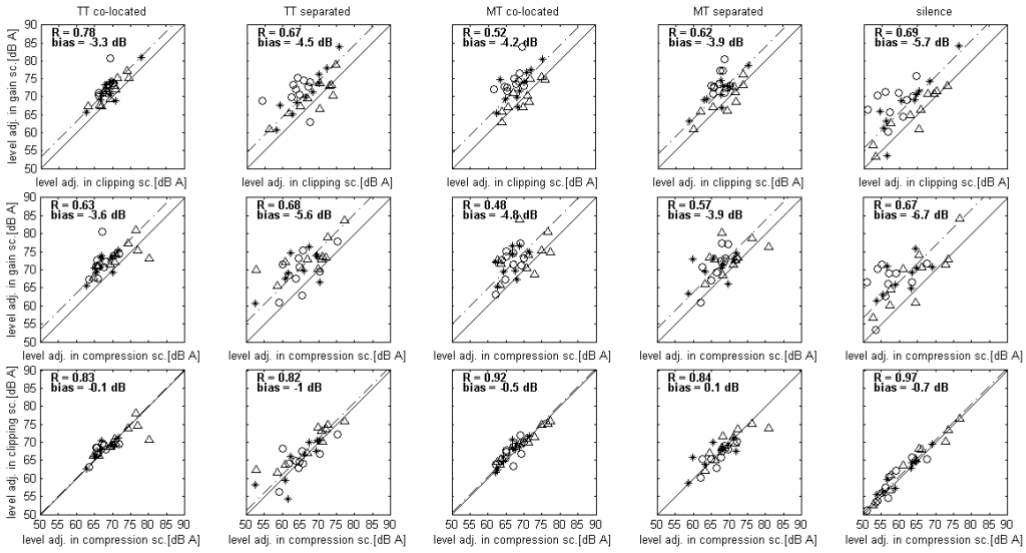

FIG. 3.2: CORRELATIONS BETWEEN ADJUSTED SPEECH LEVELS IN DIFFERENT MASKING CONDITIONS (COLUMNS). EACH ROW REPRESENTS CORRELATIONS BETWEEN THE SAME ADJUSTMENT METHODS. SYMBOLS REPRESENT INDIVIDUAL SUBJECTS (SEE TEXT).

In general, it could be observed that the gain adjustments in all three scenarios (linear, clipping, compression) were highly correlated. All fifteen correlations were significant with effect sizes ranging from R=0.48 to R=0.97. In terms of correlations as well as of bias, the comparisons between adjusted linear gains and both the adjustments involving distortions were very similar for each masker and spatial condition (compare first and second row of Fig. 3.2). The bias was always negative (ranging from -6.7 to -3.3 dB), indicating that a higher level was adjusted for linear gain without distortions. Biases for compression and clipping scenarios never differed by more than 1.1 dB for the same condition. Similarly, the difference in the correlations never exceeded 0.15 for the same condition. The third row of Figure 3.2 directly compares the adjusted levels chosen by subjects in the compression and clipping scenarios. This relation produced the highest correlations of all comparisons (ranging from R=0.82 to R=0.97) and the magnitude of the bias never exceeded 1 dB, indicating that very similar levels were adjusted in both scenarios for all masker and spatial conditions.

To investigate if a particular individual pattern could be observed in the preferred level adjustments, the mean absolute distance to the diagonal was computed for each subject as a measure of relation between scenarios involving linear gain only and gain combined with distortions. The underlying reasoning was that subjects whose preferred settings were further away from (i.e., above) the diagonal (Fig. 3.2) preferred to increase the level of target speech

only if this would not involve distortions (tending to be "distortion haters"). In contrast, subjects whose choices were located close to diagonal would choose to apply gain to the target speech even when this introduced distortions (possibly being classified as "noise haters"). Note that, in principle, it would have been possible for subjects to increase the speech level more in conditions involving distortions than without distortions (i.e., to be below the diagonal). However, inspection of Fig. 3.2 suggests that this was rarely the case and was never consistently made by any particular subjects. As a visual guide to segregating potential "noise haters" vs. "distortion haters", the ten subjects with the largest mean absolute distance to the diagonal (calculated for all panels of the two top rows) are represented by circles in Fig. 3.2, while the ten subjects with the smallest mean absolute distance are represented by triangles. The remaining nine subjects with intermediate distances to the diagonal are shown as asterisks. Qualitatively, the representations in Fig. 3.2 indicate some stable patterns throughout the cases in that it could be observed for every plot in the first two rows that circles appeared rather consistently above the diagonal, while triangles were, in general, closer to the diagonal. In order to further explore this finding, Table 3.1 was constructed representing the individual absolute distances to the diagonal for all subjects and conditions. In the table, subjects are sorted by the average absolute distance across all conditions (second column). The ten subjects with the highest average distance are marked as light gray, while the ten subjects with the lowest average distance are marked as dark gray. The nine remaining subjects are marked as white. This coding in gray scales is the same for each column, i.e., the top and bottom ten subjects of each color are always coded by light and dark gray, respectively.

TAB. 3.1: ABSOLUTE DISTANCES TO THE DIAGONAL IN THE SCATTER PLOTS OF FIG. 3.2, COMPUTED BASED ON EXPERIMENTAL DATA OBTAINED IN THE ADJS PART OF THE STUDY. THE TEN HIGHEST AND LOWEST VALUES OF EACH COLUMN ARE MARKED AS LIGHT AND DARK GRAY, RESPECTIVELY. SUBJECTS ARE SORTED ACCORDING TO THE OVERALL MEAN DISTANCE (SECOND COLUMN).

Subject number	Mean absolute distance to diagonal	Gain vs. compression scenarios					Gain vs. clipping scenarios				
		TT cl	TT sp	MT cl	MT sp	Silence	TT cl	TT sp	MT cl	MT sp	Silence
9	9.9	11.1	4.7	7.9	9.8	16.5	13.5	2.6	10.1	9.5	13.1
2	9.1	4.2	10.3	7.3	5.5	15.4	7.0	8.8	8.7	8.2	15.4
6	8.8	4.4	14.5	10.5	7.1	8.5	5.4	7.4	10.0	13.2	6.8
10	8.4	3.8	11.3	3.0	9.2	15.5	3.6	9.6	3.0	7.6	17.6
16	8.4	2.7	5.8	14.4	11.8	6.4	2.2	6.0	15.2	12.4	6.7
12	8.1	3.5	8.7	5.8	3.2	11.3	6.9	12.5	7.6	7.7	13.8
8	7.5	4.9	7.5	6.5	8.0	3.3	5.1	17.3	8.1	7.1	6.8
14	7.4	3.6	8.3	8.1	6.4	3.5	5.3	11.3	10.2	9.0	8.3
24	6.2	4.3	6.4	7.8	2.4	11.3	3.7	4.6	7.6	2.9	11.2
17	6.1	2.5	6.1	5.7	6.8	7.1	5.6	5.9	8.4	5.0	8.0
25	6.1	3.9	4.4	11.3	6.3	11.5	4.8	6.6	7.6	2.6	1.5
23	6.0	4.8	6.4	6.0	7.1	5.8	3.2	8.9	9.9	2.7	5.1
4	5.8	2.5	8.3	4.7	5.7	7.0	3.7	6.8	5.0	6.1	7.9
11	5.8	1.5	3.1	2.8	4.9	8.5	3.2	11.4	4.4	4.2	13.7
30	5.6	2.9	8.0	5.4	3.5	7.7	4.5	6.5	4.0	2.6	11.4
18	5.2	5.0	5.6	5.4	2.7	6.2	2.6	2.6	8.3	0.9	12.5
1	4.5	3.0	2.6	3.1	4.6	5.8	2.7	8.2	2.9	4.6	7.6
22	3.9	4.6	3.6	4.2	3.1	5.9	3.6	2.6	2.9	1.1	8.0
19	3.5	5.3	3.4	1.7	2.8	4.9	6.4	4.0	3.6	1.7	7.3
3	3.5	1.3	3.9	2.2	3.9	4.1	2.9	7.0	2.7	2.8	4.1
15	3.5	4.2	2.6	4.3	3.4	0.6	4.0	0.8	2.0	3.9	8.9
20	3.4	4.1	3.5	3.2	1.6	1.3	3.9	4.2	3.6	1.7	7.3
27	3.1	2.5	3.7	2.7	3.2	4.6	6.9	0.5	4.2	0.3	2.4
26	2.8	0.0	3.7	2.4	3.2	2.1	0.9	3.6	0.9	3.4	8.0
13	2.8	0.6	2.6	1.7	1.7	5.2	1.6	3.8	2.4	2.1	6.2
29	2.7	0.8	4.2	1.0	2.4	1.3	1.7	6.4	2.4	4.5	2.0
7	2.4	4.2	4.7	0.7	0.9	0.0	3.8	2.0	0.8	1.0	5.5
28	1.9	3.3	0.2	0.6	0.4	0.4	3.0	1.3	0.5	1.1	8.4
21	1.1	1.2	0.1	0.5	0.1	2.9	1.4	2.0	0.4	0.4	1.7

Looking at the color pattern in Tab. 3.1 alone, it is noticeable that subjects were in general consistent with their choices across conditions, although a few exceptions could be observed. One group of subjects (light gray) tended to apply gain to the target speech only if the resulting

change in SNR did not come at the cost of distortions (possible "distortion haters") while another group (dark gray) applied relatively similar gains to the target speech no matter whether distortion were present or not (possible "noise haters"). To statistically explore the behavior of the subjects, a three-way ANOVA was conducted with the two within-subjects factors: scenarios (two levels: gain vs. clipping and gain vs. compression) and listening condition (five levels: silence and all four combinations of masker type and spatial constellation), and one between-subjects factor being the group based on the overall mean distance as coded by the gray scales in the second column in Table 3.2. The ANOVA confirmed significant effects of both within-subject factors [scenario: $F(1,26)=8.849$, $p=0.006$; listening condition: $F(2.874,74.736)=9.524$, $p<0.001$] and a significant effect of the between-subject factor group [$F(2,26)=76.045$, $p<0.001$]. None of the interactions were significant. As post-hoc analyses separate one-way ANOVAs with between-subject factor group were conducted for each combination of scenario and listening conditions (i.e., each of the third to last column of Table 3.1). This revealed significant differences between groups in every case ($p<0.05$ for two of the ten cases, and $p<0.005$ for the other seven tests). Post-hoc paired comparisons showed that this was always due to the significant differences between groups one and three, while group two did not differ from either group. Altogether, these analyses indicate that grouping based on the overall mean distance produced consistent differences in all conditions and scenarios.

3.3.2.3. RELATIONS BETWEEN LISTENING PREFERENCES AND PERSONAL FACTORS

In this section the individual factors PTA, speech recognition performance in simple speech intelligibility tests (DTT and Goesa), and age were correlated with the individually preferred level adjustments. The goal was to explore if and to what degree the observed variance in listening preferences could be explained by those basic individual factors. Squared correlations (coefficients of determination) were computed as a measure of how much of the observed variance in adjusted levels could be explained by the individual factors. The data are presented in Tab. 3.2. For the gain scenario (top rows) the correlations were significant in the majority of cases (seventeen out of twenty), while the effect sizes of the significant correlations were relatively low in some case ($R^2=0.13$). Largest effect sizes for each personal factor occurred for

silence (up to R^2=0.49). Correlations between age and adjustments were in general weaker, with the highest significant one for the co-located MT condition (R^2 = 0.23), yet the only correlation that was not significant occurred for the spatially separated TT masker. Squared correlations between PTA and preferences in the gain scenario were slightly higher than for age. All were significant, ranging up to R^2=0.46 in silence. Correlations of personal preferences with speech intelligibility test results (DTT and Goesa) were very similar for a given condition and were again highest for silence.

TAB. 3.2: COEFFICIENTS OF DETERMINATION BETWEEN INDIVIDUAL FACTORS (AGE, PTA, DTT AND GOESA), AND PREFERENCE JUDGEMENTS (OVERALL LEVEL CHOSEN, AVERAGED ACROSS TEST AND RETEST SESSIONS) FOR THE THREE ADJUSTMENT SCENARIOS (GAIN, CLIPPING AND COMPRESSION) FOR ALL SPATIAL (SP) AND CO-LOCATED (CL) TWO-TALKER (TT) AND MULTI-TALKER (MT) MASKERS. BOLD VALUES INDICATE SIGNIFICANT CORRELATIONS (P<0.05).

	gain scenario				
	MT cl	MT sp	TT cl	TT sp	silence
Age	**0.23**	**0.13**	**0.15**	0.13	**0.16**
PTA	**0.23**	**0.14**	**0.29**	**0.33**	**0.46**
DTT	**0.23**	0.09	**0.36**	**0.35**	**0.49**
Goesa	**0.23**	0.10	**0.29**	**0.34**	**0.49**
	compression scenario				
Age	**0.28**	**0.35**	**0.28**	**0.26**	**0.27**
PTA	**0.46**	**0.54**	**0.31**	**0.53**	**0.62**
DTT	**0.50**	**0.44**	**0.34**	**0.58**	**0.66**
Goesa	**0.57**	**0.49**	**0.38**	**0.56**	**0.73**
	clipping scenario				
Age	**0.25**	**0.15**	0.09	**0.34**	**0.24**
PTA	**0.47**	**0.48**	**0.32**	**0.55**	**0.67**
DTT	**0.49**	**0.42**	**0.36**	**0.47**	**0.65**
Goesa	**0.49**	**0.47**	**0.38**	**0.49**	**0.72**

Regarding correlations in the compression scenario (mid-rows), all were significant, with PTA, DTT and Goesa holding similarly high predictive power of preferred levels in all conditions. All effect sizes were larger than for linear gain (on average R^2 was larger by 0.2) and, again, the largest effect sizes were found for correlations between individual factors and level adjustments in the silence condition. Correlations obtained for the clipping scenario (bottom rows) were similar to the compression scenario, i.e., levels adjusted in the presence of maskers

could also be explained by individual factors to a larger degree in the compression scenario than in the gain scenario, with generally smaller values for age than for the other individual factors.

3.3.2.4. RELATIONS BETWEEN LISTENING PREFERENCES AND SRTS IN THE SAME CONDITIONS

Squared correlations were also computed between listening preferences and individual SRTs measured in the same listening scenarios (Tab. 3.3). In addition, listening preferences were also correlated with SRT-difference measures, i.e., SRM for each masker type, and the MTI for each spatial constellation. As already described, SRM is defined as the difference in SRT between co-located and spatial listening conditions for the same masker type, describing the speech intelligibility benefit from the spatial separation between target and maker sources. MTI is computed as the difference in SRT for multi-talker and two-talker maskers for a given spatial constellation and describes the potential benefit from moving from a more stationary speech-shaped masker to a two-talker masker that enables better use of spectro-temporal gaps in the signal that may improve speech intelligibility of the target source (but at the same time may introduce more informational masking because the TT maskers were intelligible).

Correlations between SRT measures and preference judgements for the gain scenario (Tab. 3.3, top rows) were in the great majority significant. Exceptions were observed for MTI (co-located), which was not significantly related to adjusted levels in any condition of the present study. Dependencies between SRT measures and preferences in the different conditions were stronger for scenarios introducing distortions to the target speech due to compression or clipping (see Tab. 3.3, mid- and bottom rows) than it was the case in the gain scenario (on average R^2 was 0.23 larger for compression and 0.21 larger for clipping than for linear gain adjustments). In general, correlations with SRTs in spatially separated conditions were slightly higher than in co-located conditions. The strongest dependence occurred for relations between preferences in silence and SRTs measured in spatially separated MT maskers. SRM showed similar correlations with preferences for both masker types, being highest for silence. Correlations with MTI were only significant for separated conditions and showed rather small predictive power in comparison to other variables tested. The fields marked in gray (Tab. 3.3) indicate correlations between SRT results and preference judgments in same spatial and masker

conditions. In many cases coefficients of determination were not largest in these cases, i.e., when the SRT scenario matched the preference scenario, but rather similar for the different masking conditions.

TAB. 3.3: SQUARED CORRELATIONS BETWEEN SRT PERFORMANCE AND PREFERENCE JUDGMENTS FOR ALL EXPERIMENT CONDITIONS IN THREE SCENARIOS (GAIN, CLIPPING AND COMPRESSION) FOR ALL SPATIAL (SP) AND CO-LOCATED (CL) TWO-TALKER (TT) AND MULTI-TALKER (MT) MASKERS. SRM AND MTI INDICATE SRT DIFFERENCES (SEE TEXT). CORRELATIONS BETWEEN SAME LISTENING CONDITIONS WERE MARKED IN GRAY. SIGNIFICANT CORRELATION (P<0.05) ARE BOLDED.

	gain scenario				
	MT cl	MT sp	TT cl	TT sp	silence
SRT TT sp	**0.29**	**0.18**	**0.28**	**0.40**	**0.43**
SRT TT cl	**0.37**	**0.25**	**0.27**	**0.30**	**0.28**
SRT MT sp	**0.33**	**0.24**	**0.39**	**0.46**	**0.51**
SRT MT cl	**0.31**	**0.13**	**0.36**	**0.45**	**0.35**
SRM TT	**0.22**	**0.13**	**0.24**	**0.36**	**0.42**
SRM MT	**0.24**	**0.28**	**0.29**	**0.32**	**0.51**
MTI sp	**0.22**	0.11	**0.17**	**0.29**	**0.31**
MTI cl	0.00	0.04	0.05	0.10	0.04
	compression scenario				
SRT TT sp	**0.60**	**0.61**	**0.53**	**0.62**	**0.71**
SRT TT cl	**0.48**	**0.51**	**0.50**	**0.32**	**0.50**
SRT MT sp	**0.69**	**0.69**	**0.60**	**0.63**	**0.81**
SRT MT cl	**0.60**	**0.64**	**0.57**	**0.48**	**0.71**
SRM TT	**0.54**	**0.54**	**0.45**	**0.63**	**0.66**
SRM MT	**0.55**	**0.50**	**0.43**	**0.57**	**0.64**
MTI sp	**0.44**	**0.46**	**0.39**	**0.52**	**0.51**
MTI cl	0.06	0.06	0.04	0.10	0.13
	clipping scenario				
SRT TT sp	**0.52**	**0.53**	**0.41**	**0.67**	**0.71**
SRT TT cl	**0.47**	**0.45**	**0.36**	**0.54**	**0.49**
SRT MT sp	**0.67**	**0.68**	**0.57**	**0.74**	**0.81**
SRT MT cl	**0.61**	**0.58**	**0.52**	**0.66**	**0.66**
SRM TT	**0.45**	**0.47**	**0.36**	**0.60**	**0.67**
SRM MT	**0.52**	**0.56**	**0.43**	**0.57**	**0.69**
MTI sp	**0.34**	**0.34**	**0.24**	**0.52**	**0.53**
MTI cl	0.07	0.07	0.10	0.06	0.10

3.3.3. INTRODUCING EQUAL CHANGES TO SPEECH AND NOISE SIGNALS

This part of the study investigated the "far-end" signal modification approach, i.e., conditions in which noise and target speech were mixed and could only be modified simultaneously and undergoing the same processing. A comparable everyday scenario would be an amplifying device playing back noisy speech that was captured by a microphone (and not performing noise reduction). Subjects were presented with the same user interfaces, signals and tasks, i.e., the only difference to the adjS part of this study was that every signal modification resulted in processing not only the speech signal but speech and noise simultaneously. Remember that the SNR was fixed to the individual SRT plus 9 dB (see section 3.2.3) and no SNR modifications could be made here. Due to the fact that two subjects dropped out from the study (#5 and #19) after (partly) finishing the adjS part, the following data are results of 28 participants.

3.3.3.1. PREFERRED LINEAR BROADBAND GAIN

The individually adjusted listening levels without simultaneously introducing signal distortions for TT and MT maskers are presented in the top and bottom left panels of Fig. 3.3, respectively, complemented with mean values across all subjects plotted in the right panels. Results obtained in silence in the previous part of the study (adjS) were re-plotted as white bars for comparison. Asterisks represent the minimum adjustable level (individual SRTs + 9 dB). As in the adjS part, adjusted levels varied strongly across subjects, particularly in spatial maskers. Looking at averaged results across all subjects for TT maskers (Fig. 3.3, upper right panel) it could be observed that the difference of 9 dB between mean starting presentation levels for co-located (73 dB A) and spatially separated (64 dB A) conditions was approximately preserved after subjective preference judgements and led to the mean overall level difference between co-located and spatial conditions of about 5 dB. For co-located MT maskers the mean preferred overall level was 6 dB higher (74 dB A) than the mean preference for spatially separated maskers (on average 70 dB A; 61dB A for normal-hearing and 72 dB A for hearing-impaired subjects). Averaged results across all subjects for the MT maskers (Fig. 3.3, lower-right panel) showed that the 3-dB difference in starting level between co-located (70 dB A) and spatially separated (67 dB A) conditions was preserved in mean preferred overall level adjustments.

These trends observed in the mean data across subjects were supported by a two-way repeated measures ANOVA, which showed that the effect of spatial constellation was significant [$F(1, 27) = 53.744$, $p<0.00$], while the effect of masker type was not [$F(1, 27) = 0.280$, $p=0.601$]. The interaction between both factors was not significant [$F(1, 27) = 0.849$, $p=0.365$].

