Performance of Current Models of Speech Recognition and Resulting Challenges

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 Frau Dipl.-Phys. Wiebke Schubotz
geboren am 7.9.1984
in Rathenow

Gutachter: Prof. Dr. Dr. Birger Kollmeier
Zweitgutachter: Prof. Dr. Volker Hohmann
Tag der Disputation: 15.12.2015

Abstract

Speech signals are rarely perceived in quiet. Instead, there are usually other sound sources (so-called maskers), which appear in addition to the target speech and can severely hamper its recognition. Nevertheless, there are mechanisms and aspects of speech that allow a substantial speech recognition even in difficult listening conditions. These are, for example, the combination of across-frequency information (Bregman et al., 1990) for target identification or binaural cues to separate target speech from the masking background in spatial listening conditions (Bronkhorst and Plomp, 1988). Masking that is exhibited by background noises is usually ascribed to different masking aspects, such as energetic, amplitude modulation or informational masking (Durlach et al., 2003a; Stone et al., 2012). But there is debate over which factor is most influential and it is not always easy to entangle the influence of each single aspect (Brungart et al., 2001). A variety of complex maskers, that change consecutively in their spectro-temporal properties, is generated for the studies of this dissertation and addresses the individual masking aspects. Speech prediction models try to mimic the human auditory process and explain observed recognition data. To provide reliable predictions for realistic environments, they have to be tested in complex listening scenarios. Speech prediction models are often only tested in stationary or sinusoidally modulated maskers, but this is insufficient when predictions in realistic listening scenarios are anticipated. Therefore, current speech prediction models are used to predict speech reception thresholds (SRTs) in the same spectro-temporally complex maskers that are used for investigating the masking aspects. The models used in this dissertation are the four monaural models, *speech intelligibility index* (SII, ANSI, 1997), *extended speech intelligibility index* (ESII, Rhebergen et al., 2006), *multi-resolution speech-based envelope power spectrum model* (mr-sEPSM, Jørgensen et al., 2013), and *short-time objective intelligibility measure* (STOI, Taal et al., 2010), as well as the *binaural speech intelligibility model* (BSIM, Beutelmann et al., 2010). In detail, the studies in this dissertation address the following issues:
The study in chapter 2 investigates to which extent high-frequency envelope information can aid the identification of vowels in a lower frequency range. The studies in chapter 3 and 4 investigate binaural and monaural speech recognition in the various maskers. SRTs are measured with sentences from a German matrix sentence test (Oldenburger Satztest; Wagener et al., 1999). Observed SRTs contribute empirical data to the discussion of the different masking aspects and provide a data base for further studies on speech recog-

nition in complex maskers. The observed SRTs are also compared to the outcomes of the above speech prediction models. Thereby, it can be examined how well the different models account for the observed SRTs, as each model incorporates different signal analysis strategies. The studies designate the limits of the current model approaches, and thus, also constitute a benchmark for further studies with speech prediction models and future research on their improvement.