As before, the individual data were inspected more closely due to the large interindividual differences observed for both masker types. For the spatially separated TT maskers the overall spread in preferred listening levels reached 25 dB (similar to the 23 dB in the adjS part). The spread was about 7 dB and 23 dB for normal-hearing and hearing-impaired subjects, respectively (in comparison to 12 dB and 21 dB in the adjS part). The overall mean level across all subjects was 70 dB A (1 dB less than in the adjS case), i.e., subjects decided to apply a mean gain of 6 dB to their lowest possible presentation level, while in the adjS part of the study this mean adjustment was 15 dB when the noise level was fixed.

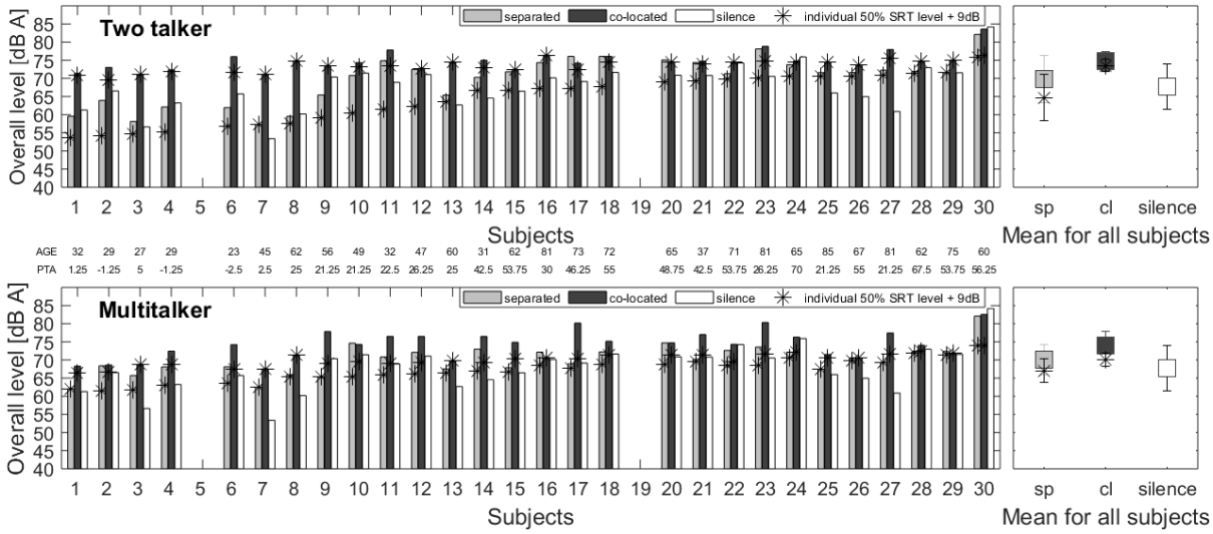

FIG. 3.3: INDIVIDUAL PREFERENCES FOR GAIN ADJUSTMENTS APPLIED SIMULTANEOUSLY TO SPEECH AND MASKERS FOR TT (TOP LEFT PANEL) OR MT MASKERS (BOTTOM LEFT PANEL) IN CO-LOCATED (BLACK) AND SPATIALLY SEPARATED CONDITIONS (GRAY) TOGETHER WITH MEAN RESULTS FOR ALL SUBJECTS (RIGHT PANELS). ASTERISK REPRESENTS THE MINIMUM ADJUSTABLE LEVEL (WHICH WAS SET TO SRT+9DB). RESULTS MEASURED IN SILENCE DURING THE PREVIOUS PART OF THE STUDY (ADJS) ARE RE-PLOTTED FROM FIG. 3.1. SUBJECTS #5 AND #19 QUIT THE STUDY BEFORE FINISHING THIS PART OF THE EXPERIMENT.

For co-located TT maskers (Fig. 3.3, upper panel), the overall preferred level varied from 71 (#1) to 84 dB A (#9) with an average level of 72 dB A for normal-hearing and 75 dB A for hearing-impaired listeners. The applied gain was on average slightly more than 1 dB but

reached up to almost 9 dB for some individuals (#30). In general, the mean spread of subjective choices was half of the spread in the previous condition (spatial TT) and the mean gain applied was smaller (1 vs. 6 dB). Individual results for spatially separated MT maskers (Fig. 3.3, lower panels) spread in overall preferred level by 20 dB, with an average adjustment to 70 dB A (similar to the 72 dB A in the adjS part). Both normal-hearing and hearing-impaired subjects preferred on average about 3 dB gain applied to the minimum presentation level (in comparison to the mean of 14 dB for both groups in the adjS part). For co-located MT maskers' participants preferred to add on average about 4 dB gain to their minimum presentation levels (compared to 11 dB in the adjS part) and the interindividual spread of preferred presentation level reached 15 dB.

3.3.3.2. PREFERRED PROCESSING IN SCENARIOS INVOLVING SIGNAL DEGRADATION

This section considers the scenarios in which both target speech and maskers underwent amplification combined with broadband compression or peak-clipping limiting. As before, two-way repeated-measures ANOVAs were conducted to assess the impact of masker type and spatial constellation of sources on adjusted levels. Results for both compression and clipping showed a significant effect of a spatial constellation [compression: $F(1, 27) = 58.694$, $p<0.001$; clipping: $F(1, 27) = 66.335$, $p<0.001$]. For the clipping scenario the impact of a masker type was significant [$F(1, 27) = 6.293$, $p=0.018$] and the interaction between both factors was significant as well [$F(1, 27) = 40.741$, $p<0.001$], while in the compression scenario the impact of the masker type was not significant [$F(1, 27) = 2.609$, $p=0.118$], although the interaction between both factors was [$F(1, 27) = 46.934$, $p<0.001$].

The preferences of individual adjustments regarding the nonlinear processing schemes were compared to the distortionless level adjustment in Fig. 3.4 in the same way as for the adjS part (Fig. 3.2). In general, it could be observed that almost all correlations were significant. In most cases, correlations were higher than in the adjS part of the study for same conditions. The resulting biases were always negative for correlations between preferred overall gain adjustments and compression and clipping scenarios (first and second row in Fig. 3.4), although they were only -1.9 dB on average, i.e., smaller than in the adjS part. The bias between compression and clipping (third row) was again very small.

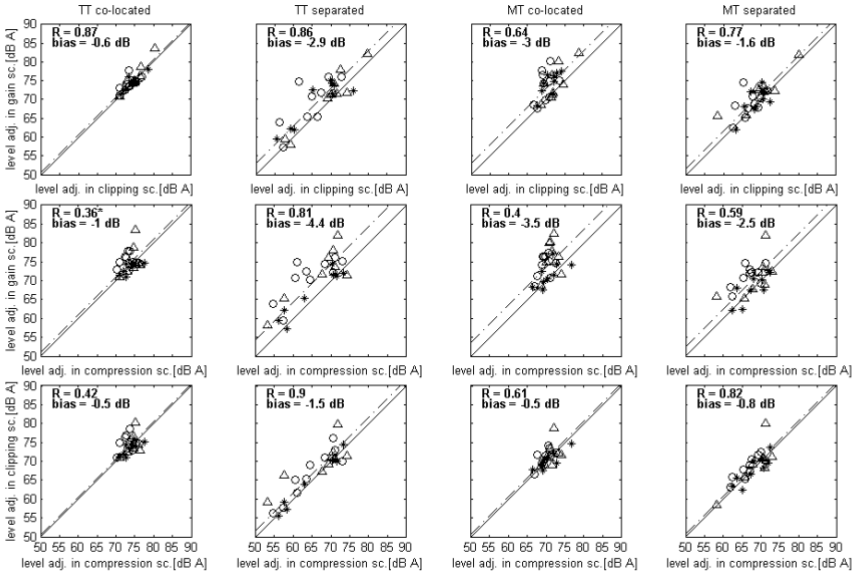

FIG. 3.4: SAME DATA REPRESENTATION AS IN FIG. 3.2, BUT FOR CORRELATIONS BETWEEN PREFERRED LEVEL ADJUSTMENTS APPLIED SIMULTANEOUSLY TO TARGET SPEECH AND MASKERS IN DIFFERENT SCENARIOS. CORRELATIONS THAT WERE NOT SIGNIFICANT ($\alpha = 0.05$) ARE MARKED WITH AN ASTERISK.

The subject coding as circles, triangles, and asterisks is the same in Fig. 3.4 as in Fig. 3.2, i.e., subjects that tolerated artifacts for the benefit of a higher SNR (triangles) in the adjS part, and subjects that preferred a lower SNR for the benefit of fewer artifacts (circles) are again marked by the same symbols. In this part of the study, however, the differences between linear gain and gain accompanied by distortions cannot be interpreted as indication for "noise haters" or "distortion haters", because the SNR did not vary and, hence, there was no trade-off between good SNR and few distortions. Rather, data points above the diagonals in Fig. 3.4 indicate that subjects would have preferred to listen at a higher overall level (as indicated by the linear gain adjustments) but did not do so because this would have introduced distortions. However, as indicated above the bias values were in general rather small, indicating that this effect was not very pronounced. Note again that, in principle, subjects could have chosen a higher playback level in the distortion scenarios than in the linear-gain scenario (i.e., being below the diagonal), but this was rarely the case and did not occur systematically for any subject.

Inspecting the different symbols in Fig. 3.4 suggests that there may again be a trend for subjects previously labeled as "distortion hates" (circles) to be further away from the diagonal

than the “noise haters” (triangles). To further evaluate this observation Tab. 3.4 was constructed similar to the Tab. 3.1, i.e., subjects were ordered according to subjective traits elicited in the adjS part of the study (“distortion haters”: light gray marking, “noise haters”: dark gray marking, intermediates: white marking). Visual inspection suggests that the color pattern was rather preserved, especially in the mean absolute distance to diagonal calculated across all masking conditions (2nd column of Tab. 3.4). To further explore this, separate one-way ANOVAs with independent variables mean absolute distance to the diagonal and between-subject factor group were conducted for each of the eight conditions. The groups were based on the classification of the adjS part, i.e., the first ten subjects of Tab. 3.1 (#9-17) were assigned to group 1 and the last ten subjects were assigned to group 3 (#3-21). Group 2 contained the remaining 8 subjects (#25-22). This grouping showed that the mean values of group 1 were always larger than those of group 3, but the differences were only significant in three of the eight cases, indicating that the effects here were trends rather than robust effects.

TAB. 3.4: MEAN ABSOLUTE DISTANCE TO DIAGONAL BEING COMPUTED BASED ON EXPERIMENTAL DATA OBTAINED IN EVERY LISTENING CONDITION TESTED IN THE ADJSN PART OF THE STUDY. THE TEN HIGHEST AND LOWEST VALUES OF EACH COLUMN ARE MARKED AS LIGHT AND DARK GRAY, RESPECTIVELY. SUBJECTS ARE SORTED ACCORDING TO THE OVERALL MEAN DISTANCE TO THE DIAGONAL IN THE ADJS PART OF THE STUDY.

Subject number	Mean distance to diagonal	Gain vs. compression scenarios				Gain vs. clipping scenarios			
		TT cl	TT sp	MT cl	MT sp	TT cl	TT sp	MT cl	MT sp
9	3.1	0.0	4.9	8.8	0.4	0.0	0.9	8.8	0.6
2	4.7	2.8	9.4	1.6	6.6	2.0	7.8	1.9	5.4
6	3.1	3.3	3.7	5.1	3.1	2.5	1.6	4.9	0.6
10	5.6	1.2	9.7	4.8	9.2	0.0	5.8	4.8	9.2
16	1.5	0.0	4.9	0.3	1.3	0.0	3.5	0.6	1.7
12	4.9	0.0	9.4	7.0	4.7	0.0	7.2	6.2	4.8
8	0.6	0.2	2.0	0.3	0.4	0.0	1.8	0.4	0.0
14	3.5	2.0	2.5	6.4	5.2	2.0	1.1	5.9	4.0
24	2.2	0.4	3.1	4.2	1.4	0.4	2.9	4.1	1.1
17	4.2	1.4	5.1	9.2	0.3	1.2	6.7	9.1	0.8
25	0.5	0.0	0.4	1.0	0.5	0.0	1.1	0.3	0.7
23	4.7	2.9	7.0	7.1	3.1	2.3	5.6	7.0	2.9
4	2.8	0.5	4.5	3.6	4.6	0.5	2.9	3.4	2.6
11	6.9	4.5	11.4	7.4	3.7	4.5	13.3	7.4	3.0
30	2.9	3.7	4.2	2.3	1.9	3.3	2.4	3.8	2.0
18	1.6	0.0	3.0	2.5	1.1	0.8	3.2	1.7	0.5
1	1.5	0.4	3.5	1.8	0.4	0.3	4.1	0.6	1.2
22	1.9	0.2	0.7	3.3	2.1	0.0	1.2	4.8	3.0
3	2.6	0.0	5.0	0.0	7.6	0.0	1.0	0.0	7.3
15	2.2	0.0	3.4	3.7	0.3	0.0	4.5	3.9	1.4
20	3.0	0.0	4.8	3.0	4.3	0.0	5.2	2.5	4.3
27	2.0	0.3	1.2	0.8	3.2	0.6	3.8	3.3	3.0
26	0.4	0.0	1.0	0.0	0.6	0.0	1.0	0.3	0.6
13	0.5	0.0	1.0	0.0	0.5	0.0	1.6	0.0	1.1
29	0.4	0.2	0.3	0.4	1.1	0.0	0.0	0.0	1.1
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
28	1.0	0.4	2.3	0.1	0.6	0.4	2.5	0.6	1.4
21	2.4	0.7	2.7	4.3	1.9	1.2	4.1	4.2	0.4

3.3.3.3. RELATIONS BETWEEN LISTENING PREFERENCES, PERSONAL FACTORS AND SPEECH INTELLIGIBILITY DATA

The same individual factors considered in the adjS part were correlated with subjective preference judgements (Tab. 3.5). In general, these dependencies were comparable to those obtained in the adjS part, especially in scenarios introducing distortions with the highest correlations (like in adjS) between personal factors and preferences in separated TT maskers. In particular, it was observed that age generally correlated worse with individual preferences than PTA, DTT, or Goesa.

TAB. 3.5: SQUARED CORRELATION BETWEEN INDIVIDUAL FACTORS (AGE, PTA, DTT AND GOESA), AND PREFERENCE JUDGEMENTS (OVERALL LEVEL CHOSEN AVERAGED ACROSS TEST AND RETEST SESSIONS) FOR THE THREE EXPERIMENT SCENARIOS (GAIN, CLIPPING AND COMPRESSION) FOR ALL SPATIAL AND MASKER CONDITIONS. VALUES FOR SIGNIFICANT CORRELATIONS (P<0.05) ARE BOLDED.

	Gain scenario			
	MT cl	MT sp	TT cl	TT sp
Age	0.06	0.09	0.12	**0.31**
PTA	**0.20**	**0.37**	0.11	**0.50**
DTT	**0.20**	**0.38**	**0.15**	**0.42**
Goesa	**0.23**	**0.42**	**0.21**	**0.52**
	Compression scenario			
Age	**0.29**	**0.37**	**0.45**	**0.51**
PTA	**0.36**	**0.58**	**0.30**	**0.68**
DTT	**0.45**	**0.57**	**0.34**	**0.61**
Goesa	**0.56**	**0.67**	**0.44**	**0.69**
	Clipping scenario			
Age	**0.29**	**0.28**	**0.36**	**0.48**
PTA	**0.43**	**0.52**	**0.23**	**0.59**
DTT	**0.46**	**0.54**	**0.29**	**0.58**
Goesa	**0.58**	**0.63**	**0.38**	**0.63**

Correlations between SRT measures and listening preferences are presented in Tab. 3.6. In general, dependencies were stronger in the adjSN part than in the adjS part (on average R^2 values were 0.13 higher), especially for TT and MT maskers in spatial conditions. Correlations for co-located conditions were roughly comparable to the adjS part. As for adjS, correlations between preferences and MTI in co-located conditions were never significant. Highest correlations generally occurred between SRTs and listening preferences in same conditions (gray fields in Tab. 3.5), which was not the case for the adjS part (Tab. 3.3), where correlations were rather similar in all conditions. This is likely an artefact of the fact that many subjects performed only small adjustments from their starting presentation point of SRT + 9 dB, which was naturally highly correlated to the SRT.

TAB. 3.6: SQUARED CORRELATIONS BETWEEN SRT PERFORMANCE AND PREFERENCE JUDGMENTS FOR ALL EXPERIMENTAL CONDITIONS IN THE THREE ADJUSTMENT SCENARIOS (GAIN, CLIPPING AND COMPRESSION) FOR THE ADJSN PART. VALUES FOR SIGNIFICANT CORRELATIONS (P<0.05) ARE BOLDED.

	Gain scenario			
	MT cl	MT sp	TT cl	TT sp
SRT TT sp	**0.30**	**0.47**	**0.35**	**0.72**
SRT TT cl	**0.19**	**0.36**	**0.46**	**0.45**
SRT MT sp	**0.33**	**0.54**	**0.42**	**0.68**
SRT MT cl	**0.32**	**0.46**	**0.42**	**0.53**
SRM TT	**0.28**	**0.43**	**0.27**	**0.69**
SRM MT	**0.23**	**0.44**	**0.29**	**0.60**
MTI sp	**0.23**	**0.35**	**0.25**	**0.64**
MTI cl	0.08	0.05	0.00	0.04
	Compression scenario			
SRT TT sp	**0.63**	**0.83**	**0.60**	**0.96**
SRT TT cl	**0.59**	**0.68**	**0.85**	**0.62**
SRT MT sp	**0.72**	**0.92**	**0.70**	**0.90**
SRT MT cl	**0.81**	**0.74**	**0.73**	**0.71**
SRM TT	**0.54**	**0.74**	**0.44**	**0.92**
SRM MT	**0.41**	**0.79**	**0.44**	**0.79**
MTI sp	**0.46**	**0.64**	**0.44**	**0.87**
MTI cl	0.12	0.03	0.00	0.04
	Clipping scenario			
SRT TT sp	**0.66**	**0.74**	**0.55**	**0.91**
SRT TT cl	**0.60**	**0.59**	**0.76**	**0.64**
SRT MT sp	**0.75**	**0.84**	**0.63**	**0.86**
SRT MT cl	**0.85**	**0.67**	**0.68**	**0.73**
SRM TT	**0.57**	**0.67**	**0.40**	**0.85**
SRM MT	**0.43**	**0.73**	**0.39**	**0.71**
MTI sp	**0.49**	**0.55**	**0.40**	**0.80**
MTI cl	**0.14**	0.04	0.00	0.04

3.4. DISCUSSION

3.4.1. TEST – RETEST RELIABILITY

Profiling listeners based on their preferred processing schemes is possible only if such preferences are stable over time for the same or similar acoustic conditions. The current data from the adjS part of the study showed high test-retest reliability of subjective choices for gain, clipping, and compression scenarios under all acoustic conditions for the widely diversified group of participants in regards of their hearing abilities and age. It was observed that the gain scenario scored slightly lower test-retest correlations than the two scenarios involving distortions, which was often caused by a single or couple of outliers. One possible reason for the small decrease in test/re-test stability could be that the linear-gain scenario may have encouraged participants to explore the perceptual space accessible via UI over a broader range due to the lack of signal degradations that could follow. It can be assumed that listeners adjusted the SNR until their listening effort was sufficiently low, but there could be a rather wide range of SNRs at which this was the case. Presumably, the upper limit to the speech level in the linear-gain scenario was the speech loudness, while in the distortion scenarios the individual tolerance to distortions was likely the limiting factor, at least for subjects designated as “distortion haters”.

Stability of individual preferences was also reported in previous studies testing the trade-off between good SNR and high speech fidelity (low degree of distortions) for noise reduction (NR) algorithms. Neher and Wagener (2016) showed stability of personal judgements by inviting subjects that had shown clear preferences for either weak or strong NR in two previous studies. Participants performed (among others) measures of maximally acceptable background noise levels, detection thresholds for speech distortions caused by NR, and adjusted the strength of the binaural coherence-based NR algorithm to their preferred level. Their results confirmed preservation of noise-reduction strength throughout the studies. The present data suggest that this stability of noise / distortion tolerance may extend also over a broader range of listening conditions with varying degrees of spatial unmasking, informational masking, and dip listening, supporting the basic assumption of the concept proposed by Völker et al. (2018)

that subjective traits remain constant over time as long as the acoustical context is the same or similar.

High test-retest correlations were also observed in the adjSN part of the study. However, this may not be a measure of subjective preference stability, but rather reflect the subjects' tendency to do not introduce large changes relative to the starting level, because this did not produce improvements in SNR. Such effects are likely to have occurred here, because the starting level was not randomized over the entire range of available gains to prevent uncomfortable loudness at presentation onset. In contrast to the gain adjustments (with and without distortions), results from the equalization scenario from both the adjS and the adjSN part showed very limited test-retest reliability. This is in contrast to observations reported previously for different groups of normal-hearing or hearing-impaired subjects and similar signal modifications (Rennies et al., 2016; Ciba, et al., 2014). One possible reason is that in the current study subjective choices regarding frequency shaping were not stable over time, which would limit the applicability of a personal trait along the dimension "high-frequency lover" vs. "low-frequency lover" to tuning speech playback in practical applications. Another reason could be that the employed user interface was a too big challenge for the participants to reliably find their individual preferences. One particular difference to the previous studies was that the direction and rotation of the "X" and "Y" axes were randomized, which had not been done in the previous studies.

Overall, it can be concluded that not all dimensions, along which individual preferences differ, can be considered stable without further considerations. This general observation as well as the possible role of the method used for assessing listening preferences has to be accounted for when attempting to apply knowledge of individual preferences to tuning of audio playback devices or audiological applications.

3.4.2. PREFERRED SPEECH PROCESSING IN FIXED NOISE CONDITIONS

This part of the study investigated preferences in the "near-end" listening enhancement approach, when the maskers in the scene were not modifiable, but the target speech was. In the current study preferred overall gain adjustments of all subjects in different masker and spatial conditions varied considerably. As expected, hearing-impaired listeners generally preferred

higher overall gain than normal-hearing subjects. The lower end of the gain adjustment scale in the present study was set to the individual SRTs. The SRTs differed strongly between co-located and spatially separated conditions due to spatial unmasking. The self-adjusted speech levels measured in the present study were always clearly above the individual SRT, which is also expected since listening to a speech at or close to SRT is very effortful according to combined measurements of speech intelligibility and listening effort (e.g., Rennies et al., 2014; Schepker et al., 2016). These studies used diotitically presented speech-shaped noise maskers in categorical listening effort scaling and found that NH and HI listeners judged listening to matrix sentences (as used in this study) at an SNR 11 dB above SRTs to be "low effort" (German: wenig anstrengend). In the present study, 11 dB above SRTs was the mean adjusted speech level for the co-located multi-talker masker (which is closest to the SSN used in the previous studies). This suggests that, on average, subjects self-adjusted the speech level to create low listening effort. Comparing adjusted speech levels in spatially separated and co-located maskers, it was observed that the differences were smaller than at SRT, although the differences were still significant. This indicates that the impact of binaural unmasking transferred to supra-threshold listening conditions, but that its contribution to the target perception (e.g., with respect to listening effort) is smaller than at SRT (in line with recent listening-effort data of Rennies and Kidd, 2018)

The adjS part furthermore allowed investigating the noise vs. distortion trade-off by comparing the linear gain adjustments to gain adjustments accompanied by clipping and compression artifacts. It was observed that subjects generally adjusted less gain in the two latter scenarios when the SNR advantage came at the cost of distortions (see Fig. 3.2, first and second row). The trade-off between artefacts and SNR seemed to be very similar for the two different processing schemes (see Fig. 3.2, third row), with a basically absent bias and high correlations of individual adjustments in the two scenarios. According to some related research it is not yet clear which dynamic range limiting technique (clipping or compression) is preferred by end users. Some studies suggested that hearing-impaired subjects (both underage and adults) preferred clipping to compression when the presentation level was relatively high (80 dB SPL), while their normal-hearing counterparts showed a clear preference for compression limiting at this presentation level (Stelmachowicz et al., 1999). In contrast, a study on subjects with mild-to-moderate hearing impairment (Hawkins, Naidoo, 1993) reported a significant preference for

compression limiting among this group. The current study does not reveal any preference for either of the processing schemes in any of the tested conditions.

Apart from the general and rather intuitive finding that, on average, the influence of distortions was reduced by the subjects by selecting a lower speech level at higher fidelity, the individual assessments of the present study allowed for a more detailed analysis. Previous studies (Völker et al., 2018; Marzinzik, 2000) suggested that some listeners show a higher tolerance to distortions or artifacts generated by NR algorithms ("noise haters"), while others would rather accept residual noise to avoid artifacts ("distortion haters"). The data of the present study supported this claim in that the formation of three subgroups ("noise haters", "distortion haters", "indifferent") based on the overall difference between linear gain and "distorted gain" produced significant differences between groups for each of the investigated scenarios. This provides direct evidence that classifying subjects along this perceptual dimension is a valid approach, which extends to more complex listening scenarios with different degrees of spatial unmasking, informational masking, and dip listening. This is also in line with the approach and conclusions of Neher et al. (2014) that preference for strong or weak noise reduction can be measured consistently, although it remains to be shown that the preference for the strength of noise reduction and gain adjustments, as investigated in this study, results in the same personal traits of individual subjects. Overall, the personal trait of being a "noise hater" or" distortion haters" seems to have potential to be used in, e.g., audiological applications.

From a practical point of view, the method employed here may deserve further attention in future research, because it can be administered very quickly, especially compared to paired-comparisons, as employed for noise reduction preference in previous studies (Neher, 2014; Neher et al., 2016). For determining the personal trait, the gain adjustment with and without distortions may be advantageous over adjusting the strength of noise reduction schemes, because the available range of SNRs and the corresponding distortions are wider and can be controlled more easily. For example, the noise reduction algorithm investigated by Neher and Wagener (2016) allowed for an SNR-benefit of at most 3.8 dB even in the most aggressive setting. In contrast, the gain adjustment of the present study theoretically allowed for a 30-dB range, and it is also applicable at negative SNRs, where many noise reduction schemes fail to operate as desired. If a consistent transfer between preferred noise reduction strength and the measure proposed in this study can be established, it might be possible to classify subjects as "noise haters" or "distortion haters"with a very limited number of simple gain adjustments.

3.4.3. PREFERRED SIMULTANEOUS PROCESSING OF TARGET SPEECH AND MASKERS

This part of the study investigated listening preferences when the processing adjusted by the subjects affected both target speech and maskers equally. In general, subjects introduced a larger linear gain than distorted gain (see Fig. 3.4), but the differences were smaller than in the adjS part, which was expected since no SNR benefit could be reached. The subjective variability in the linear gain scenario may give a hint at preferences for higher or lower overall listening level (at the same SNR) and possibly also with respect to the tolerance for noise. In general, subjects labeled as "distortion haters" in adjS part of the study in the majority (70%) chose higher overall level throughout conditions in the adjSN part, while previous "noise haters" tended to prefer lower levels. This latter trend may be due to a tendency to reduce the amount of noise perceived, but further investigation is required to explore if "noise haters" consistently prefer lower playback levels (at a given SNR) because this reduces the audibility of the noise.

3.4.4. RELATION OF PREFERENCES TO INDIVIDUAL FACTORS AND SRTS

In the adjS part of the study correlations between preferences and personal factors were in the great majority significant (more than 90% of all cases). Age had the smallest predictive power with squared correlations ranging from R^2=0.13 to R^2=0.35 throughout all experimental scenarios. In the gain scenario, the other individual factors (PTA, DTT and Goesa) held similar, relatively small predictive power as age ($R^2 \leq 0.36$ for masker conditions and $R^2 \leq 0.49$ in silence). Correlations between PTA, DTT and Goesa and individual preferences for the two other scenarios (compression and clipping) were on average higher (R^2 increased by on average 0.2) and were also highest for silence. In general, squared correlations between preferences and SRTs in the linear gain scenario showed similar predictive power as individual factors previously. In scenarios involving distortions, however, squared correlations between SRTs and preferences were considerably higher (by 0.25 on average) than in the gain scenario. Scenarios involving gain at the cost of distortions caused by compression or clipping may be comparable to investigating noise reduction processing preferences. Several previous studies (Neher et al.,

2014; Brons et al., 2012; Brons et al.2013; Brons et al., 2014) reported a dependence, or rather a tradeoff, between overall preference for given noise reduction processing and intelligibility scores. It was observed that both normal-hearing as well as hearing-impaired participants often favored noise reduction schemes causing reduction of speech intelligibility scores but at the same time reducing noise annoyance. Similar correlations between scenarios involving distortions and two different intelligibility measures (DTT and Goesa) may support those findings. Other factors, such as spatial constellation had a limited impact on the current correlations – revealing slightly higher correlations for spatially separated conditions, which was possibly due to a generally larger spread of SRTs in spatial conditions than in co-located conditions. Regarding masker type - silence brought the highest correlation results, and none of the other maskers seemed to have a noticeable influence on the outcome.

Correlations between preferences and personal factors as well as between preferences and SRT measures in adjSN part were roughly comparable with those obtained in adjS part of the study, yet with stronger dependencies in compression and clipping scenarios, especially with speech intelligibility test results (particularly Goesa) and PTA, which may also be related to the abovementioned dependence between noise reduction preferences and intelligibility scores. The highest correlations occurred between SRTs and listening preferences in the same conditions (see Tab. 3.6, cells marked in gray). This can probably be explained by the fact that many subjects kept the adjusted levels rather close to the starting presentation level, which was the individual SRT+9 dB.

3.5. CONCLUSIONS

The main findings of the current study can be summarized as follows:

- Subjects show high test-retest reliability in their adjustments of target speech level in fixed-noise scenarios for all investigated masker types and spatial conditions. In contrast, individual preferences for frequency shaping varied strongly from test to retest, possibly due to a more complex method of adjustment. This indicates that the stability of individual listening preferences and the role of their assessment method has to be accounted for when attempting to exploit individual preference profiles for hearing device fitting.

- The comparison of the speech level preferences between the linear gain scenario and the two different gain scenarios involving distortions (clipping and compression) experimentally confirmed the notion of the subjective trait to be rather a “noise hater” or a “distortion hater” as introduced by Völker et al. (2018). The grouping of the present subjects into either of these groups (or into a third, intermediate group) produced significant group differences that remained relatively stable through the tested conditions.
- SRT measures in complex listening conditions were reasonably good predictors for listening preferences in most tested conditions with R^2 values between 0.60 and 0.70. These amounts of explained variance in the preference data was larger than for age, PTA, or simple speech intelligibility test results (DTT, Goesa) used as predictors.

4. IMPACT OF SPORT EXERCISES ON INDIVIDUAL LISTENING PREFERENCES

This study investigated the influence of physical fatigue on listening preferences regarding playback volume and frequency shaping and how performing sport exercises accompanied with music could affect hearing thresholds. First, a survey-based study was conducted (N=138) on perceived changes in sound perception with growing physical fatigue. Next, ten normal-hearing participants underwent pure-tone threshold examination and preference adjustments both without a physical load present (at resting condition) as well as at different stages of induced fatigue. The main goals of the study were to investigate if and how sound perception changes during physical exercises, if such changes can be induced in a controlled laboratory setting, and if hearing thresholds are affected by listening to music during exercises. The survey results indicated that 52% of the participants experienced changes in their sound perception during sport. These changes caused 81% of them to modify sound settings during the exercises. The laboratory experiment suggested a consistent shift in listening preference of subjects when moving from resting condition through growing fatigue, resulting in an initial volume increase which continued up to an individual level of maximum fatigue, when the majority of subjects (70%) preferred to turn the volume down or to reduce high-frequency content. Additionally, there was no impact of physical exercise on pure-tone thresholds after music exposure at the levels tested here.

4.1. INTRODUCTION

Millions of people exercise to music every day. Yet very little is known about how physical fatigue affects listening preferences and hearing abilities. Previous studies have reported certain changes in auditory perception, such as temporary modification of hearing thresholds especially at high frequencies (Vittitow et al.,1994; Hutchinson et al., 1991), as results of sport exercises conducted while listening to loud music or noise, but no studies have yet investigated individual listening preference change with growing fatigue. This study therefore aimed to explore how listening preferences may change when listening to music is combined with a physical load, and also to re-examine previous findings on thresholds' change in such conditions (Vittitow et

al., 1994). A better understanding of the underlying effects may be relevant for adjusting playback settings following the physical state of the user and to provide hearing protection if needed. The approach of this study was to first explore a possible influence of physical exercises on sound perception by means of a survey and subsequently, based on the survey results, to conduct a psychoacoustic study to further investigate if such effects can be induced in a laboratory setting and lead to systematic patterns of individual listening preferences.

It is a well-known phenomenon that pure-tone thresholds increase after being exposed to high-intensity auditory stimuli, which is typically referred to as temporary threshold shift (TTS). The magnitude of TTS, the degree and rate of recovery depend on the exposure level, duration, frequency, and characteristics of the individual (Schlauch and Nelson , 2015). Threshold shifts of up to about 50 dB immediately after a single noise exposure have been reported to recover completely, while more intense immediate hearing losses are likely to result in permanent hearing loss (e.g., Ryan and Bone, 1978; Clark and Bohne, 1999). Several studies reported an increased effect of noise-induced TTS in response to physical activity in noise or music for both low-intensity (Hutchinson et al., 1991; Lindgren and Axelsson, 1988) and high-intensity exercises (Vittitow et al., 1994) among normal-hearing subjects. The first two studies (Hutchinson et al., 1991; Lindgren and Axelsson, 1988) tested subjects cycling on an ergometer at 40% of their maximal workload for 10 minutes while being exposed to high intensity noise (1/3-octave-wide noise centered at 2 kHz in both cases, at 104 and 105 dB SPL, respectively). The results clearly showed increased thresholds, especially for high frequencies. On average, thresholds increased by about 10 to 15 dB after noise exposure only (classic TTS), especially for frequencies of 6 and 8 kHz, relative to thresholds measured without previous noise exposure. When physical exercises were performed in the presence of the same noise, a change in TTS was observed. For some frequencies a lowering of hearing thresholds (i.e., a reduction in TTS) by about 1-3 dB occurred (e.g., 2 or 3 kHz,), but for other frequencies an increase by more less same amount was found. Another study by Vittitow et al. (1994) also employed cycling to induce physical workload, which was set to 70% of the maximum level of the maximum work capacity. Participants were asked to control the heart rate between 140 to 160 beats per minute. Therefore, they had to cycle with 50 rotations per minute and the workload was increased until the heart rate was constant in the target range. The physical exercise was accompanied with music stimuli at a level of 94.4 dB(A). The results showed a mean increase in thresholds of between 4-6 dB as a reaction to the exposure to the loud music only (classic

TTS) at the most impacted frequencies of 3, 4, and 6 kHz. When participants performed sport exercises while listening to that stimulus, a further mean threshold increase of about 1-3 dB in comparison to music exposure only condition was found for the frequencies of 3, 4, 6 and 8 kHz. In addition to increased TTS, the studies of Hutchinson et al. (1991) and Vittitow et al. (1994) indicated two further aspects: firstly, physical activity performed in silence could lower pure-tone thresholds by on average almost 2 dB, hence improve hearing abilities. Secondly, increased TTS due to physical exercise was not found for all conditions, but noise or music exposure alone could lead to similar TTS as such exposure combined with physical exercise, i.e., here reports are conflicting. Vittitow et al. (1994) found that TTS introduced by exposure to loud music only was (for the majority of tested audiogram frequencies) lower than TTS due to combined music exposure and physical exercises. In contrast, the Hutchinson et al. (1991) reported higher TTS for noise exposure alone for the majority of cases (hence, no detrimental effect of additional physical exercise). The reason for this discrepancy is unclear. It is also currently unknown if TTS may also be influenced by physical activity at lower – and possibly more common – noise or music exposure levels. Hence, the available knowledge base for this effect is very limited.

The mentioned studies focused on tone-detection thresholds and did not investigate any other auditory phenomena. The primary goal of this study was to test if and how physical fatigue may affect supra-threshold hearing, i.e., individual listening preferences as observed in preferred playback volume and frequency-shaping adjustments. Most music players allow users to adjust not only volume but also frequency-shaping settings, but to the best of our knowledge it is currently unknown whether there is a measurable relation between personal playback preferences and sensory data on the physical status of a listener. On the basis of the positive influence of physical arousal on perceptual-cognitive tasks (Biagini et al., 2012) and the results shown by Vittitow et al. (1994), the association of physical exercise and hearing should be investigated in more depth. As a control measurement, the possible influence of physical exercise on pure-tone thresholds was also measured. Combining measurements of individual hearing thresholds and listening preferences with physiological sensor data monitoring the physical status of subjects (pulse oximetry, heart rate, etc.) may reveal possible interdependencies between these factors and provide a basis for future audio applications such as automatically adjusting playback settings according to the current status of the listener.

4.2. SURVEY-BASED ASSESSMENT OF SOUND PERCEPTION DURING PHYSICAL ACTIVITIES

Before the laboratory experiment reported in this paper was designed, an evaluation of the notion that sound perception may depend on physical fatigue and that automation of the playback in accordance to this change could be beneficial to sportsmen was tested by means of online surveys. Their outcome was the main motivation, as well as a guide, in the design of the subsequent listening preference test under laboratory conditions.

4.2.1. METHODS

This part of the study consisted of two surveys: the first survey aimed at answering the fundamental question of whether or not the phenomenon of a change in sound perception during sports activities was perceptible. The second survey had three additional questions further exploring the trends indicated by the results of the first one. The number of participants in the first and second survey was 94 and 44 responders, respectively. The survey questions are presented in the Tab. 4.1, with the ones used only in the second survey highlighted in gray. Both surveys were announced to address those who listen to music during sport exercises. The surveys differed slightly in their implementation. In the first one only questions no. 1, 2, 3, 5, 6, 7 were asked, and questions no. 6 and 7 did not contain the answer “other”. There was an open field at the end of the survey dedicated for comments and answers different than those proposed. In the implementation of the second survey all nine questions were asked. For each question from no. 6 to 8 subjects were encouraged to indicate responses not provided as multiple-choice options in the open field marked as “other”. The surveys were web-based and conducted over the course of about one week. The language of the first survey was Polish (the translation of which is reported in the Tab. 4.1), and the second survey was conducted in English.

TAB. 4.1: QUESTIONS FROM SURVEY NO.1 (WHITE) AND SURVEY NO. 2 (WHITE AND GRAY). SOME OF THE QUESTIONS DID NOT PROVIDE PREDEFINED RESPONSE ALTERNATIVES (OPEN QUESTIONS), OTHERS OFFERED MULTIPLE CHOICE OPTIONS.

No.	Question	Response alternatives
1	Sport discipline to which you listen to music most often.	open question
2	What is your age?	Single choice: 18-25, 25-35, 35-45, 45-55, 55-65, 65-75, 75-85
3	Which devices do you use most frequently for listening to music during sport?	Single choice: headphones, loudspeakers
4	How much time do you spend on sports per week?	open question
5	Do you notice changes in sound perception while exercising?	Single choice: yes, no; option to quit the survey after this question in case of answering no
6	What kinds of sounds bother you? (if 5 answered YES)	multiple choice: high-frequency sounds, low-frequency sounds, impulse sounds, long-lasting sounds, loud sounds, other
7	What action do you undertake? (if 5. answered YES)	multiple choice: turn the music off, change volume settings (up or down), change EQ settings, withstand, change playlist, other
8	If you experience hearing inconvenience, how would you describe it?	multiple choice: I hear squeaks and ringing in my ears, I hear noises, I feel pain or pressure in my ears, other
9	How long does your inconvenience last? (in minutes)	open question

4.2.2. RESULTS

The results of both surveys are reported here jointly. In total 36 out of 138 subjects (26%) explicitly reported experiencing a change in the sound perception (i.e., they answered "yes" to question no. 5 in Tab. 4.1). Interestingly, another group of 36 individuals (in both surveys in total) answered that they did NOT notice the phenomenon but reported being bothered (question no. 6) or experiencing hearing inconvenience (question no. 8). Since this implied that they judged their sound perception to be affected by doing sports, they were counted as part of the total pool of results as those who "experience changes in sound perception or related effects". When these two groups were combined, the number of responders who either directly reported experiencing the phenomenon per se or indicated related effects was 72 responders, i.e., 52% of the participants.

There are some general results (questions 1 to 4) to be mentioned before a more detailed description of results relating to changes in sound perception is given below. First, the phenomenon seemed to not be related to any particular sport discipline. 48% of the responders indicated running as their primary discipline, but several others were also indicated (e.g., fitness, aerobic, cycling, gym). Second, the phenomenon occurred for both playback modes, where 44% of loudspeakers user and 57% of headphones users were affected indicating that the effect was somewhat more common for the latter group. Third, there was no clear link between the time spent on sport weekly and the occurrence of the perceived changes in sound perception. The surveys were mainly filled in by young people (91% were aged below 36 years). However, the effect tended to be reported more frequently among older participants (>35 years): seven out of seven and three out of four participants in the age groups 36-45 years and 46-55 years, respectively, reported that physical exercise affected their sound perception.

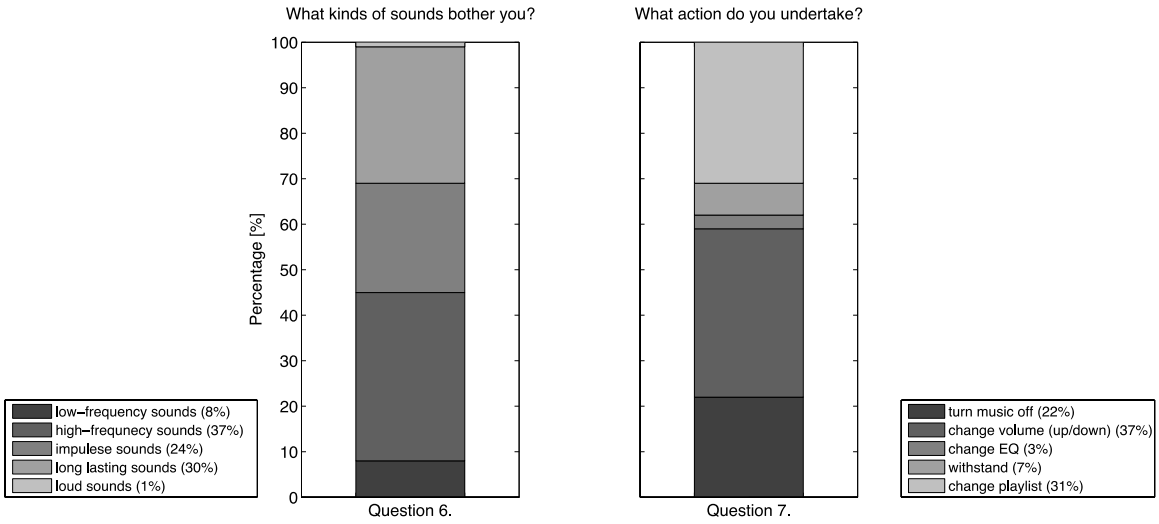

FIG. 4.1: PROPORTION OF RESPONSES TO QUESTIONS NO. 6 (LEFT) AND 7 (RIGHT).