Zusammenfassung

Die Wahrnehmung von Sprache findet selten in Ruhe statt, sondern meist in einer Umgebung mit vielen verschiedenen Hintergrundgeräuschen. Diese Störgeräusche (oder Maskierer) behindern das Verstehen eines Zielsprechers zum Teil in erheblichem Maße. Dennoch gibt es Mechanismen, wie die frequenzübergreifende Analyse von Sprachanteilen (Bregman et al., 1990) oder die Auswertung von binauralen Merkmalen, die in schwierigen Hörsituationen ein gutes Sprachverstehen ermöglichen oder zur Trennung von Zielsprecher und Maskierer (Bronkhorst and Plomp, 1988) genutzt werden. Die Maskierwirkung von Störgeräuschen wird meist durch die drei Aspekte energetische Maskierung, Amplitudenmodulationsmaskierung und "informational masking" beschrieben (Durlach et al., 2003a; Stone et al., 2012). Allerdings ist die Abgrenzung zwischen diesen Aspekten nicht immer einfach und nicht abschließend geklärt, welcher Aspekt den größten Einfluss auf das menschliche Sprachverstehen hat (Brungart et al., 2001). Um den Einfluss der einzelnen Maskieraspekte näher zu beleuchten, wurden für diese Dissertation verschiedene komplexe Maskierer generiert, die konsekutiv ihre spektro-temporalen Merkmale ändern und die einzelne Aspekte ansprechen. Sprachverständlichkeitsmodelle versuchen die Prozesse des menschlichen Hörens nachzubilden und so gemessene Sprachverständlichkeitsschwellen zu erklären. Um Modelle in realistischen Hörsituationen anzuwenden, ist es nötig, sie auch in komplexen Störgeräuschsituationen zu testen. Derzeit werden viele Modellen in "einfachen" Störgeräuschen, wie stationärem oder sinusförmig moduliertem Rauschen, angewendet. Aber diese Herangehensweise ist nicht ausreichend, wenn Sprachverständlichkeitsmodelle das Sprachverstehen in realistischen Situationen vorhersagen sollen. Daher werden in dieser Arbeit aktuelle Sprachverständlichkeitsmodelle in denselben komplexen Störgeräuschen getestet, die zur Untersuchung der verschiedenen Maskieraspekte genutzt werden. Die Modelle, die in den Studien dieser Dissertation angewendet werden sind die monauralen Modelle *speech intelligibility index* (SII, ANSI, 1997), *extended speech intelligibility index* (ESII, Rhebergen et al., 2006), *multi-resolution speech-based envelope power spectrum model* (mr-sEPSM, Jørgensen et al., 2013) und *short-time objective intelligibility measure* (STOI, Taal et al., 2010), sowie das binaurale Modell *binaural speech intelligibility model* (BSIM, Beutelmann et al., 2010). Im Detail betrachten die Studien dieser Arbeit folgende Aspekte:
Die Studie in Kapitel 2 untersucht, inwiefern hochfrequente Einhüllendeninformation die Identifikation von Vokalen in einem tieferen Frequenzbereich un-

terstützen kann. Die Studien in den Kapiteln 3 und 4 untersuchen binaurales und monaurales Sprachverstehen in den verschiedenen spektro-temporalen Maskierern. Sprachverständlichkeitsschwellen werden mit Sätzen aus dem Oldenburger Satztest (Wagener et al., 1999) bestimmt. Die Studien steuern empirische Daten zur Diskussion der verschiedenen Maskieraspekte bei und liefern eine Datenbasis, die in weiteren Sprachverständlichkeitsstudien genutzt werden kann. Desweiteren werden die ermittelten Sprachverständlichkeitsschwellen mit den Vorhersagen der verschiedenen Sprachverständlichkeitsmodelle verglichen. Dadurch kann untersucht werden, wie gut die Analysestrategien der einzelnen Modelle die gemessenen Daten erklären und in welchen Störgeräuschsituationen die angewandten Modelle an ihre Grenzen stoßen. Damit setzen die Daten aus den Kapiteln 3 und 4 auch einen Maßstab, der für die Weiterentwicklung bestehender Modelle und zukünftige Studien zur Vorhersage von Sprachverstehen genutzt werden kann.

Contents

List of Figures

List of Tables

List of Abbreviations

AMM:	amplitude modulation masking
ANSI:	American National Standard Institute
CLUE:	conversational language understanding evaluation test
CVC:	consonant-vowel-consonant (logatome)
dB:	decibel
EM:	energetic masking
ERB:	equivalent rectangular bandwidth
ESII:	extended speech intelligibility index
ESIIsen:	extended speech intelligibility index with OLSA sentence as target input
Hz:	hertz
HI:	hearing-impaired
HL:	hearing level
ICRA:	International Collegium for Rehabilitative Audiology
ILD:	interaural level differences
IM:	informational masking
IMBM:	ideal monaural better-ear mask
IPD:	interaural phase differences
NH:	normal-hearing
OLSA:	Oldenburger Satztest (Oldenburger sentence test)
OLLO:	Oldenburger logatome speech corpus
OLnoise:	stationary noise spectrally matched to the OLSA material
mr-sEPSM:	multi-resolution speech-based envelope-power spectrum model
RMSE:	root-mean-square error
sEPSM:	speech-based envelope-power spectrum model
SI:	speech intelligibility
SII:	speech intelligibility index
SNR:	signal-to-noise ratio
SPL:	sound pressure level
SRT:	speech reception threshold
STOI:	short-time objective intelligibility measure

Chapter 1

General introduction and overview

A complex communication is one of the most unique characteristics of human nature; in fact, it is the en- and decoding of acoustic speech signals, the understanding and responding to them that make us human. Although animals such as frogs, birds, and primates (e.g., Suzuki, 2014; Seyfarth and Cheney, 2003) vocalize to change the behavior of other individuals, i.e., inform about potential predators (Suzuki, 2014) or organize collaborative defense (Elie et al., 2010), they are not involved in acoustic communication in a way humans are from early childhood until old age. Speech signals are essential to everyday communication of humans and investigating the recognition of such signals is in the focus of this dissertation.