Figure 4.1 summarizes the results of the survey related to perceived auditory effects while doing sports. These data are based on the responses of the 72 participants who had indicated changes in auditory perception (questions 6 and 8). The left bar shows responses as to the type of sounds that were perceived as bothering during sports (question 6 answered by 66 responders, multiple choice). The most frequently stated type of sound was high-frequency sounds (37%), followed by long-lasting ones (30%). Other sound types mentioned by the participants were low-frequency sounds, impulse sounds and loud sounds. With respect to manual changes introduced to the playback settings (question 7, answered by 64 participants, multiple choice), the combined results showed that 59 responders introduced changes to the music playback during sport activities - meaning they answered question no.7 but did not choose “withstand” option as their only choice, or chose it as an alternative to other answers (1 person indicated turning volume down or withstand as answers). These 59 responders make up about 81% of the whole 72-participants’ group that indicated noticing changes in auditory perception. Remaining 5 out of 64 subjects chose “withstand” as their only answer to the question 7. Among the playback changes introduced (see right bar of Fig. 4.1), the most common was changing the volume (up or down, 37% of answers in total, with only two of them indicating changing the volume up versus eleven replies of changing the volume down) or changing the playlist (31%), 22% of answers reported turning the music off entirely.

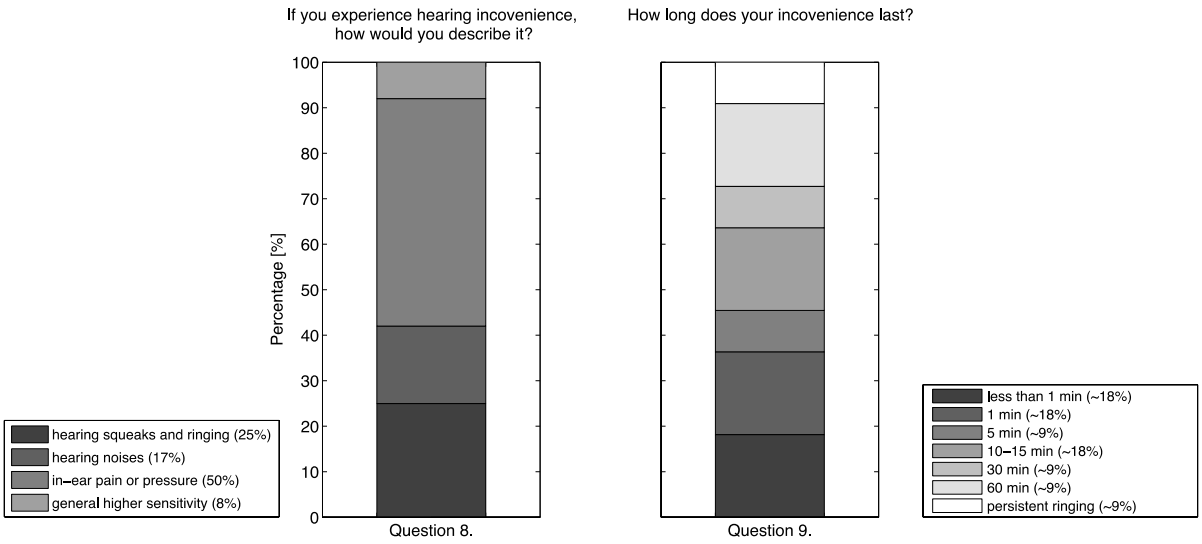

FIG. 4.2: PROPORTION OF RESPONSES TO QUESTIONS NO. 8 (LEFT) AND 9 (RIGHT)

Figure 4.2 illustrates the data related to possible hearing inconvenience and its duration. Eleven out of 44 responders of the second survey answered question no. 8 and reported that they experienced hearing inconvenience during sport activities (left bar). The most frequently indicated answer was that they perceived pain or pressure in their ears (chosen by 6 participants), followed by squeaks and ringing (3 participants) and hearing noises (2 participants). The duration of such inconveniences (right bar) was also reported by 11 participants of the second survey, and in about 70% of answers lasted no longer than 30 minutes, and in 50% of answers no longer than 15 minutes (this question was not multiple choice).

In summary, the results of the surveys indicated that about half of the participants perceived some (mostly uncomfortable) auditory effects while listening to music during sport activities. The lack of systematic effects of playback device and sports discipline may indicate that this is a rather general phenomenon not specific to any particular group. The fact that the most common type of sounds causing annoyance was high-frequency sounds may be related to the results of previous studies, where exposure to loud sounds during physical fatigue resulted in increase in hearing thresholds that occurred mainly for higher audiogram frequencies (see Vittitow et al., 1994). It is interesting to observe that the majority of subjects did some action to overcome hearing inconveniences, e.g., by turning the volume down or changing the playlist. This indicates that the hearing inconveniences can be bothersome enough to make the listener react (but not bothersome enough to make the user quit listening to music while doing sports). This may point to some potential for automation in future audio devices if a reliable relation

between individual sound perception and physical status could be established. A combined monitoring of the individual exercise status and a system that learns the responses of the listener could introduce these changes automatically and thereby increase the listening comfort of music playback. Some of the indicated responses might even raise more serious concerns related to hearing health in addition to comfort and preferences. The indicated longer lasting sounds persisting several minutes after training or the perception of pain or pressure suggest that hearing protection should at least be considered when investigating sound perception during physical exercises. In the following laboratory-based study, we therefore measured not only individual listening preferences, but also pure-tone thresholds to explore if undesired effects of elevated thresholds occurred during the conditions of the experiment.

4.3. LABORATORY-BASED MEASUREMENT OF INDIVIDUAL SOUND PREFERENCES AND HEARING THRESHOLDS

4.3.1. SUBJECTS

Ten healthy and normal-hearing subjects (self-declared) - four women, six men, aged 22 to 48 years (median: 26.5 years) - participated in the study. Subjects were asked to fill out a short survey asking for their age and the number of hours spent for sport weekly, which turned out to be from 2 to 10 hours (median 4.5 hours). Participants took part in both auditory and physical exercises (maximum workload (WL) measurement) and were paid on an hourly basis. The sex, age and heartbeat tempo for maximum WL as well as for 70% of the individual WL are presented in Tab. 4.2.

TAB. 4.2: PHYSIOLOGICAL DATA OF THE SUBJECTS. HEART RATES ARE INDICATED AS BEATS PER MINUTE (BPM), SEX IS INDICATED AS F – FEMALE, M - MALE.

Subject number	Sex	Age / years	Heart rate for maximum WL estimation / BPM	Heart rate for 70% of WL estimation / BPM	Maximum WL / Watt	70% of maximum WL / Watt
1	F	25	174	122	120	84
2	F	29	171	120	180	126
3	M	24	168	118	180	126
4	F	48	134	94	180	126
5	F	23	172	120	180	126
6	M	28	174	122	240	168
7	M	28	195	137	260	182
8	M	24	191	134	280	196
9	M	22	180	126	300	210
10	M	34	189	132	340	238

4.3.2. APPARATUS AND MEASUREMENT CONDITIONS

The experiment was performed in the Sport Motor Behaviour Laboratory (Sportmotorik-Labor) of the University of Oldenburg, Germany. The laboratory was surrounded by a "silence zone" announced on written notes. All doors in a radius of ca. 50 m were kept closed or were blocked from smashing, no sport activities were held on the sport fields nearby when the experiment was running. The experiment was held during the summer break, almost no pedestrian traffic occurred in the corridors or on the staircase nearby the lab. In addition to these precautions, the experimenter was present in the laboratory at all times to ensure that no audible noise from outside affected the experiments. The study was conducted using a Matlab software environment installed on a personal laptop computer (DELL LATITUDE E7450), an RME Babyface USB 2.0 Audio Interface sound card, and a Lenovo ThinkVision LT 1423p touch screen. The acoustic stimuli were presented to the subjects via Sennheiser HDA 200 headphones that were calibrated to dB SPL using a Bruel&Kjær (B&K) 4153 artificial ear, a B&K 4134 microphone, a B&K 2669 preamplifier, and a B&K 2610 measuring amplifier. The impact of the headphones was free-field equalized. The maximum WL examination was performed using a cycling ergometer (Ergoline select 200 P) conforming to the DIN-norms for medical devices. During the step test on the ergometer the participants' heart frequency was controlled using a mobile heart rate computer (Polar FS-400).

The study consisted of three sessions (described in detail below), with not less than 48 hours break in-between, during which subjects were asked to not engage in loud activities such as concerts or sports events. The first session lasted about 2h and consisted of three blocks: 1) hearing threshold and listening preferences measurements in resting condition (no sport exercises), 2) exposition to loud music stimuli followed by hearing threshold measurements, and 3) estimation of the individual maximum WL of the subject. The goal of this session was to collect reference data on thresholds, preferences, and 100% personal WL, as well as to measure a possible threshold shift after loud music exposure without exercises. The second session (about 30 min) consisted of sport exercises in the presence of loud music stimuli, followed by hearing threshold measurements to test a possible threshold shift for combined sport and music exposure. During the third session (also about 30 min), subjects performed sport exercises while listening to music, and adjusted the playback settings to their individual

listening preferences (see below). Afterwards, hearing thresholds were measured to test if they were affected by the playback as adjusted by the subjects.

4.3.2.1. SESSION NO.1 (S1)

During the first session subjects were asked to fill out a survey on their age and the number of hours they spend on sport weekly. Next, participants performed the listening test, during which pure-tone hearing thresholds at 2, 3, 4, 6, and 8 kHz were measured (session one audiogram one - S1AU1) using the procedure described in the next section. These frequencies were selected based on previous studies, which had reported the main effect of TTS to be present in this frequency range (Vittitow et al., 1994; Hutchinson et al., 1991; Lindgren and Axelsson, 1988). The subsequent task of the subjects was to adjust playback parameters to their own listening preferences for equalization and volume adjustment of ten music pieces (session one, preference measurement: S1PM). The procedure for listening preference adjustment is described in the next section. Then, subjects were asked to adjust the level of one randomly chosen music piece from those used in the study (see Table 4.3) to their sensation perceived as "loud" and listened to all music pieces with that volume setting for 20 minutes. Afterwards, within less than 60 seconds break subjects performed a second hearing-threshold measurement (S1AU2) for the same five frequencies. Finally, the individual maximum WL was assessed by cycling on an ergometer and monitoring the WL as described in the next section.

4.3.2.2. SESSION NO.2 (S2)

The second session consisted of 20 minutes of cycling with a WL set to 70% of each subject's personal WL (estimated during the first session). The heart rate frequency was controlled to remain at a submaximal condition of 70-75% of the maximal heart rate frequency determined in session 1. If the heart rate increased higher than 75% of the individuals' maximum, the WL on the ergometer was decreased to control the submaximal exercise condition using the heart frequency. While cycling, subjects listened to the same ten music pieces with the volume set to their level of sensation perceived as "loud" (also determined during the first session).

Afterwards, within less than 60 seconds break, subjects performed the hearing-threshold measurement (S2AU) using the same method and frequencies as in the first session.

4.3.2.3. SESSION NO.3 (S3)

During the third session subjects cycled on the ergometer with the load again set to 70% of their personal WL and were asked to either adjust equalization and volume of the music pieces they were listening to or were instructed to listen without performing any adjustments. In total, 18 song excerpts were used, which were presented in three blocks of songs (A, B, and C). Block A consisted of four songs (see Table 4.3, rows 2-5), blocks B and C consisted of three different songs (see Table 4.3, rows 6-8 and 9-11). The presentation order was A-B-A-C-A. The selection of songs for each block and their order within a block was the same for each block presentation. For each song of block A, subjects were asked to adjust equalization and volume to their preferred settings using the procedure described below. The adjustments are indicated S3PM1, S3PM2, and S3PM3 in the following. During the presentation of blocks B and C both equalization and volume setting were fixed to the individually selected settings for the same music pieces obtained in the first session, and no further adjustments were possible during playback (subjects listened only). After listening to all 18 songs, within less than 60 seconds break, subjects performed the hearing-threshold measurements (S3AU) using the same method and frequencies as previously.

4.3.3. STIMULI AND MEASUREMENT PROCEDURES

4.3.3.1. HEARING-THRESHOLD MEASUREMENTS

Hearing thresholds were measured using the procedure proposed by Lecluyse and Meddis (2009). This procedure was selected because it is relatively quick, which was considered important for the present study to be able to perform the measurements within a very short period after music exposure. Stimuli were presented monaurally to the right ear and consisted of pure-tone bursts with frequencies of 2, 3, 4, 6, and 8 kHz and a duration of 10 ms. The

different frequencies were interleaved and randomized during the measurement to minimize sequential effects. Each frequency was measured with 20 trials. During each trial, the stimulus consisted of a leading tone pulse (called “cue”) and “test” tone pulse following after 30 ms at the same frequency, but with a 10-dB lower level. The task of the subject was to indicate how many tones were audible: one (only the “cue”), two (both “cue” and “test” tones), or none. No catch trials (only “cue” tones) were included in this study to minimize measurement duration, i.e., the presentation always consisted of two-tone bursts (the subjects were not aware of this). If the subject answered correctly that he or she had heard two tones, both tones were decreased in level, otherwise the level was increased. The initial step size was 10 dB and was reduced to 2 dB after the first reversal employing a 1-up, 1-down alternative-forced choice (AFC) procedure. The starting presentation levels of the procedure were 40 and 30 dB SPL for “cue” and “test” tone, respectively.

4.3.3.2. LISTENING PREFERENCE MEASUREMENTS

Music stimuli were used to assess individual listening preferences. Stimuli were excerpts of ten songs with tempi not slower than 80 beats-per-minute (BPM) (see Tab. 4.3) and uplifting, motivating character. They were cut in length (preserving melodic structure) to last approximately one minute and presented without any pauses in between. Music pieces were equalized in digital level to -50dB full scale (FS), which corresponded to an average level of 65 dB SPL in the calibrated experimental setup. Subjects were asked to adjust the music according to their listening preferences with respect to volume and EQ. The volume adjustment was implemented as a horizontal slider, along which subjects could adjust the broadband level over a range of 20 dB (when subjects were instructed to set the gain adjustment to the sensation corresponding to “loud”, this range was increased to 30 dB). The equalization was performed by moving a point (the cursor) on a two-dimensional plane. The 2D plane represented a matrix of 19x19 equalization presets, where each preset corresponded to equalizer settings which applied up to 15 dB boost or attenuation to the signal in eight frequency bands, i.e., a lowpass band with a cut-off frequency of 250 Hz, an high-pass band with a cut-off frequency of 6 kHz, and six bandpass bands centered at 500, 1000, 1500, 2000, 3000, and 4000, respectively. The filter gains were set such that a perceptually continuous transition from low-frequency boost to high-frequency boost was achieved by moving the cursor along one axis, and a perceptually

continuous transition from mid-frequency attenuation to mid-frequency boost was achieved by moving the cursor along the other axis. This kind of user interface had been applied in previous studies investigating listening preferences (Ciba et al., 2014; Rennies et al., 2016) and was chosen because it typically enables users to find their individual preferences within one minute or less. In general, it is possible for these kinds of user interfaces that the visible position of the slider bar (for the volume adjustment) and the cursor point on the 2D plane (for the equalizer settings) may affect the adjustments of the subjects, for example, by biasing them into ending up at similar visual positions irrespective of the underlying auditory perception. To minimize such bias effects, the direction of the level change while manipulating the slider was randomized to increase from left to right or vice versa. Similarly, the starting point of the slider was randomized for each adjustment. However, the latter was limited to be within 0% to 10% of the slider range to avoid too loud levels at stimulus onset. Similarly, the axes rotation and direction in the 2D user interface was randomized, i.e., low- to high-frequency boost could be along the x- or the y-axis, and it could be in both directions (e.g., starting with low-frequency boost at the left / bottom or vice versa). No labels were indicated in the UI, i.e., subjects had to try out the changes in sound processing by moving the cursor and listening in real time. At the beginning of each equalizer adjustment the starting point of the cursor on the 2D plane was randomized. Naming convention for both audiogram and preference measurements are listed in the Tab. 4.4.

TAB. 4.3: EXCERPTS FROM MUSIC PIECES USED IN THIS STUDY. THE PRESENTATION BLOCK FOR EACH SONG IS INDICATED IN PARENTHESES; BPM STANDS FOR BEATS PER MINUTE AND RELATES TO MUSIC TEMPO HERE.

Song	Artist	BPM	Duration (min:sec)
Always on My Mind (A)	Pet Shop Boys	125	1:13
Call Me (A)	Blondie	142	1:08
Wake Me Up Before You Go-Go (A)	Wham!	82	0:59
Ray of Light (A)	Madonna	127	1:07
Can't Buy Me Love (B)	The Beatles	171	1:11
The Final Countdown (B)	Europe	118	1:13
Maria (B)	Blondie	160	1:02
Waterloo (C)	ABBA	148	1:04
Lady Madonna (C)	The Beatles	110	1:10
What else is there (C)	Röyksopp (Trentemoller RMX)	123	1:02

TAB. 4.4: MEASUREMENTS NAMING CONVENTION

Audiogram measurements naming convention	
Session 1 [S1AU1]	reference (no sport, no preceding loud acoustic stimuli)
Session 1 [S1AU2]	measurement after exposure to loud music (to measure classic TTS by comparing to S1AU1)
Session 2 [S2AU]	measurement after sport exercises combined with exposure to loud music (to test the influence of physical exercise on TTS by comparing to S1AU1 and S1AU2)
Session 3 [S3AU]	after sport exercises combined with music playback set to preferred settings (to measure if TTS occurs when preferred adjustments are enabled)
Preference judgement naming convention	
Session 1 [S1PM]	reference (no sport, no loud acoustic stimuli prior the measurement)
Session 3 [S3PM1]	beginning of sport exercises (first occurrence of block A)
Session 3 [S3PM2]	middle of sport exercises (second occurrence of block A)
Session 3 [S3PM3]	end of sport exercises (third occurrence of block A)

4.3.3.3. MAXIMUM WORKLOAD ASSESSMENT

From sports research it is known that endurance performances (WL capacity) and the heart rate (amount of heart beats per minute) are very individual parameters. Vittitow et al. (1994) used the individual WL capacity (70% of individual's maximum WL) by controlling the heart rate in a span of 140 to 160 beats per minute for all participants. In contrast to that, in the present study besides tracking the individual's WL capacity adjustment, the control of the heart rate of each participant was in main focus as a precise guide to adjust load. The all-out measurement was oriented on recommendations of the German association of sports medicine and prevention

(Boldt et al., 2002); benchmarks given by the association for untrained persons were adapted on the ergometer used in this study. Due to the large variance in individual maximum heart rate this span was individually defined for each participant using a maximum watt test (performed in 20-Watt steps) in the first session for defining the maximal heart speed. In line with (Vittitow et al., 1994), the exercise duration was set to 20 minutes with an individual heart rate span from 70% to 75% of the individually set maximum. For warming up, participants were allowed to cycle up to 5 minutes on the ergometer without any WL. Prior to the measurements the seat height for every participant was individually adjusted. For measuring the maximum heart rate, a test was performed with each participant to define the individual zone of a defined submaximal workload of 70%. Participants started to perform such examination with 60 watts for two minutes. Afterwards, each two minutes the workload increased by 20 watts and participants were asked to cycle keeping the heart rate in-between 60 to 80 beats per minute. The test was terminated when participants were not able to cycle with the prescribed number of beats per minute and therefore the WL increased too high. Afterwards, participants were asked to cycle without any load to regenerate. Finally, the data were downloaded from the wireless computer and the maximum heart rate was defined for each participant. On the basis of that the submaximal WL capacity of 70% to 75% was calculated for session two and three. In addition, the maximum watt capacity for each participant was measured for defining the thresholds in session two and three.

4.4. RESULTS

4.4.1. PREFERENCE ADJUSTMENTS

In order to assess the individually preferred equalizer settings for the four songs of block A (see Tab. 4.3), the spectra of the songs were analyzed in third-octave bands, and then the adjusted level changes for each frequency band were averaged across songs for each subject and adjustment block. The resulting frequency-dependent gains are shown in Fig. 4.3, where each panel shows the preferred gains of one subject, and different line styles represent the different adjustments. Comparing the reference listening preferences collected without sport exercises and without being seated on the ergometer (the solid black line marked as S1PM) with

preference shift during the third session at the beginning of the exercise and after about 2 minutes warmup (dashed line marked as S3PM1) it can be observed that for the majority of cases a significant increase of level (upwards-shift of gains) could be observed accompanied with a boost of low frequencies (<1000 Hz). The second adjustment in session 3, taken in the middle of the physical exercise (indicated by the dashed line marked S3PM2) generally showed a further increase of preferred gains at this increased level of fatigue. However, some subjects (no. 5, 7 and 9) also selected a significant reduction in high-frequency gain. The last frequency judgement at the end of the exercise in session 3 is plotted as dashed lines marked S3PM3. Six out of the ten subjects adjusted the gain to produce a reduction in level or/and in high-frequency content in comparison to the previous adjustment (S3PM2). Subjects no. 1, 6 and 8 decided to further increase the gain of the signal.

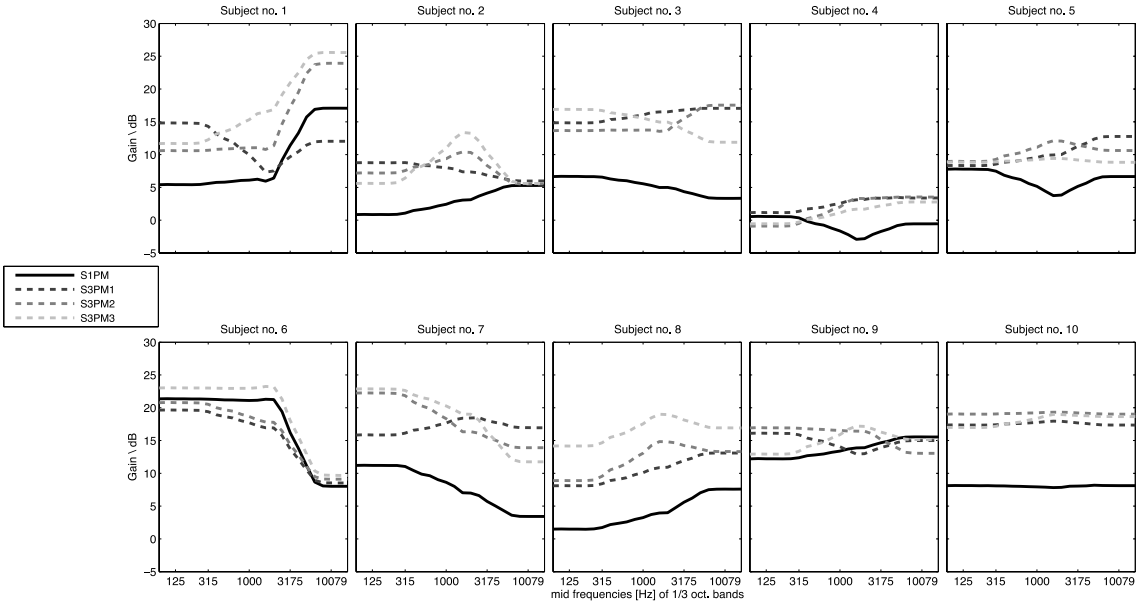

FIG. 4.3: INDIVIDUAL LEVEL CHANGES INTRODUCED BY THE SUBJECTS TO THE AUDIO MATERIAL RELATIVE TO THE ORIGINAL MIX, AVERAGED ACROSS SONGS AND PRESENTED IN 1/3 OCTAVE FREQUENCY BANDS. SOLID BLACK LINES REPRESENT ADJUSTMENTS IN THE FIRST SESSION; DASHED GREY LINES REPRESENT THE THREE ADJUSTMENTS IN SESSION 3.

To further analyze the observed effects, especially in the frequency range where the biggest change in hearing threshold may be expected (based on previous studies), the adjusted gains in third-octave frequency bands covering the high-frequency range in which equalizer adjustments were possible (2 to 6 kHz) were analyzed and plotted in Fig. 4.4. Comparing the preferred gain for these frequencies it could be observed that the first level increase at the

beginning of the sport exercises (compare first and second bar) continued with increasing fatigue up to a certain point (subsequent bars), at which the majority of subjects decided to reduce either the overall level, the high-frequency content, or both (except for subject no. 1 and 8, and slightly subject no. 6). These data were analyzed by means of separate repeated-measures ANOVAs for each frequency (2, 3, 4, and 6 kHz) to asses if the factor adjustment (S1PM, S3PM1, S3PM2, S3PM3) had a significant effect on adjusted gains. Degrees of freedom were Greenhouse-Geisser corrected and the significance level was set to $\alpha=0.05$. For each frequency, the effect of adjustment was significant [2 kHz: $F(2.004,18.040)=16.346$, $p<0.001$; 3 kHz: $F(1.863,16.765)=12.021$, $p=0.001$; 4 kHz: $F(1.901,17.109)=9.936$, $p=0.002$; 6 kHz: $F(2.025,18.223)=6.171$, $p=0.009$]. Post-hoc comparisons with Bonferroni correction for multiple comparisons showed that this was due to the differences between S1PM and S3PM3, and between S1PM and S3PM2, which differed significantly for each frequency, indicating that gain preferences at these frequencies changed at the two highest levels of fatigue. The remaining comparisons were not significant.

To further explore the changes in preferred gains due to increasing fatigue, the difference between adjustments after 20 minutes of exercises at 70 % of maximum personal workload (S3PM3) and the preferred listening settings in the reference conditions (S1PM) were derived. These differences are shown for each subject in the bottom panels of Fig. 4.4. This representation illustrates that all subjects preferred to introduce a gain >0 dB to this frequency region when music was accompanied by sport exercises compared to adjustments without sport exercises. Nevertheless, in every case it could be observed that the level increase declined with increasing frequency. The degree of this downwards sloping gain was very pronounced for some subjects (no. 2, 5, 7, and 8), and flatter for others (no. 1, 4, 6, 10), yet the trend in this subjective behavior was rather consistent. As a measure for this change across frequency, first-order polynomials were fitted to the gain-change-vs.-frequency data for each subject. The resulting slopes were always negative, ranging from about -0.1 dB/kHz (subject no. 6) to -2.2 dB/kHz (subject no. 2). After confirming that normality of the derived slope data could be assumed, a t-test was conducted to test if the mean slope differed significantly from 0. This was the case ($t(9)=-3.528$, $p=0.006$), indicating that, despite considerable interindividual differences, a significant group effect of the frequency-dependence on the adjusted gains differences between the reference conditions and the condition with maximum fatigue was observed. Analogous analyses were also conducted for the differences between the first and

second adjustment of session 3 (S3PM1 and S3PM2) and the reference adjustment (S1PM) (not shown in Fig. 4.4). Here, the slopes did not differ significantly from 0, indicating that no frequency-dependence of the gain differences could be observed.

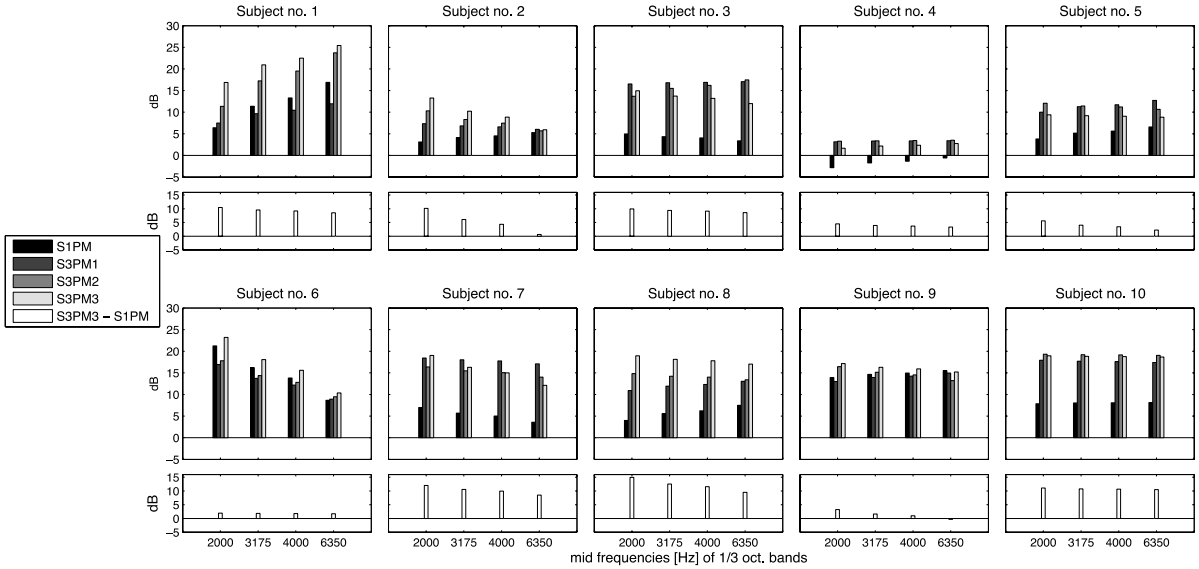

FIG. 4.4: THE TOP PANELS IN EACH ROW SHOW INDIVIDUAL CHANGES INTRODUCED TO THE AUDIO MATERIAL (COMBINED EQ AND VOLUME CHANGES) AVERAGED ACROSS SONGS IN 1/3-OCTAVE FREQUENCY BANDS (RMS LEVEL PRESENTED), COVERING THE HIGH-FREQUENCY RANGE THAT COULD BE ADJUSTED IN THE PREFERENCE MEASUREMENT (2 TO 6 KHZ). THE BOTTOM PANELS IN EACH ROW REPRESENT THE CHANGE BETWEEN THE MEASUREMENT AT THE HIGHEST FATIGUES (S3PM3) AND THE REFERENCE PREFERENCE ADJUSTMENT IN THE FIRST SESSION WITHOUT ANY PHYSICAL EXERCISE (S1PM).

4.4.2. HEARING THRESHOLDS

The average hearing thresholds across subjects are shown in Fig. 4.5. Error bars represent standard errors. In general, the hearing threshold measurements revealed the expected variance between subjects. Black bars labelled as S1AU1 represent the reference audiogram of the subjects measured at the beginning of the first session. With a single exception, the measured references pure-tone thresholds (without noise exposure and physical exercise) confirmed the self-reported normal hearing status of the subjects (thresholds ≤ 20 dB HL). Only subject #8 had thresholds of 25 and 36 dB HL at 6 and 8 kHz, respectively. The second bar (S1AU2)

indicates threshold values after being exposed to loud music for 20 minutes (with no sport included). Here, it is important to note that each subject was asked to indicate the volume level individually perceived as “loud” before exposure. This self-adjusted level varied from 66 to 90 dB SPL. The mean “loud” level of sensation chosen by subjects during the first session was 81 dB SPL. These individual levels were also employed in the second session, where listening was combined with physical exercise (third bars, S2AU). For the final threshold measurement after sport and self-adjusted playback settings (S3AU), the individual playback levels depended on the preference adjustments of each subject. These levels (averaged across songs) ranged from 64 to 90 dB SPL, with a mean level of 79 dB SPL across subjects.

Comparing the mean thresholds measured in the different sessions, only small differences were observed, which in general did not exceed 3 dB at any frequency, although there was a trend for thresholds to be slightly lower in the first session than in the second session. This observation was confirmed by separate repeated-measures ANOVA conducted for each frequency (significance level $\alpha=0.05$, degrees of freedom Greenhouse-Geisser corrected). The independent factor was measurement session. This analysis revealed that the main effect of session on thresholds was only significant for 3 kHz. However, subsequent pairwise comparisons showed that the individual sessions did not differ significantly from each other when Bonferroni corrections for multiple comparisons were applied.

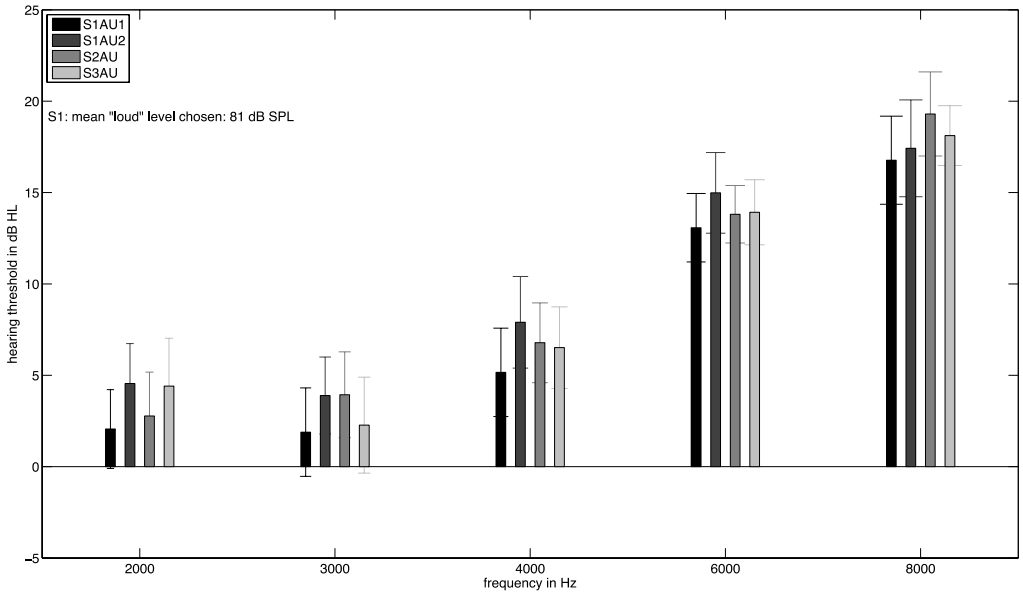

FIG. 4.5: MEAN PURE-TONE THRESHOLDS WITH STANDARD ERRORS ACROSS SUBJECTS FOR ALL MEASUREMENTS. THE MEAN LEVEL CHOSEN BY THE SUBJECTS DURING FIRST SESSION AS CORRESPONDING TO A SENSATION OF "LOUD" WAS 81 DB SPL, WHILE DURING THE THIRD SESSION THE PLAYBACK ADJUSTMENTS (GAIN AND EQ) RESULTED IN A MEAN LEVEL OF 79 DB SPL.

4.5. DISCUSSION

Exercises are well known for their positive impact on health and wellbeing (Norris et al., 1992) and music is an important factor associated with their enjoyment (Wininger and Pargman, 2003). The synchronization of body movements with music can lead to extended exercise duration, particularly among non-highly trained participants, being an ergogenic aid in numerous aerobic activities (Anshel and Marisi, 1978; Ramji et al., 2016). Music is also known for its impact on the rate of perceived exertion (RPE). Numerous studies suggest that sport accompanied with music leads to lower RPE than sport performed in silence, especially for untrained subjects (Mohammadzadeh et al., 2008; Szmedra and Bacharach, 1998). Among the positive impact of music, the studies list an increased ability to relax, hence reducing muscle tension and thereby increasing a blood flow, which accelerates lactate clearance and leads to lower lactate production in working muscle. Jia et al. (2016) investigated the impact of music on the autonomous nervous system and found that exercising with music could lead to better post-exercise recovery by the body as music attenuated a decrease in parasympathetic nervous system activity after physical load occurred. Knowing how often music and sport go together and that music serves as a means to increase the enjoyment of sport exercises, it seems important to look at this topic also from the perspective of hearing research and listening comfort. The major goal of this study was to investigate a possible change in listening preferences due to the physical fatigue, but also to explore a possible need for protection from (potentially increased) TTS reported in the literature when sport exercises are accompanied by music (Vittitow et al., 1994).

4.5.1. LISTENING PREFERENCE ASSESSMENTS

A possible shift in listening preferences due to the growing fatigue of the listener was investigated by the means of online surveys as well as in the laboratory experiment. To the author's best knowledge, reports like these have not yet been published. The survey data indicated that a considerable portion of people listening to music while exercising experience some effects related to their sound perception. The majority of the reported effects were negative, ranging from discomfort to significant inconvenience or even pressure and pain.

Based on the survey data alone, it could be expected that for about a half of the participants of the laboratory experiment a noticeable change in listening preferences could be measured. In fact, seven out of the ten subjects expressed a behavior that may point towards a consistent shift in their preferences. In general, the present data indicate that music indeed serves as an important energizer in that all the subjects preferred an increase of either overall (frequency-independent) gain or low-frequency content of the signal when training began (S3PM1 compared to S1PM). The average self-adjusted overall level at the beginning of the exercise was very similar to the level determined in the first session (no exercise) as corresponding to a sensation of "loud" (79 vs. 81 dB SPL). For individual subjects, however, the listening level preferred during physical exercises (in all three phases tested: beginning, middle and end of exercises) differed considerably from the level corresponding to "loud" without physical exercise. The total change introduced ranged from a +15 dB boost of middle frequencies (in the 1/3 octave band with a mid-frequency of 1587 Hz, subject no. 4) to a -5 dB reduction in gain at high frequencies (in 1/3 octave bands with center frequencies >6 kHz, subject no. 6). Overall, there was no significant correlation ($R^2<0.1$, $p>0.05$) between the two sets of individual levels, indicating that the listening levels selected during exercises and the levels corresponding to "loud" music without exercises were similar only on average, but not on an individual basis.

With respect to frequency shaping subjects again showed markedly different preferences. Some subjects clearly preferred high-frequency gain (e.g., no. 1 and 8) and other clearly preferred low-frequency gain (no. 6 and 7). Such strong inter-individual differences in preference adjustment have also been found in previous studies using a similar user interface (Rennies et al., 2016). It is worth noting at this point that these individual trends were similar within subjects across adjustment sessions (apart from a possible influence of growing fatigue which is described below), even though the starting point of the frequency shaping settings as well as the axes arrangement and orientation were randomized for each trial. This means that subjects could not make use of a visual anchor to find their preferred parameters (which could have happened in Rennies et al., 2016). It is, therefore, reasonable to assume that a consistent individual pattern reflects individual listening preferences without any bias caused by the effects of the user interface design.

When fatigue began to increase after the initial adjustment at the onset of the exercise (S3PM2) the majority of participants introduced further level increases, yet some of them decided to decrease high-frequency content of the signal (>2000 Hz). During the last preference

adjustment measured at the highest degree of fatigue included in this study (S3PM3), seven out of ten subjects further reduced either overall gain or high-frequency content, while three subjects further increased the listening level. The group effect of this trend was statistically significant, despite large interindividual differences with respect to how fatigue appeared to affect listening preferences. One possible explanation for the differences between subjects is that it may take some time for a change in auditory perception to build up during exercise. This time may depend on individual fatigue, or it may be related to the individually perceived exertion rate. Hence, subjects with higher fatigue or higher perceived exertion rate could have introduced more pronounced changes in the audio material, here being expressed by the lowering of overall level of high-frequency content of the signal. Similarly, subjects whose perceived exertion rate is lower could still require a further increase in fatigue before a clear change in their preferred settings becomes obvious. However, the present data do not allow drawing firm conclusions at this point because individual fatigue and perceived exertion rates were not controlled. Further research is needed to better understand the mechanism underlying modified listening preferences during physical exercises. The duration was selected in this study to be comparable to previous studies on TTS, but it may have been too short to induce preference changes in all subjects. On a group level, the changes introduced by subjects to the playback within different fatigue stages of session 3 turned out as not significant, which may be expected since the survey data indicated that auditory changes were consciously perceived by only about half the participants. Nevertheless, it is interesting to note the parallels between survey data and laboratory experiments. Not only was the proportion of subjects to show a trend of changed preferences comparable to the proportion of participants in the survey to indicate auditory effects. In addition, the changes in listening preferences observed in the experiment consisted of a reduction in high-frequency content and/or overall level, which matched the most common action undertaken by the survey participants to counter the perceived auditory changes or hearing inconveniences. This compatibility could be a motivation for further studies into this phenomenon.

4.5.2. HEARING THRESHOLD MEASUREMENTS

The threshold measurements revealed that the measurement session had a significant effect on thresholds at one of the measured frequencies, but the effect size was very small and the significance of differences disappeared on the post-hoc test when correcting for multiple comparisons. The general observation was that the threshold data were dominated by large interindividual differences, which averaged out and resulted in minimal mean effects. On average subjects experienced between 1-2 dB higher threshold shift after being exposed to the music only (at their individual "loud" level, at 2, 4, and 6 kHz), than after being exposed to the music combined with sport exercises. At frequencies of 3 and 8 kHz the opposite trend was observed with mean subjective results indicating slightly higher thresholds after loud music exposure and sport exposure than after exposure to loud music only (see Fig. 4.5), with differences being < 2dB. In any case, both threshold measurement sessions S1AU2 and S2AU led to the same small trend of increase in thresholds in comparison to the reference test S1AU1. In summary, the present data are in contrast to results of (Vittitow et al., 1994), who used the same music exposure duration, in two ways. First, no significant TTS was observed, i.e., thresholds after 20 minutes of music exposure did not differ significantly from thresholds measured without prior music exposure. Second, there was no measurable impact of physical exercise on thresholds after music exposure, while (Vittitow et al., 1994) reported that for all frequencies (except for 2 kHz) - exercise with music produced on average 2-3 dB higher hearing thresholds than music exposure alone. One reason for the conflicting results could be the different subjects and methods employed. Vittitow (1994) tested younger subjects (mean age of 24.1 years), female participants only, and performed threshold measurements using the Hughson-Westlake-method (down 4 dB, up 2 dB with counterbalanced presentation order of frequencies). It is, however, not obvious why a different threshold measurement method should result in different trends. It is more likely that the considerable level difference between both studies is the main contributor to the differences in observed TTS. In (Vittitow et al., 1994), subjects could not individually choose the level of acoustic stimuli and were exposed to music at a level of 94.4 dB A, while in the current experiment subjects were allowed to choose the level corresponding to their individual "loud" sensation. This self-adjustment was introduced to avoid uncomfortable loudness for the subjects. The resulting levels ranged from 66 to 90 dB SPL and averaged to 81 dB SPL. Likewise, the RMS levels of self-adjusted songs (gain and

EQ) in the third session varied considerably between subjects (64-90 dB SPL), and averaged to 79 dB SPL. Hence, considerably lower levels were employed in the present study. It is possible that TTS would have occurred in the present study if higher exposure levels had been employed.