1.1 Factors influencing speech recognition

The recognition of speech signals in everyday listening situations is influenced by many factors (e.g., Kollmeier, 1990; Bronkhorst, 2000, 2015). First of all, speech signals (termed *target* hereafter) are never perceived in quiet and there are many aspects that hamper a perfect recognition of target speech in a noisy background (termed *masker*). The recognition is influenced by masker aspects such as the frequency content, amplitude modulations, the duration of temporal gaps in the masker (Festen, 1987; Drullman et al., 1994; Brungart, 2001), gender differences of target and masker or the general absence or presence of interfering talkers. But since speech is a wide-band signal, information from different frequency regions can be combined to form an auditory object (Darwin, 1997) or to separate the target from a masking background. Moreover, humans perceive signals with both ears and this leads to fundamentally new signal aspects that help to localize a sound source and distinguish a target signal from a masker.

1.1.1 Across-frequency processing

Psychoacoustic studies found that the detection of single tones is easier when the filters adjacent to the target tone share the same amplitude modulations

(e.g., Moore, 1990; Verhey et al., 2003). Such decrease in detection thresholds is termed *co-modulation masking release* (CMR; Hall et al., 1984; Hall III and Grose, 1988) and demonstrates that an across-frequency processing takes place during the analysis of auditory signals in the human auditory system. This is supported by more recent studies, such as Gordon (1997) and Griffiths and Warren (2004), stating that there are also grouping mechanisms based on synchronicity and harmonic structure of the across-frequency components. Such processing does not only affect signal detection, but also the intelligibility of speech signals. Among others, Festen and Plomp (1990) and Lorenzi et al. (2006) found that speech reception thresholds (sometimes also termed recognition thresholds; SRTs) are lower in maskers that have temporal gaps. Thus, modulations across all auditory filters allow for *listening in the dips* (e.g., Bronkhorst, 2000; Brungart et al., 2001; Lorenzi et al., 2006) and increase speech recognition in fluctuating maskers. A recent study, using automatic speech recognition (ASR) approaches to explain psychoacoustic discrimination and speech intelligibility experiments, emphasizes the importance of across-frequency processing. Schädler et al. (2015a) show that observed SRTs can only be explained with feature sets that incorporate some kind of across-frequency processing. In summary, across-frequency processing is an important principle in the signal analysis of the human auditory system and it is especially important to investigate this in more detail in situations with speech processing.
The study in chapter 2 examines possible cues that could facilitate such across-frequency processing in speech stimuli. In contrast to most studies, which investigate across-frequency aspects with artificial stimuli such as synthesized vowels or complex tones (e.g., Bregman et al., 1985; Gordon, 1997), the study in chapter 2 utilizes natural speech signals. It makes use of a strobed-integration approach, presented in Patterson et al. (1995), and applies this to the identification of vowels in consonant-vowel-consonant logatomes, taken from the Oldenburg Logatome Corpus (OLLO; Wesker et al., 2005). It is investigated to what extent findings from artificial setups can be transferred to real speech signals.

1.1.2 Binaural Hearing: $1 + 1 \neq 2$

Since humans listeners have ears at both sides of the head, all auditory input signals are received and analyzed binaurally (at least for the case of normal-hearing listeners). In fact, binaural listening cannot be described as "listening twice with a single ear", but provides fundamentally new signal aspects that arise only when sounds are perceived with two ears. The sound arrives earlier at the ear facing the sound source, which establishes a time delay between the ears that is termed *interaural time* or *phase difference* (ITD, IPD). Besides, the listeners head leads to a decrease in the amplitude of the sounds at the ear opposite to the sound source, which introduces an *interaural level difference* (ILD). Both binaural cues (ILD and ITD) allow a sound localization and the separation of a target from interfering sound sources (a situation termed "cocktail party" problem by Cherry, 1953). These cues are especially helpful in acoustically challenging situations with many

interfering sources, reverberation or generally difficult *signal-to-noise ratios* (SNRs). The spatial distribution of signals also greatly influences binaural speech recognition and leads to decreasing SRTs, termed *spatial release from masking* (SRM), when the masking source is separated from the target (e.g., Plomp and Mimpen, 1981; Vom Hövel, 1984; Bronkhorst and Plomp, 1988, 1992; Litovsky, 2012). However, studies that examine SRM are often designed such that only one masker is presented in addition to a target signal and observed speech recognition can then be explained by the improved SNR that is present at the "better-ear", opposite to the masker (Edmonds and Culling, 2006; Kidd Jr et al., 2008). This leads to the question how binaural speech recognition is influenced in situations with two symmetrically positioned maskers. Recently, it was proposed that a binaural "better-ear glimpsing" (Brungart and Iyer, 2012; Best et al., 2015) takes place in such listening conditions. This suggests that those parts of the spectro-temporal signal representation in both ears are utilized that provide a favorable SNR.
The study in chapter 3 presents data on binaural speech recognition and SRM that is gained when maskers are placed symmetrically at both sides of the listener's head. It investigates SRTs in conditions where ILD and IPD cues are presented separately and in combination. The study uses a variety of maskers (see next paragraph) and thus provides a data base of SRTs for further studies on binaural speech recognition. Observed SRTs are compared to outcomes of the binaural speech intelligibility model (BSIM, Beutelmann et al., 2010) and serve as a test case for the analysis strategies implemented in this model. In addition, BSIM is run in various configurations to examine if observed SRTs can be explained by different binaural analysis strategies, such as the equalization-cancellation mechanism proposed by Durlach (1963), a possible binaural "better-ear glimpsing" or a binaural summation approach.

1.1.3 Masking aspects

Bronkhorst (2000) states that "speech intelligibility depends in a complex manner on the properties of the interfering signal(s), the number of signals, the spatial configuration of the sources, and the acoustic environment". Due to the variety of noise scenarios listeners constantly encounter, it is unclear to which degree speech recognition is hampered in such situations. Research has brought up different masking aspects that are thought to account for reduced speech recognition in a masker. The distinction is often made with respect to the stages of the auditory process in which the masker is thought to be analyzed. Currently, masking aspects in the literature appear as three-flavored and are termed *energetic masking* (EM), *amplitude modulation masking* (AMM), and *informational masking* (IM). The separation between the three is not always clear and continues to be subject of debate (Drullman, 1995; Brungart et al., 2001; Durlach et al., 2003a; Stone et al., 2012). However, the basic findings can be summarized as follows:
EM arises when masker energy falls into the same auditory filter at the same time as the target energy and exceeds the level of the target signal. Delgutte (1990), Moore and Vickers (1997), and Moore (2012) define EM as occurring when the neural response to the masker alone is not different

from the response to the mixture of target signal and masker. Thus, EM can be described by the SNR of target speech in a masking background. EM is the most prominent aspect and can explain much of the masking in noisy backgrounds.
AMM is mostly discussed for fluctuating maskers, such as amplitude modulated noise or interfering speech, and is especially pronounced when modulation arise over the entire frequency range of the masker. But following Stone et al. (2012), even a masker that is usually thought to convey EM (i.e., a stationary masker) can, when analyzed in narrow auditory filters, imply small modulations (Drullman, 1995) that convey AMM. AMM can be determined by the SNR_{env} of the envelope power in the modulation filter domain (Ewert and Dau, 2000; Dubbelboer and Houtgast, 2008; Jørgensen and Dau, 2011) and thus be described analogously to EM. Such approach takes into account the modulation strength of the masker in comparison to the target modulations.
While the definitions of EM and AMM are rather clear and established, this is less the case for IM. Often, masking effects, which cannot clearly be ascribed to either one of the aforementioned aspects, are broadly categorized as being informational masking effects. Thus, IM often serves as a "wastebasket" category, but this disregards the research that has been done to investigate this aspect more closely. As stated in Durlach et al. (2003a) and Rosen et al. (2013), the two most influential factors on IM are stimulus uncertainty (e.g., the stimulus complexity or rapid masker variations) and target-masker similarity. Consequently, strong IM is provided when target and masker material have the same semantical structure, as is the case for the Coordinate Response Measure corpus (CRM; Bolia et al., 2000), which is often used to study IM. But this speech corpus is not suitable to examine speech recognition (of German listeners) in everyday listening conditions, as it firstly consists of English words and secondly, does not have a semantically correct sentence structure. Thus, speech recognition in this dissertation is measured with a German speech corpus, the Oldenburger Satztest (OLSA, Wagener et al., 1999). This corpus also provides nonsense test sentences, but with a semantically correct sentence structure.
It is clear that more than one single masking aspect comes into play in natural listening conditions and it is often not possible to strictly separate between the three. Nevertheless, an understanding of the influence of the individual masking aspects is important to explain observed speech recognition. Therefore, speech recognition of OLSA sentences is assessed in a variety of maskers that change consecutively in their spectro-temporal properties. The generated maskers range from stationary noise to single, interfering talkers and address the individual masking aspects (see chapter 4). In this study, SRTs observed with those complex maskers are discussed with respect to the influence of the different masking aspects and compared to the outcomes of various speech prediction models that incorporate different analysis strategies. This study uses the same masker types as the binaural study (chapter 3) and thus expands the data base that is provided in this dissertation. The observed SRTs from both studies contribute empirical data to the discussion