It was reported previously (Vittitow et at., 1994, Hutchinson et al. 1991, Lindgren and Axelsson, 1988) that exercises alone could slightly lower hearing thresholds. In the current study, such a comparable trend was observed for individual subjects and frequencies when comparing thresholds after listening to loud music alone and after loud music accompanied by exercises. However, this was not statistically significant. As stated above, it is possible that larger differences could have occurred if higher music levels had been employed. Further research is needed to verify if an increased TTS due to physical exercise occurs at higher listening levels, longer exercise and exposure duration, or both.

4.6. CONCLUSIONS

The present study investigated possible changes in listening preferences and hearing thresholds for amateur athletes while exercising to music. The following conclusions can be drawn:

- Survey data indicated that physical exercises affect sound perception in a considerable portion of participants (72 out of 138). The reported effects ranged from preference-related effects (expressed as changing songs, volume, or frequency shape of the playback) to discomfort and hearing inconvenience (e.g., abandoning listening to music, perceiving pressure or pain). This may indicate that listening comfort could be enhanced if music playback devices were capable of monitoring the listener's exercise status and performing the individually preferred changes in audio settings automatically.
- In a controlled laboratory experiment, all subjects preferred louder music during sport exercises than in resting condition. Additionally, a significant trend for increased sensibility to high-level and/or high-frequency sounds was observed in relation to growing fatigue, although the degree of this trend differed strongly across subjects and was found to be pronounced for seven out of ten subjects. This proportion was in line with the survey data indicating high-frequency sounds as bothering for fatigued subjects in 37% of cases.

- The potentially dangerous impact of sport exercises when accompanied with loud music on hearing thresholds reported in previous studies was not found, possibly due to personalized and considerably lower signal levels employed here. In the light of the current study and the available literature, further research into the topic is needed to assess the dependence between the acoustic signal level and exposure duration, physical fatigue, and hearing thresholds to ensure safety and sufficient hearing protection.

5. CONCLUSION AND FURTHER RESEARCH

In this chapter the main aspects of the thesis are summarized and an outlook for further research is provided.

5.1. CONCLUSIONS

The subject of the thesis is basic psychoacoustic research on personalization of audio playback to individual hearing abilities and listening preferences of normal-hearing subjects and those with hearing loss. Connecting user abilities with preferences has the potential to improve the accuracy of personalization algorithms and the living conditions of those with not yet treated hearing loss that could easily adjust the playback on their electronic devices (TVs, mobile phones, etc.) to their needs and comfort. The research scenarios addressed in this thesis included subjective speech recognition abilities and preferred listening settings for speech signals on the background of different maskers and their spatial constellations as well as an investigation of how hearing thresholds and listening preferences are affected by physical strain.

General conclusions from this thesis are:

- A better understanding is achieved of which personal factors play a decisive role in predicting subjective speech recognition threshold performance under different masker sources and their spatial placement, and also how this performance can be predicted by a binaural speech intelligibility model for groups of subjects spread in age, PTA and speech recognition abilities,
- A better understanding of supra-threshold listening preferences is achieved for basic signal processing (gain, gain accompanied with distortions caused by clipping or compression, and frequency-shaping) in complex listening conditions and preferences' relation to the abovementioned personal factors and threshold hearing performance. Listening preferences in complex listening conditions can be predicted based on SRT performance, which also finds application in the field of speech enhancement and hearing support technologies.

- A novel method of profiling subjects along the perceptual dimension of noise-vs.-distortion sensitivity is presented. This subjective trait proved to be stable over time and across conditions and together with the measurement method developed may find its way into hearing-support technologies or clinical applications.
- Proof-of-concept data was gained on whether listening preferences change under physical fatigue (sport). This had not been fully explored before – similar to hearing threshold changes in relation to growing fatigue where virtually no data existed. Questionnaire data suggest a relation between perception and growing fatigue. The laboratory study found corresponding effects indicating a common shift in listening preferences when a certain fatigue level is reached. Future research based on these results should develop a playback automatization for subjects undergoing physical strain in order to provide hearing protection and increase comfort.

In chapter 2 the investigation into speech recognition performance in complex listening scenarios has shown how certain personal factors can explain subjective performance in complex acoustic listening conditions. The group of subjects tested was chosen to be highly diversified in age, audiogram and the mentioned standard diotic speech-in-noise test results in order to test the possible impact of a wide range of subjective factors on speech recognition performance. Factors that best explained the performance of the whole group were PTA, SRT and Goesa results. This indicates that speech intelligibility tests employing more complex material (e.g., everyday sentences like in case of Goesa) is more meaningful for predicting speech perception in complex scenarios than tests employing simpler meaningless speech material (like in case of DTT), even if both are measured monaurally. Age had the lowest predictive power for SRTs measured for the MT masker (in all spatial constellations) and for TT masker in the separated masker constellation. At the same time, age showed the highest correlations with SRTs measured for the TT co-located condition, i.e., the one with presumably the highest degree of IM - so the highest demand for cognitive processing to segregate the target speech from the maskers.

Participants with good baseline performance (low SRT) benefited more from spatial unmasking (separated maskers), strong masker envelope modulations (TT masker), and release from IM. Those who did not benefit from the abovementioned factors had the poorest baseline SRTs and had either relatively high PTA (but not too advanced age) or close-to-normal PTA but advanced age (81-86 years), providing evidence that both age and hearing loss may exert

an effect on SRT. The results showed that for a subgroup of subjects, age was a factor stronger related to speech recognition performance than PTA. These subjects had close-to-normal hearing at their better ear, yet their SRT performance (as well as SRM, especially for TT maskers) was much poorer than that of normal-hearing young subjects. Their SRT performance placed them among the results of other subjects similar in age but with considerably higher hearing loss. In general, both average and individual performance of the oldest group of subjects revealed decreasing performance with increasing age, which was not consistently observed for increasing hearing loss (PTA). This indicates that the older "close-to-normal-hearing" subjects were much stronger affected by IM and could not obtain a spatial benefit like their younger counterparts.

In order to assess how well interindividual variability could be predicted by a state-of-the-art speech intelligibility model, a BSIM model was used due to its capability to predict the influence of the factors involved in the experimental conditions of the current study, such as the benefit from masker amplitude modulation (also known as "dip listening"), spatial unmasking (as well as the combination of both factors). Such accurate predictions based on a limited set of subjective measures are important information for diagnostic purposes, hearing aid fitting and individualized speech enhancement strategies and offer the possibility to model speech perception in a given condition for a diversified set of subjects without the need of costly and lengthy listening tests. This thesis states that the biggest improvements in the model's predictive power was achieved by using individualized reference SII values (matched to one of the experimental conditions for each subject) and individual audiograms instead of only individualization cues. The model provided accurate predictions for the majority of experimental conditions, although larger discrepancies between experimental data and predictions remained for conditions with a high degree of IM - e.g., TT masker co-located or SRM for TT maskers. Nevertheless, testing the model on a diversified group of subjects in complex listening conditions showed the model's applicability for clinical purposes in conditions involving both energetic masking and spatial unmasking.

Chapter 2 provides an investigation on what personal factors explain subjective performance in complex listening conditions, and what would be the minimum set of such factors needed to model subjective performance at the level of SRT. Question addressed within Chapter 3 was how these personal factors and subjective performance measured in Chapter 2 (SRTs) can influence or define supra-threshold listening preferences tested with the same

groups of subjects, in the same listening conditions with various signal processing strategies involved. This study was designed to measure individually preferred playback settings along specific perceptual dimensions - namely loudness rating (linear gain scenario), noise-vs.-distortion trade off (gain at the cost of clipping or compression distortions) and frequency-shaping. The data was intended to serve as a basis for creating individual profiles (also referred to as "personal traits") that could be used for personalization of hearing devices and other hearing support technologies. This approach relies on the assumption that individual listening preferences are stable over time in similar listening conditions. A focus was put on the noise-vs.-distortion trade-off – by including scenarios of gain at the cost of clipping or compression distortions – since the existence of such trade-off was previously reported by several studies (Völker et al, 2018, Marzinzik, 2000) and is a typical problem in hearing-aid fitting related to, e.g., the strength of noise reduction.

The study was divided into two parts. During the first part, subjects performed sound adjustments only on the speech (target) signal, while the respective masker remained the same. In the second part, the adjustments were made for the combined speech-masker mixture, i.e., with constant SNR. This diversification was made in order to mimic near-end and far-end processing where either only speech or both speech and noise were processed. This experimental design was motivated by not yet well understood mechanisms underlying preference judgement on the individual basis, or how such preferences vary in complex listening conditions (offering binaural unmasking and dip listening and involving challenges such as energetic masking and IM). The conclusion of this chapter was a general high test-retest reliability of adjustments for all modification's schemes (except for frequency-shaping) throughout the conditions and for both parts of the study (adjustable SNR vs. fixed SNR). Three consistent categories of subjects were observed, that formed along the noise-vs.-distortion dimension - "noise haters" and "distortion haters" and "indifferent". This individual preference trait remained stable throughout all the test conditions and over time (test/retest sessions). This provided a direct evidence that classifying subjects along this perceptual dimension can be a valid approach and has the potential to be used in audiological applications. The comparison of the spatial benefit obtained in chapter 2 with the individually preferred sound adjustments indicates that the binaural unmasking effect (SRT difference between co-located and spatial maskers) was transferred to supra-threshold listening conditions. In other words, when maskers were spatially separated, subjects adjusted the speech level to be softer (i.e., the SNR to be

lower) than when maskers were co-located. This is in line with recent findings that illustrate a binaural release from listening effort also in conditions where intelligibility is at ceiling (Rennies and Kidd, 2018). Furthermore, it was found that SRTs measured in complex listening conditions were reasonably good predictors for listening preferences in the same conditions (R^2 between 0.60 and 0.70), better than any other individual measures used in the study.

The method of adjustment used in order to determine personal traits for gain and gain accompanied with distortions (compression or clipping) turned out to be very quick and easy to administer for both researchers and subjects (especially in comparison to e.g. paired-comparisons). Additionally, the range of adjustments was chosen to be comparatively wide (30 dB) and applicable also for negative SNRs – where many noise reduction schemes fail to operate as designed. If a consistent link between the personal traits measure proposed in this study and preferences for noise reduction strength in an end product (e.g., hearing aids) can be established in future work - the proposed method can be successfully used to easily classify subjects on the noise-vs.-distortion perceptual dimension, letting them self-express their needs in a very fast and playful way. At the same time, providing an easy access to preference estimation of this subjective trait would be a valuable information for advanced tuning of hearing systems and could significantly reduce e.g., the "try and error" phase of hearing aids fitting (Jenstad et al., 2003).

Investigations described in the chapters 2 and 3 concentrated on finding the relation between personal factors, hearing abilities (e.g., SRT performance) and listening preferences for speech processing and succeeded in finding a "personal trait" along the noise-vs.-distortion perceptual dimension as well as described age as a highly relevant factor related to speech recognition performance in the presence of informational masking. The presence of such stable traits in scenarios involving everyday listening conditions raised another question whether other such scenarios exist where preferences could be tracked and even modelled. These scenarios needed to be specific and possibly involve other (or additional) personal factors to help in profiling. Next to speech communication, another common condition that would involve a completely new additional set of personal factors is a scenario of listening to music while doing sport, due to involving rapid physiological changes caused by physical strain. Are such traits also stable throughout different physiological conditions – e.g., different stages of physical arousal from onset to maximal strain – or maybe they change alongside? This led to a more general question formulation, such as: if and how does physical fatigue change listening

abilities (e.g. on threshold levels) or influence listening preferences? There was very little to no scientific reports on these topics (to the author's best knowledge - only three, and contradicting). This led to the development of an online survey study on subjective listening preferences (for music) under physical strain. Due to its interesting, consistent results this preliminary study was later followed by a lab-based study.

In chapter 4, the concept of individual sound preferences and possible changes in hearing thresholds due to physical fatigue was investigated in a specific scenario, which had been motivated by informally reported hearing phenomena, and online questionnaire results. Specifically, the influence of physical fatigue on hearing thresholds as well as on supra-threshold listening preferences regarding audio playback was explored. As mentioned above, one source of the motivation for the current study were previous findings reporting possible bigger TTS after exercising to music than after being exposed to loud music alone (Vittitow et al., 1994), which was in line with the preliminary survey data obtained here. Another motivation was the possibility to further investigate "personal traits" existence in different listening scenario. Based on survey data, it was hypothesised that a sport scenario can reveal a consistent shift in preferences, similar to one described in chapter 3, but rather along the "loudness" dimension– e.g., "loudness hater" vs. "power junkie" (Marzinzik, 2000) or to be related to a specific frequency shaping (annoyance for high-frequency sounds). Another important aspect was an attempt to fulfil a complete lack of data – as mentioned above – on how listening preferences change during physical exercises as well as very limited and contradicting findings about how and if hearing thresholds are affected by listening to music during growing physical fatigue. Possible changes in perception were first investigated via an online survey, followed by an experiment in controlled laboratory settings with a group of ten normal-hearing amateur athletes, who could adjust the music playback at different states of a defined physical exercise (cycling on an ergometer). In general, the results of both the questionnaire and the experiment were consistent in that about half of the subjects either reported (questionnaire) or showed (experiments) changes in how they preferred music to be processed over the course of their exercise. The questionnaire data indicated that physical arousal affected sound perception of a considerable portion of participants (72 out of 138) and that 81% of them introduced some changes to the playback during sport (which indicates an apparent need for automation of this process). The reported effects ranged from preference change (in relation to volume, EQ, tempo or content of music material) to discomfort and hearing inconveniences either during or soon

after the exercises (such as turning music off, feeling pain or pressure in the ears). The results of the controlled laboratory experiment showed a preference for louder music during sport exercises than in resting conditions for all subjects. Moreover, the data suggested an increase in preferred volume settings until a certain level of fatigue is achieved, as the majority of subjects tended to initially increase and then decrease the overall volume and/or the high-frequency content of the music towards the end of the exercise. Only three participants (possibly the ones with the highest endurance, based on maximum oxygen uptake tests) did not make such a reduction towards the end of their exercise. In general, a significant trend for increased sensibility to high-frequency or high-level sounds was observed. Possible higher TTS in scenarios where exercises were accompanied with loud music than in scenarios where subjects were exposed to loud music alone was not confirmed – possibly due to the individualized (to one's "loud" sensation) and, as a result, lower playback levels used in the study, but this issue requires further research. While the physiological as well as the perceptual reasons underlying these individual effects remain unclear at this point, these data indicate that a better understanding of auditory perception under physical strain may be relevant, e.g., to adjusting playback settings in relation to physical state of the user and provide hearing protection (if needed) or signal personalization leading to increased comfort of athletes. The results obtained (both experimental as well as survey data) indicate a consistent shift in personal preferences ("traits") successfully proving the experimental hypothesis.

5.2. OUTLOOK AND FURTHER RESEARCH

The thesis aimed at gaining knowledge on the influences that various personal factors may have on both speech recognition performance as well as listening preferences (regarding simple signal modifications) of both normal-hearing and hearing-impaired subjects. Each of the studies brought outcomes that can be enhanced by further clinical and applied research. Experimental results of the study presented in chapter 2 indicated that the group of older subjects exhibited considerably poorer performance in segregating target speech from maskers compared to their younger counterparts with similar hearing abilities. This finding showed potential for further research on cognitive processes related to speech perception in complex listening scenarios and their possible age-related decline. Additionally, research directed on incorporating cognitive and age-related measures into speech intelligibility models could result in improving their accuracy. Finding the connection between age and poor speech recognition performance raised the question whether these or other personal factors could also lead to a diversification of subjects at supra-threshold levels, regarding their listening preferences.

Experimental results described in chapter 3 proved the stability of listening preferences in certain speech processing techniques, throughout different listening conditions and test/retest sessions, and let for describing a stable preference profile along the perceptual dimension of noise tolerance (offering a method of adjustment to elicit such preference). Further research is needed to confirm such preferences' stability - that let for subjects' classification either as "noise hater", "distortion hater" of "indifferent" - in various listening conditions and on a broader group of subjects. Additional investigation on the method proposed is needed too, in order to test its clinical applicability and generalization towards different kinds of speech enhancement schemes.

The observed stability of listening preferences for speech processing throughout different listening conditions led to the question of whether such stability of judgments or a consistent shift of preferences could also be observed in very different listening situations, involving physical exercises. The study presented in chapter 4 can be considered as a first step into the "sport-audio" field. Further research, among a bigger group of normal-hearing participants, is needed to confirm the consistent preference shift observed. At the same time, there is a complete lack of scientific reports – again, to author's best knowledge – on listening

preferences of hearing-impaired subjects in the same conditions. Moreover, the results obtained from the current study did not confirm a higher TTS occurring when exercising to loud music then TTS after being exposed to the loud music only - this alone requires further examination as the levels used in the current study were self-adjusted to subjective “loud” sensation level, yet much lower than ones in the previous reports. It is also important to mention that all the studies found on the topic targeted amateur athletes and no data on professional sportsmen and sportswomen regarding their listening preferences were found. Continuing research in this new “sport-audio” field may lead to increasing comfort and safety of both amateur and professional athletes. Gaining a better understanding of the underlying physiological effects of physical strain and their implication for threshold hearing and supra-threshold listening preferences could lead to automatization of sport playback devices or the development of specific hearing aid programs.

The findings presented in this thesis show that individual listening preferences are measurable in a reliable way, that they can be stable both over time and across different listening conditions and are (most likely) rooted in individual health- and age-related factors, as well as in the physiological state of a person. Creation of audio consumer profiles based on such individual factors seems possible yet requires further research.

BIBLIOGRAPHY

Anshel, M., & Marisi, D. (1978). Effects of music and rhythm on physical performance. *Research Quarterly, 49*, 109-113.

ANSI 1969, Methods for the calculation of the articulation index. American National Standards Institute, New York.

ANSI, 1997, Methods for Calculation of the Speech Intelligibility Index, ANSI/ASA S3.5-1997 (R2017).

Arbogast, T., Mason, C., & Kidd Jr, G. (2005). The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am., 117*(4), 2169-2180.

Arbogast, T., Mason, C., & Kidd Jr., G. (2002). The effect of spatial separation on informational and energetic masking of speech. *J. Acoust. Soc. Am., 112*(5), 2086-2098.

Best, V., Marrone, N., Mason, C., & Kidd Jr., G. (2012). The influence of non-spatial factors on measures of spatial release from masking. *J. Acoust. Soc. Am., 131*, 3103-3110.

Beutelmann, R., & Brand, T. (2006). Prediction of speech intelligibility in spatial noise and reverberation for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am., 120*, 331-342.

Beutelmann, R., Brand, T., & Kollmeier, B. (2010). Revision, extension, and evaluation of a binaural speech intelligibility model. *J. Acoust. Soc. Am., 127*(4), 2479–2497.

Biagini, M., Brown, L., Coburn, J., Judelson, D., Statler, T., Bottaro, M., Longo, N. (2012). Effects of self-selected music on strength, explosiveness, and mood. *J. Strength. Cond. Res., 26*, 1934-1938.

Boldt, F., Berbalk, A., Halle, M., Hoffmann, G., Löllgen, H., Schmidt, A., & Trucksäß, M. (2002). Leitlinien zur Belastungsuntersuchung in der Sportmedizin. *Expertenkommission und Präsidium der Deutschen Gesellschaft für Sportmedizin und Prävention. Aktueller Stand: 03/2002*.

Brand, T., & Hohmann, V. (2001). Effects of hearing loss, centre frequency, and bandwidth on the shape of loudness functions in categorical loudness scaling. . *Audiology, 40*(2), 92-103.

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J. Acoust. Soc. Am., 111*(6), 2801-2810.

Brand, T., Hauth, C., Wagener, K., & Neher, T. (2017). Predicting the benefit of binaural cue preservation in bilateral directional processing schemes for listeners with impaired hearing. *Proceedings of the International Symposium on Auditory and Audiological Research.* Nyborg: The Danavox Jubilee Foundation.

Bronkhorst, A. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Atten. Percept. Psychophys.*

Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *J. Acoust. Soc. Am., 83*(4), 1508-1516.

Bronkhorst, A., & Plomp, R. (1992). Effect of multiple speech like maskers on binaural speech recognition in normal and impaired hearing. *J. Acoust. Soc. Am., 92*(6), 3132-3139.

Brons, I., Houben, R., & Dreschler, W. (2012). Perceptual effects of noise reduction by time-frequency masking of noisy speech. *J. Acoust. Soc. Am., 132*(4), 2690–2699.

Brons, I., Houben, R., & Dreschler, W. (2013). Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort. *Ear Hear, 34*, 29-41.

Brons, I., Houben, R., & Dreschler, W. (2014). Effects of Noise Reduction on Speech Intelligibility, Perceived Listening Effort, and Personal Preference in Hearing-Impaired Listeners. *Trends Hear*, 1-10.

Carhart, R. T. (1968). Effects of Interaural Time Delays on Masking by Two Competing Signals. *The Journals of the Acoustical Society of America, 43*(6), 1223-1230.

Chabot-Leclerc, A., MacDonald, E., & Dau, T. (2016). Predicting binaural speech intelligibility using the signal-to-noise ratio in the envelope power spectrum domain. *J. Acoust. Soc. Am., 136*, 192-205.

Chen, Y.-A., Wang, J.-C., Yang, Y.-H., & Chen, H.-H. (2017). Component Tying for Mixture Model Adaptation in Personalization of Music Emotion Recognition. *IEEE/ACM Trans. Audio, Speech, Language Process, 25*(7), 1409-1420.

Cherry, E. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am., 25*, 975-979.

Chung, T., Rust, R., & Wedel, M. (2009). My Mobile Music: An Adaptive Personalization System for Digital Audio Players. *Marketing Science, 28*(1), 52-68.

Ciba, S., Volgenandt, A., Bruns, T., Asendorf, R., Oetting, D., & Rennies, J. (2014). Evaluation of interfaces for the self-fitting of personalized communication systems by hearing-impaired users. *Deutsche Jahrestagung für Akustik* (pp. 68-69). Oldenburg: Deutsche Gesellschaft für Audiologie .

Clark, W., & Bohne, B. (1999). Effects of noise on hearing. *Journal of the American Medical Association, 281,* 1658-1659.

Cord, M., Walden, B., Surr, R., & Dittberner, A. (2007). Field Evaluation of an Asymmetric Directional Microphone Fitting. *J Am Acad Audiol, 18*, 245-256.

Culling, J. S. (1995). Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *J. Acoust. Soc. Am., 98*(2), 785-797.

Culling, J., Hawley, M., & Litovsky, R. (2004). The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *J. Acoust. Soc. Am., 116*(2), 1057-1065.

Dau, T., Püschel, D., & Kohlrausch, A. (1996). A quantitative model of the ‘‘effective’’ signal processing in the auditory system. I. Model structure. *J. Acoust. Soc. Am., 99*(6), 3615-3622.

Dau, T., Püschel, D., & Kohlrausch, A. (1996). A quantitative model of the ‘‘effective’’ signal processing in the auditory system. II. Simulations and measurements. *J Acoust Soc Am, 99*(6), 3623-3631.

Dau, T., Verhey, J., & Kohlrausch, A. (1999). Intrinsic envelope fluctuations and modulation-detection thresholds for narrowband noise carriers. *J. Acoust. Soc. Am., 106*, 2752–2760.

Dawson, P., Dillon, H., & Battaglia, J. (1990). Output limiting compression for the severe-profoundly deaf. . *Aust. J. Audiol., 13*, 1-12.

De Andrade, K., De Lemos Menezes, P., Carnaúba, A., De Sousa Rodrigues, R., De Carvalho Leal, M., & Pereira, L. (2013). Non-flat audiograms in sensorineural hearing loss and speech perception. *Clinics, 68*(6), 815-819.

Derleth, R., Dau, T., & Kollmeier, B. (2001). Modeling temporal and compressive properties of the normal and. *Hear. Res., 159*, 132-149.

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *J. Acoust. Soc. Am., 97*, 585–592.

Drullman, R., & Bronkhorst, A. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *J. Acoust. Soc. Am., 107*, 2224-2235.

Dubbelboer, F., & Houtgast, T. (2008). The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. *J. Acoust. Soc. Am., 124*(6), 3937–3946.

Dubno, J., Dirks, D. D., & Morgan, D. E. (1984). Effects of age and mild hearing loss on speech recognition in noise. *J. Acoust. Soc. Am., 76*(1), 87-96.

Duquesnoy, A. (1983). Effect of a single interefering noise or speech source upon the binaural sentence intelligibility of aged persons. *J. Acoust. Soc. Am., 74*(3), 739-743.

Durlach, N. (1963). Equalization and cancellation theory of binaural masking-level differences. *J. Acoust. Soc. Am., 35*(8), 1206–1218.

Durlah, N. (1963). Equalization and cancellation theory of binaural masking-level differences. *J Acoust Soc Am, 35*, 1206-1218.

Durlah, N. (1972). Binaural signal detection: Equalization and cancellation theory. In J. Tobias, *Foundations of Modern Auditory Theory* (pp. 371-462). New York: Academic.

Ellinger, R., Jakien, K., & Gallun, F. (2017). The role of interaural differences on speech intelligibility in complex multi-talker environments. *J. Acoust. Soc. Am., 141*(2).

EuroTrak Germany, Anovum. (2018), https://www.ehima.com/wp-content/uploads/2018/06/

EuroTrak_2018_GERMANY.pdf.

Ewert, S., & Dau, T. (2000). Characterizing frequency selectivity for envelope fluctuations. *J. Acoust. Soc. Am., 108*(3), 1181–1196.

Ewert, S., Schubotz, W., Brand, T., & Kollmeier, B. (2017). Binaural masking release in symmetric listening conditions with spectro-temporally modulated maskers. *J. Acoust. Soc. Am., 142*(1), 12-28.

Ezzatian, P., Li, L., & Pichora-Fuller, K. (2011). The effect of priming on release from informational masking is equivalent for younger adn older adults. *Ear Hear, 32*, 84-96.

Füllgrabe, C., Moore, B., & Stone, M. A. (2015). Age-group differences in speech indentification despite matched audiometrically normla hearing: contributions from auditory temporal processing and cognition. *Front Aging Neurosci, 6*, Article 347.

Fastl, H., & Zwicker, E. (2007). *Psychoacoustics Facts and Models.* Berlin: Springer.

Festen, J., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am., 88*(4), 1725-1736.

Fletcher, H. (1940). Auditory patterns. *Rev. Mod. Phys., 12*, 47–65.

Frank, J., & Massey, J. (1951). The Kolmogorov-Smirnov Test of Goodness of Fit. *Journal of the American Statistical Association, 46*(253), 68-78.

French, N., & Steinberg, J. (1947). Factors governing the intelligibility of speech sounds. *J. Acoust. Soc. Am, 19*(1), 90– 119.

Freyman, R., Balakrishnan, U., & Helfer, K. (2001). Spatial release from informational masking in speech recognition. *J. Acoust. Soc. Am., 109*(5), 2112-2122.

Freyman, R., Balakrishnan, U., & Helfer, K. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *J. Acoust. Soc. Am., 115*(5), 2246-2256.

Gelfand, S., Ross, L., & Miller, S. (1988). Sentence reception in noise from one versus two sources: Effects of aging and hearing loss. *J. Acoust. Soc. Am., 83*(1), 248-256.

George, E. L., Goverts, S. T., Festen, J. M., & Houtgast, T. (2010). Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners. *J. Speech Lang. Hear. Res., 53*, 1429-1439.

Goverts, S. T., & Houtgast, T. (2010). The binaural intelligibility level differences in hearing-impaired listeners: The role of supra-threshold deficits. *J. Acoust. Soc. Am., 127*(5), 3073-3084.

Hauth, C., Ewert, S., & Brand, T. (2017). Blind modeling of binaural unmasking of speech in stationary maskers. *J. Acoust. Soc. Am. 141*(5), 3969-3970.

Hawkins Jr., J. (1950). The masking of pure tones and of speech by white noise. *J. Acoust. Soc. Am., 22*(1), 6-13.

Hawkins, D., & Naidoo, S. (1993). Comparison od sound quality and clarity with asymmetrical peak clipping and output limiting compression. *J. Am. Acad. Audiol., 4*, 221-228.

Hawley, M. (2000). Speech intelligibility, localization adn binaural hearing in listenters with normal and impaired hearing. *Ph. D. dissertation, Biomedical Engineering, Boston University, Boston, MA.*

Helfer, K., & Freyman, R. (2008). Aging and speech-on-speech masking. *Ear&Hearing, 29*, 87-98.

Helfer, K., & Freyman, R. (2014). Stimulus and listener factors affecting age-related changes in competing speech perception. *J. Acoust. Soc. Am., 136*(2), 748-759.

Hochmuth, S., Jürgens, T., Brand, T., & Kollmeier, B. (2015). Talker- and language-specific effects on speech intelligibility in noise assessed with bilingual talkers: Which language is more robust against noise and reverberation? *Int J Audiol, 54*, 23-34.

Hohmann, V., & Kollmeier, B. (1995). The effect of multichannel dynamic compression on speech intelligibility. *J. Acoust. Soc. Am., 97*, 1191–1195.

Hopkins, K., & Moore, B. (2011). The effects of age and cochlear hearing loss on temporal fine structure sensitivity, frequency selectivity, and speech reception in noise. *J. Acoust. Soc. Am., 130*(1), 334–349.

Houtgast, T., Steeneken, H., & Plomp, R. (1980). Predicting speech intel- ligibility in rooms from the modulation transfer function. I. general room acoustics. *Acoustica, 46*, 60–72.

Huber, R., & Kollmeier, B. (2006). PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception. *IEEE Trans Audio Speech Lang Process, 14*, 1902–1911.

Humes, L. (2002). Factors underlying the speech-recognition performance of elderly hearing-aid wearers. *J Acoust Soc Am, 112*(3), 1112-1132.

Humes, L., & Coughlin, M. (2009). Aided speech-identification performance in single-talker competition by older adults with impaired hearing. *Scand J Psychol, 50*, 485-494.

Humes, L., Lee, J., & Coughlin, M. (2006). Auditory measures of selective and divided attention in young and older adults using single-talker competition. *J. Acoust. Soc. Am., 120*(5), 2926-2937.

Hutchinson, K., Alessio, H., Spadafore, M., & Adair, R. (1991). Effects of low-intensity exercise and noise exposure on temporary threshold shift. *Scand Audiol, 20*, 121-127.

IEC (2003) *Sound System Equipment - Part 16: Objective rating of speech intelligibility by speech transmission index, Interantional Standard IEC 60268-16 (International Electrotechnical Commission)*.

International Standard Association. International Standard Association, ISO 389-1:2017, Acoustics - Reference zero for the calibration of audiometric equipment - Part 1: Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones, (12/2017).

ITU-R BS.1116-3, Methods for the subjective assessment of small impairments in audio systems, (02/2015).

ITU-R BS.1387-1, Method for objective measurements of perceived audio quality, (1998-2001).

ITU-R BS.1534-3, Method for the subjective assessment of intermediate quality level of audio systems, (10/2015).

ITU-T P.861, SERIES P: TELEPHONE TRANSMISSION QUALITY, Objective quality measurement of telephone- band (300 - 3400 Hz) speech codecs, (08/96).

ITU-T, SERIES P: TELEPHONE TRANSMISSION QUALITY, Methods for objective and subjective assessment of transmission quality, P.800 (08/96).

Jørgensen, S., & Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *J. Acoust. Soc. Am., 130*(3), 1475–1487.

Jansen, S., Koning, R., Wouters, J., & van Wieringen, A. (2014). Development and validation off the Leuven intelligibility sentence test with male speaker (LIST-m). *Int J Audiol*(53), 55-59.

Jansen, S., Luts, H., Dejonckere, P., van Wieringen, A., & Wouters, J. (2014). Exploring the sensitivity of speech-in-noise tests for noise-induced hearing loss. *Int. J. Audiol., 53*, 199-205.

Jenstad, L., Van Tasell, D., & Ewert, C. (2003). Hearing aid trouble-shooting based on patents' descriptions. *J Am Acad Audiol, 14*, 347-360.

Jia , T., Ogawa, Y., Miura, M., Ito, O., & Kohzuki, M. (2016). Music Attenuated a Decrease in Parasympathetic Nervous System Activity after Exercise. *PLOS ONE, 11*, 1-12.

Kayser, H., Ewert, S., Anemüller, J., Rohdenburg, T., Hohmann, V., & Kollmeier, B. (2009). Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP J Adv Signal Process, 2009*, Article ID 298605.

Keidser, G., Dillon, H., Flax , M., Ching, T., & Brewer, S. (2011). The NAL-NL2 Prescription Procedure. *Audiol Res, 1:e24*(1), 88-90.

Kidd Jr, G., Mason, C. R., Rohtla, T. L., & Deliwala, P. S. (1998). Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. *J. Acoust. Soc. Am., 140*(1), 422-431.

Kidd Jr., G., & Colburn, H.S. (2017). Informational Masking in Speech Recognition. In J. Middlebrooks, J. Simon, A. Popper, & R. Fay, *The Auditory System at the Cocktail Party* (pp. 75-109). Springer International Publishing.

Kidd Jr., G., Mason, C., Swaminathan, J., Roverud, E., Clayton , K., & Best, V. (2016). Determining the energetic and informational components of speech-on-speech masking. *J. Acoust. Soc. Am., 140*(1), 132-144.

Kießling, J., Schubert, M., & Archut, A. (1996). Adaptive fitting of hearing instruments by category loudness scaling (ScalAdapt). *Scand. Audiol., 25*, 153-160.

Kim, B., & Oh, S. (2013). Age-related changes in cognitioin and speech perception. *Koren J Audiol, 17*, 54-58.

Kjems, U., Boldt, J., Pedersen, M., Lunner, T., & Wang, D. (2009). Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *J. Acoust. Soc. Am., 126*(3), 1415–1426.

Kollmeier, B., & Kiessling, J. (2018). Functionality of hearing aids: state-of-the-art and future model-based solutions. *Int J Audiol, 57*(3), 3-28.

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

Kollmeier, B., Schädler, M., Warzybok, A., Meyer, B., & Brand, T. (2016). Sentence Recognition Prediction for Hearing-impaired Listeners in Stationary and Fluctuation Noise With FADE: Empowering the Attenuation and Distortion Concept by Plomp With a Quantitative Processing Model. *Trends in Hearing, 20*, 1-17.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M., Uslar, V., Brand, T., & Wagener, K. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *Int J Audiol, 54*, 3-16.

Kowalk, U., Kissner, S., von Gablenz, P., Holube, I., & Bitzer, J. (2017). An improved privacy-aware system for objective and subjective ecological momentary assessment. *International Symposium on Auditory and Audiological Research (Proc. ISAAR)* (pp. 25-32). Nyborg, Denmark: The Danavox Jubilee Foundation.

Kryter, K. (1962). Methods for the calculation and use of the articulation index. . *J. Acoust. Soc. Am., 34*(11), 1689–1697.

Kubiak, A.M., Rennies, J., Ewert, S.D., Kollmeier, B. (2020). Prediction of individual speech recognition performance in complex listening conditions. *J. Acoust. Soc. Am., 147*(3), 1379–1391.

Kubiak, A., Saft, M., Rennies, J., & Kollmeier, B. (2014). Listener's preferences with respect to frequency shaping and loudness adjustments for music and speech reproduction. *40. Jahrestagung für Akustik*, (pp. 495-496). 10.-13. März 2014 in Oldenburg.

Kucirkova, N., & Flewitt, R. (2018). The future-gazing potential of digital personalization in young children's reading: views from education professionals and app designers. *Early Child Dev Care*, 1-15.

Lavandier, M., & Culling, J. (2010). Prediction of binaural speech intelligibility against noise in rooms. *J. Acoust. Soc. Am., 127*, 387-399.

Lavandier, M., Jelfs, S., Culling, J., Watkins, A., Raimond, A., & Makin, S. (2012). Binaural prediction of speech intelligibility in reverberant rooms with multiple noise sources. *J. Acoust. Soc. Am., 131*, 218-231.

Lebo, C., & Reddell, R. (1972). The presbycusis component in occupational hearing loss. *Laryngoscope, 82*, 1399-1409.

Leclère , T., Lavandier, M., & Culling, J. (2015). Speech intelligibility prediction in reverberation: Towards an integrated model of speech transmission, spatial unmasking, and binaural de-reverberation. *J. Acoust. Soc. Am., 137*(6), 3335-3345.

Lecluyse, W., & Meddis, R. (2009). A simple single-interval adaptive procedure for estimating thresholds in normal and impaired listeners. *J. Acoust. Soc. Am., 126*(5), 2570-2579.

Lee, J., & Humes, L. (2012). Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background. *J. Acoust. Soc. Am., 132*(3), 1700–1717.

Levitt, H. R. (1967). Binaural Release From Masking for Speech and Gain in Intelligibility. *J. Acoust. Soc. Am., 42*(3), 601-608.

Licklider, J. (1948). Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. *J. Acoust. Soc. Am., 20*(1), 42-51.

Lindgren, F., & Axelsson, A. (1988). The influence of physical exercise on susceptibility to noise-induced temporary threshold shift. *Scand. Audiol., 17*, 11-17.

Lloyd, S. (1982). Least Squares Quantization in PCM. *IEEE Transactions on information theory, 28*(2), 129-137.

Ludvigsen, C., Elberling, C., & Keidser, G. (1993). Evaluation of a noise reduction method - comparison between observed scores and scores predicted from STI. *Scand. Audiol., 38,* 50-55.

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Buechler, M., Spriet, A. (2010). Multicenter evaluation of signal enhancement algorithms for hearing aids. *J Acoust Soc Am, 127*(3), 1491-1505.

Lyzenga , J., & Smits, C. (2011). Effects of coarticulation, prosody, and noise freshness on the intelligibility of digit triplets in noise.(Report). *J. Am. Acad. Audiol., 22*(4), 215-221.

Marrone , N., Mason, C., & Kidd Jr, G. (2008). Tuning in the spatial dimension: Evidence from a masked speech identification task. *J. Acoust. Soc. Am., 124*(2), 1146-1158.

Martini, A. European Working Group on genetics of hearing impairment, European Commission Directorate, Biomedical and Health Research Programme (HEAR) Infoletter 2, November 1996.

Marzinzik, M. (2000). *Noise Reduction Schemes for Digital Hearing Aids and their Use for the Hearing Impaired.* Oldenburg: Shaker-Verlag. .

Micheyl, C., Arthaud, P., Reinhart, C., & Collet, L. (2000). Informational masking in normal-hearing and hearing-impaired listeners. *Acta Oto-Laryngologica, 120*(2), 242-246.

Middelweerd, M., Festen , J., & Plomp, R. (1990). Difficulties with speech intelligibility in noise in spite of a normal pure-tone audiogram. *Audiology, 29,* 1-7.

Miller, G. (1947). The masking of speech. *Psychological Bulletin, 44,* 105-129.

Miller, G., & Licklider, J. (1950). The intelligibility of interrupted speech. *J. Acoust. Soc. Am., 22*(2), 167-173.

Mohammadzadeh, H., Tartibiyan, B., & Ahmadi, A. (2008). The effects of music on the perceived exertion rate and performance of trained and untrained individuals during progressive exercise. *FACTA UNIVERSITASIS, Series: Physical Education and Sport, 6,* 67-74.

Moore, B., Johnson, J., Clark, T., & Pluvinage, V. (1992). Evaluation of a dual-channel full dynamic range compression system for people with sensorineural hearing loss. *Ear Hear., 13*(5), 349-370.

Neher, T. (2014). Relating hearing loss and executive functions to hearing aid users' preference for, and speech recognition with, different combinations of binaural noise reduction and microphone directivity. *Front. Neurosci.*, 1-14.

Neher, T., & Wagener, K. (2016). Investigating differences in preferred noise reduction strehgth among hearing aid users. *Trends Hear*, 1-14.

Neher, T., Grimm, G., & Hohmann, V. (2014). Perceptual consequences of different signal changes due to binaural noise reduction: do hearing loss and working memory capacity play a role? *Ear. Hear., 20*(10), 1-15.

Neher, T., Laugesen, S., Søgaard Jensen, N., & Kragelund, L. (2011). Can basic auditory and cognitive meaures predict hearing-impaired listeners' localization and spatial speech recognition abilities? *J. Acoust. Soc. Am., 130*(3), 1542-1558.

Neher, T., Wagener, K., & Fischer, R. (2016). Directional processing and noise reduction in hearing aids. Individual and situational influences on preferred setting. *J Am Acad Audiol, 27*(8), 628-646.

Nielsen, J., Nielsen, J., & Larsen, J. (2015). Perception-Based Personalization of Hearing Aids Using Gaussian Processes and Active Learning. *IEEE/ACM Trans. Audio, Speech, Language Process., 23*(1), 162-173.

Noordhoek, I., & Drullman, R. (1997). Effect of reducing temporal in- tensity modulations on sentence intelligibility. *J. Acoust. Soc. Am., 101*, 498–502.

Norris, R., Carroll, D., & Cochrane, R. (1992). The effects of physical activity and exercise training on psychological stress and well-being in an adolescent population. *J. Psychosom. Res., 36*, 55-56.

Oetting, D., Brand, T., & Ewert, S. (2014). Optimized loudness-function estimation for categorical loudness scaling data. *Hear. Res., 316*, 16-27.

Palmer, C., Bentler, R., & Mueller, H. (2006). Evaluation of a Second-Order Directional Microphone Hearing Aid: II. Self-Report Outcomes. *J Am Acad Audiol, 27*, 190-201.

Patterson, R., & Moore, B. (1986). Auditory filters and excita- tion patterns as representations of frequency resolution. In *Frequency Selectivity in Hearing.* Academic, London.

Peissing, J., & Kollmeier, B. (1997). Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *J. Acoust. Soc. Am., 101*(3), 1660-1670.

Perez, E., & Edmonds, B. (2012). A Systematic Review of Studies Measuring and Reporting Hearing Aid Usage in Older Adults since 1999: A Descriptive Summary of Measurement Tools. *Plos One, 7*(3), 1-8.

Pittman, A., & Stelmachowicz, P. (2003). Hearing Loss in Children and Adults: Audiometric Configuration, Asymmetry, and Progression. *Ear and Hearing, 24*(3), 198-205.

Plomp, R., & Mimpen, A. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology, 18*, 43-52.

Pollack, I. (1948). Effects of high pass and low pass filtering on the intelligibility of speech in noise. *J. Acoust. Soc. Am., 20*(3), 259-266.

Rajan, R., & Cainer, K. (2008). Ageing without hearing loss or cognitive impairement causes a decrease in speech intelligibility only in informational maskers. *Neuroscience, 154*, 784-795.

Ramji, R., Aasa, U., Paulin, J., & Madison, G. (2016). Musical information increases physical performance for synchronous but not asynchronous running. *Psychol. Music, 44*, 984-995.

Reddy, C., Shankar, N., Bhat, G., Charan, R., & Panahi, I. (2017). An Individualized Super-Gaussian Single Microphone Speech Enhancement for Hearing Aid Users With Smartphone as an Assistive Device. *IEEE SIGNAL PROCESSING LETTERS, 24*(11), 1601-1604.

Rennies, J., & Kidd Jr, G. (2018). Benefit of binaural listening as revealed by speech intelligibility and listening effort. *J. Acoust. Soc. Am., 144*(4), 2147-2159.

Rennies, J., Brand, T., & Kollmeier, B. (2011). Preditions of the influence of reverberation on binaural speech intelligibility in noise and in quiet. *J. Acoust. Soc. Am., 130*, 2999-3012.

Rennies, J., Kubiak, A., & Doclo, S. (October 2014). Personalization of audio playback using intuitive self-fitting interfaces . *The 48th annual conference of the German Society of Biomedical Engineering.* Hannover, Germany.

Rennies, J., Oetting, D., Baumgartner, H., & Appell, J.-E. (2016). User-interface concepts for sound personalization in headphones. *AES International Conference on Headphone Technology* (pp. 2-6). Aalborg: Audio Engineering Society.

Rennies, J., Schepker, H., Holube, I., & Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected ny noise and reverberation. *J. Acoust. Soc. Am., 136*(5), 2642-2653.

Rennies, J., Warzybok, A., Brand, T., & Kollmeier, B. (2014). Modeling the effects of a single reflection on binaural speech intelligibility. *J. Acoust. Soc. Am., 135*, 1556-1567.

Rhebergen, K., & Versfeld, N. (2005). A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. , 117(4),. *J Acoust Soc Am, 117*(4), 2181-2192.

Rhebergen, K., Versfeld, N., & Dreschler, W. (2006). Extended speech intelligibility index for the prediction of the speech reception thresh- old in fluctuating noise. *J. Acoust. Soc. Am., 120*(6), 3988–3997.

Rhebergen, K., Versfeld, N., & Dreschler, W. (2009). The dynamic range of speech, compression, and its effect on the speech recep- tion threshold in stationary and interrupted noise. *J. Acoust. Soc. Am., 126*, 3236–3245.

Ryan, A., & Bone, R. (1978). Noise-induced threshold shift and cochlear pathology in the Mongolian gerbil. *J. Acoust. Soc. Am., 63*(4), 1145-1151.

Schädler, M., Warzybok, A., & Kollmeier, B. (2018). Objective Prediction of Hearing Aid Benefit Across Listener Groups Using Machine Learning: Speech Recognition Performance With Binaural Noise-Reduction Algorithms. *Trends in Hearing, 22*, 1-21.

Schepker, H., Haeder, K., Rennies, J., & Holube, I. (2016). Perceived listening effort and speech intelligibility in reverberation and noise for hearing impaired listeners. *Int. J. Audiol., 55*, 738-747.

Schlauch, R., & Nelson , P. (2015). *Handbook of clinical audiology.* Philadelphia: Wolters Kluwer Health.

Schubotz, W., Brand, T., & Ewert, S. (2017). Speech intelligibility and spatial release from masking in maskers with different spectro-temporal modulations. *J. Acoust. Soc. Am., 141*(5).

Simpson, S., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N (L). *J. Acoust. Soc. Am., 118*(5), 2775–2778.

Standardization, ISO, I. 3.-1. Acoustics -- Reference zero for the calibration of audiometric equipment -- Part 1: Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones.

Steeneken, H., & Houtgast, T. (1980). A Physical method for measuring speech-transmission quality. *J. Acoust. Soc. Am., 67*(1), 318-326.

Stelmachowicz, P. G., Lewis, D. E., Hoover, B., Keffe, D. E. (1999). Subjective effects of peak clipping and compression limiting in normal and hearing-impaired children and adults. *J. Acoust. Soc. Am.,* 105(1), 412-422.

Stuckenberg, M., Nayak, C., Meyer, B., Völker, C., Hohmann, V., & Bendixen , A. (2018). Age Effects on Concurrent Speech Segregation by Onset Asynchrony. *J Speech Lang Hear Re*, 1-13.

Summers, V., & Molis, M. (2004). Speech reception in fluctuating and continuous maskers: Effects fo hearing loss and presentation level. *J. Speech Hear. Res., 47*, 245-256.

Swaminathan, J., Mason, C., Streeter, T., Best, V., Kidd Jr., G., & Patel, A. (2015). Musical training and the cocktail party problem. *Scientific Reports, 5*, 1-10.

Szmedra, L., & Bacharach, D. (1998). Effect of music on perceived exertion, plasma lactate, norepinephrine and cardiovascular hemodynamics during treadmill running. *Int. J. Sports Med., 19*, 32-37.

Taal, C., Hendriks, R., Heusdens, R., & Jensen, J. (2010). A short- time objective intelligibility measure for time-frequency weighted noisy speech. . *Acoustics Speech and Signal*

Processing (ICASSP), 2010 IEEE International Conference (pp. 4214-4217). Dallas, TX, USA: IEEE.

Taitelbaum-Swead, R., & Fostick, L. (2016). The effect of age and type of noise on speech perception under conditions of changing context and noise levels. *Folia Phoniatr Logop, 68*, 16-21.

Takahashi, G., & Bacon, S. (1992). Modulation detenction, modulation masking, and speech understanding in noise in the elderly. *J. Speech Hear. Res., 35*, 1410-1421.

Thorndike, R. (1953). Who belongs in the family? *Psychometrica, 18*(4), 267-276.

Tominaga, T., Hayashi, T., Okamoto, J., & Takahashi, A. (2010). Performance comparisons of subjective quality assessment methods for mobile video. *Proceedings of QoMEX*, Trondheim, Norway.

Torcoli, M., Herre, J., Fuchs, H., Paulus, J., & Uhle, C. (2018). The Adjustment/Satisfaction Test (A/ST) for the Evaluation of Personalization in Broadcast Services and Its Application to Dialogue Enhancement. *IEEE Trans. Broadcast, 64*(2), 524-538.

Völker, C., Ernst, S., & Kollmeier, B. (2018). Hearing aid fitting and fine-tuning based on estimated individual traits. *Int J Audiol*, 57, 139-145.

Völker, C., Warzybok, A., & Ernst, S. (2015). Comparing Binaural Pre-processing Strategies III: Speech Intelligibility of Normal-Hearing and Hearing-Impaired Listeners. *Trends Hear, 19*, 1-18.

Van Esch, T., & Dreschler, W. (2015). Relations between the intelligibility of speech in noise and psychophysical meaasures of hearing measured in four languages using auditory profile test battery. *Trends Hear, 19*, 1-12.

Verhaeghen, P., & Cerella, J. (2002). Aging, executive control, and attention: A review of meta-analyses. *Neurosci. Biobehav. Rev., 26*, 849-857.

Versfeld, N., & Dreschler, W. (2002). The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. *J. Acoust. Soc. Am., 111*(1), 401-408.

Vittitow, M., Windmill, J., Yates, J., & Cunningham, D. (1994). Effect of simultaneous exercise and noise exposure (music) on hearing. *J. Am. Acad. Audiol., 5*, 343-348.

Vongpaisal, T., & Pichora-Fuller, M. (2007). Effect of age on F0 difference limen and concurrent vowel identification. *J. Speech Hear. Res., 50*, 1139-1156.

Wagener, K., Bräcker, T., Brand, T., & Kollmeier, B. (2006). Evaluation des Ziffern-Triplet-Test über Kopfhörer und Telefon. *Proceedings of 9th congress of the German Society of Audiology*.

Wagener, K., Kühnel, V., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests in deutscher Sprache I: Design des Oldenburger Satztests. *Z Audiol, 38*, 4-15.

Walden, B., Surr, R., Cord, M., & Dyrlund, O. (2004). Predicting Hearing Aid Microphone Preference in Everyday Listening. *J Am Acad Audiol, 15*, 365-396.

Walden, B., Surr, R., Grant, K., Van Summers, W., Cord, M., & Dyrlund, O. (2005). Effect of Signal-to-Noise Ratio on Directional Microphone Benefit and Preference. *J Am Acad Audiol, 16*, 662-676.

Walton, T., Evans, M., Kirk, D., & Melchior, F. (2018). Exploring object-based content adaptation for mobile audio. *Personal and Ubiquitous Computing, 22*, 707-720.

Wan, R., Durlach, N., & Colburn, H. S. (2014). Application of short-time version of the equalization-cancellation model to speech intelligibility experiments. *J. Acoust. Soc. Am., 136*, 768-776.

Wininger, S., & Pargman, D. (2003). Assessment of factors associated with exercise enjoyment. *J. Music Ther., XL*, 57-73.

Wu, Y.-H., Stangl, E., Zhang, X., & Bentler, R. (2015). Construct Validity of the Ecological Momentary Assessment in Audiology Research. *J Am Acad Audiol, 26*, 872-884.

Xia, J., Noorale, N., Kalluri, S., & Edwards, B. (2015). Spatial release of cognitive load measured in a dual-task paradigm in normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am., 137*, 1888-1898.

Zokoll, M., Wagener, K., Brand, T., Buschermöhle, M., & Kollmeier, B. (2012). Internationally comparable screening tests for listening in noise in several European languages: The German digit triplet test as an optimization propotype. *International Journal of Audiology, 51*, 697-707.

APPENDIX

	AUDIOMETRIC PURE-TONE THRESHOLDS LEFT EAR / dB HL										
Subj. #	125 [Hz]	250 [Hz]	500 [Hz]	750 [Hz]	1 [kHz]	1.5 [kHz]	2 [kHz]	3 [kHz]	4 [kHz]	6 [kHz]	8 [kHz]
1	-5	-5	0	5	0	0	5	10	10	15	15
2	0	-5	0	0	0	0	-5	-5	0	0	0
3	0	-5	0	0	0	0	5	10	15	15	10
4	10	0	-5	0	-5	5	5	5	0	10	10
5	0	0	5	5	5	5	5	0	-5	10	0
6	0	0	-5	-5	0	-5	0	0	-5	10	15
7	0	5	0	5	0	0	0	0	10	15	15
8	10	5	10	20	20	30	35	35	40	50	50
9	0	5	15	15	20	20	30	20	20	45	45
10	0	0	5	10	5	10	25	35	55	75	75
11	10	10	20	30	35	30	25	20	10	30	40
12	0	0	0	5	5	30	40	50	60	70	70
13	25	25	30	30	25	40	45	35	25	40	30
14	10	15	35	50	50	50	50	55	40	30	20
15	40	45	45	50	55	60	60	60	60	70	70
16	20	20	20	20	15	30	40	45	50	70	75
17	25	30	45	50	55	55	55	55	70	80	80
18	30	30	45	45	50	55	60	60	65	85	75
19	25	30	40	45	45	50	55	55	60	70	80
20	20	10	15	25	40	55	65	75	75	100	100
21	55	60	55	50	45	45	35	30	35	40	70
22	15	10	15	45	45	70	75	80	80	75	80
23	15	15	5	5	5	20	35	55	90	100	90
24	50	55	60	70	60	65	75	75	85	85	75
25	15	10	15	15	10	10	15	30	45	70	90
26	35	45	50	50	45	55	65	65	70	75	70
27	5	5	15	15	10	25	30	50	70	80	90
28	45	50	55	65	70	75	75	65	70	70	75
29	20	15	20	30	30	60	80	95	100	*[1]	*[1]
30	55	55	55	55	55	60	65	70	80	80	80

	AUDIOMETRIC PURE-TONE THRESHOLDS RIGHT EAR / dB HL										
Subj. #	125 [Hz]	250 [Hz]	500 [Hz]	750 [Hz]	1 [kHz]	1.5 [kHz]	2 [kHz]	3 [kHz]	4 [kHz]	6 [kHz]	8 [kHz]
1	0	0	0	0	0	-5	5	0	0	10	0
2	0	5	5	5	0	-5	10	0	0	10	20
3	5	5	10	10	5	10	10	20	25	15	10
4	5	0	0	5	5	5	15	5	0	15	5
5	0	-5	-5	0	0	0	5	5	0	10	0
6	-5	-5	-5	0	0	5	10	0	0	-5	15
7	0	5	0	0	0	5	5	5	15	15	10
8	5	5	10	15	20	30	35	35	35	35	35
9	0	5	15	20	20	25	30	20	20	40	40
10	5	-5	5	5	0	20	25	45	55	70	70
11	20	20	20	30	35	35	30	20	15	35	35
12	5	5	5	5	10	25	40	50	70	80	80
13	25	20	25	25	25	35	35	25	15	25	30
14	25	30	40	50	50	55	50	45	30	20	15
15	40	45	55	55	50	55	55	55	55	70	65
16	20	15	20	25	15	25	35	55	50	70	80
17	15	25	35	40	40	50	50	55	60	75	85
18	35	45	50	55	55	65	65	65	75	85	80
19	25	40	45	45	50	55	60	60	65	90	90
20	25	10	15	30	40	60	65	75	85	105	105
21	75	80	65	60	60	45	50	45	45	100	90
22	15	10	25	50	75	85	75	70	80	90	90
23	15	15	10	10	5	15	25	45	65	70	75
24	55	60	65	70	75	75	80	85	95	95	80
25	10	15	20	20	15	15	20	35	35	85	90
26	30	30	35	35	40	55	70	70	75	75	70
27	20	20	10	5	0	20	25	40	50	80	85
28	50	55	65	75	75	75	75	65	70	70	75
29	20	20	25	30	30	45	65	75	95	*[1]	*[1]
30	40	40	45	50	50	55	55	65	75	80	80

	INDIVIDUAL FACTORS			
Subj. #	PTA @ better ear/ dB HL	Age / years	DTT: SRT / dB SNR	Goesa: SRT / dB SNR
1	1.25	32	-9.5	-5.7
2	-1.25	29	-9.5	-7.6
3	5.00	27	-9.5	-5.0
4	-1.25	29	-9.5	-6.0
5	0.00	27	-9.2	-5.5
6	-2.50	23	-10.0	-4.6
7	2.50	45	-9.3	-6.1
8	25.00	62	-9.2	-3.8
9	21.25	56	-7.8	-5.3
10	21.25	49	-6.7	-3.0
11	22.50	32	-9.0	-4.4
12	26.25	47	-6.7	-2.2
13	25.00	60	-8.8	-3.0
14	42.50	31	-7.0	-5.7
15	53.75	62	-3.8	1.6
16	46.25	73	-1.2	-0.3
17	30.00	81	-6.7	-1.4
18	55.00	72	4.2	0.0
19	50.00	71	3.0	3.5
20	48.75	65	-2.5	1.7
21	42.50	37	-4.8	-0.2
22	53.75	71	4.0	0.6
23	26.25	81	-5.5	-1.0
24	21.25	85	-6.0	-2.1
25	70.00	65	3.7	5.6
26	55.00	67	-3.5	0.9
27	21.25	81	-5.0	-1.6
28	67.50	62	7.3	3.8
29	53.75	75	-1.3	1.0
30	56.25	60	4.8	5.6

	EXPERIMENTAL SRTs / dB SNR*[2]			
Subj. #	MT cl	MT sp	TT cl	TT sp
1	-7.6	-11.9	-3.1	-20.3
2	-7.3	-12.5	-4.5	-19.8
3	-5.3	-12.3	-3.0	-19.2
4	-5.3	-10.9	-2.1	-18.8
5	-5.1	-9.9	-1.2	-18.0
6	-6.4	-10.4	-2.4	-17.1
7	-6.4	-11.5	-2.9	-16.7
8	-2.7	-8.6	0.7	-16.4
9	-5.0	-8.7	-0.5	-14.8
10	-4.5	-8.6	-0.8	-13.6
11	-4.9	-8.0	-0.6	-12.4
12	-4.6	-7.7	-1.2	-11.7
13	-4.2	-7.5	0.4	-10.4
14	-4.7	-7.0	-1.0	-7.4
15	-3.6	-7.2	-1.5	-7.3
16	-3.7	-6.2	-1.5	-6.7
17	-3.3	-5.5	2.4	-6.7
18	-2.7	-5.1	0.4	-6.3
19	-3.3	-4.5	-0.1	-5.9
20	-2.5	-5.1	0.6	-5.0
21	-2.6	-4.4	0.0	-4.8
22	-4.5	-5.6	0.4	-4.3
23	-2.3	-5.6	0.8	-4.0
24	-3.5	-6.5	0.5	-3.5
25	-1.9	-3.4	0.4	-3.5
26	-3.4	-4.1	-0.3	-3.4
27	-2.3	-4.6	1.5	-3.2
28	-1.5	-2.2	0.7	-2.6
29	-2.2	-2.7	1.0	-2.4
30	0.1	0.1	2.2	1.8

	EXPERIMENTAL SRM / dB		EXPERIMENTAL MTI / dB	
Subj. #	MT	TT	Cl	sp
1	4.3	17.2	-4.5	8.4
2	5.2	15.3	-2.8	7.3
3	7.0	16.2	-2.3	6.9
4	5.6	16.7	-3.2	7.9
5	4.8	16.8	-3.9	8.1
6	4.0	14.7	-4.0	6.7
7	5.1	13.8	-3.5	5.2
8	5.9	17.1	-3.4	7.8
9	3.7	14.3	-4.5	6.1
10	4.1	12.8	-3.7	5.0
11	3.1	11.8	-4.3	4.4
12	3.1	10.5	-3.4	4.0
13	3.3	10.8	-4.6	2.9
14	2.3	6.4	-3.7	0.4
15	3.6	5.8	-2.1	0.1
16	2.5	5.2	-2.2	0.5
17	2.2	9.1	-5.7	1.2
18	2.4	6.7	-3.1	1.2
19	1.2	5.8	-3.2	1.4
20	2.6	5.6	-3.1	-0.1
21	1.8	4.8	-2.6	0.4
22	1.1	4.7	-4.9	-1.3
23	3.3	4.8	-3.1	-1.6
24	3.0	4.0	-4.0	-3.0
25	1.5	3.9	-2.3	0.1
26	0.7	3.1	-3.1	-0.7
27	2.3	4.7	-3.8	-1.4
28	0.7	3.3	-2.2	0.4
29	0.5	3.4	-3.2	-0.3
30	0.0	0.4	-2.1	-1.7

	GROUPING*[3]	
Subj. #	Hearing loss group	Age group
1	1	1
2	1	1
3	1	1
4	1	1
5	1	1
6	1	1
7	1	2
8	2	2
9	2	2
10	2	2
11	2	1
12	2	2
13	2	2
14	3	1
15	3	2
16	2	3
17	3	3
18	3	3
19	3	3
20	3	2
21	3	1
22	3	3
23	2	3
24	3	2
25	2	3
26	3	2
27	2	3
28	3	2
29	3	3
30	3	2

*[1]: audiogram data not available due to output limit of audiometer, see section 2.2.1.

*[2]: MT: multitalker, TT: two-talker, cl: co-located, sp: spatial; see section 2.2.3.

*[3]: see description in section 2.3.6.

ACKNOWLEDGEMENTS

This thesis has been written at the Medical Physics Group in the School of Mathematics and Science (current School of Medicine and Health Sciences) of the Carl von Ossietzky Universität Oldenburg, Germany.

I would like to take the opportunity to thank all the people who contributed to its completion.

First, I would like to express my sincere gratitude to my supervisor, Birger Kollmeier, for his continuous support, valuable suggestions, guidance, and freedom he gave me to pursue my scientific interests.

Furthermore, I would like to thank Jan Rennies-Hochmuth, the leader of the group: Hearing, Speech and Audio Technology at Fraunhofer IDMT in Oldenburg, for giving me a chance to join his team where I learned a lot. Your support, patience, and enthusiastic ‘can do’ attitude have made this thesis possible.

I would like to also thank Stephan Ewert for his support with study design and many valuable suggestions.

Special thanks go to all my colleagues from the Fraunhofer IDMT and the University - for always being great colleagues and creating a friendly working environment. I would like to thank in particular: Lena Schell-Majoor, Dirk Oetting, Andreas Volgenandt, Tobias Bruns, and Rene Asendorf - I learned a lot from you!

Aleksandra Maria Kubiak

STATEMENT OF AUTHORSHIP

I hereby declare that I am the sole author of this thesis and that I have not used any sources other than those listed in the bibliography. I further declare that I have not submitted this thesis at any other institution in order to obtain a degree.

Poznań, 17. September 2020

……...

Aleksandra Maria Kubiak