of different types of masking and can be used as reference for other studies on speech recognition in complex maskers.

1.2 Speech prediction models

Speech prediction models are based on knowledge from speech recognition measurements with human listeners and incorporate this, in order to predict human SRTs. They try to mimic the human auditory process and ideally should do this to such a degree that long and costly measurements with human listeners can be replaced.
The investigation of speech perception and its modeling traces back as far as the 1920's and was done in the Bell Telephone Laboratories. Research done there focused on the distortion of speech transmitted over telephone (French and Steinberg, 1947; Fletcher and Galt, 1950) and investigated how the audibility of speech, the SNR or the sensitivity of the auditory system influence the recognition of speech signals. This led to the first representative speech prediction model, the articulation index (Kryter, 1962; ANSI, 1969), which was further revised in the 1990's and then given the name speech intelligibility index (SII, ANSI, 1997). This "new SII" included a decomposition of the signal in frequency bands with different center frequencies and varying filter width as found in the human auditory system (Glasberg and Moore, 1990). Moreover, it included a band-importance function that weights the contributions of the individual frequency bands to the overall intelligibility differently.
This approach became a standard tool for the prediction of speech recognition, but is mostly suited for predicting speech intelligibility in stationary maskers, when the masker energy is nearly constant over time. More recent models are designed to explain speech recognition also in fluctuating maskers (Rhebergen et al., 2006) or to account for the influence of amplitude modulations on the recognition of speech (Jørgensen and Dau, 2011). Using these approaches, human speech recognition can be well predicted, but usually, predictions are only done for stationary maskers or maskers that show coherent modulations across the entire frequency spectrum. But these maskers are clearly not complex enough to apply speech prediction models in order to simulate the hearing process in realistic listening scenarios. Therefore, it is interesting and necessary to search for masking conditions, where established models are challenged and where their predictions fail, as only this provides information on the need of new analysis strategies. Moreover, masker conditions in which the predictions of speech prediction models fail provide an ideal environment to test new analysis solutions.
The study in chapter 4 applies four current speech prediction models, the SII (ANSI, 1997), the extended SII (ESII, Rhebergen and Versfeld, 2005; Rhebergen et al., 2006), the multi-resolution speech-based envelope power spectrum model (mr-sEPSM, Jørgensen et al., 2013), and the short-time objective intelligibility measure (STOI, Taal et al., 2010). Observed SRTs are predicted with the same spectro-temporally complex maskers as used in chapter 3. As each model accounts differently for the three masking aspects, comparing observed data to the predicted SRTs allows to test the

assumptions those prediction models are based on and to investigate how well those assumptions can explain observed SRTs. Moreover, deviations between observed and predicted SRTs demonstrate the limits of the current models in these challenging listening scenarios. The study chapter 4 provides a benchmark for the four speech prediction models and shows that predictions can fail even in fairly "easy" situations, such as spectral mismatch of target and masker material.

1.3 What this dissertation provides (and what not)

In this dissertation, many aspects of speech recognition (e.g., across-frequency processing, the influence of binaural cues or the influence of different masking aspects) are examined and the performance of selected speech prediction models is investigated in detail. The current models are challenged in predicting SRTs in various masking backgrounds and observed SRTs (from monaural and binaural experiments) serve as benchmarks for the models. The strengths and limits of the individual analyses are revealed, which allows a discussion of the current understanding of speech perception as it is implemented in the individual models.
Investigating speech recognition in detail can involve a large number of parameters, because it is not only the type of target or masker material that influences the SRTs. There are also factors such as age, attention or other cognitive factors that affect the recognition of speech signals. A single dissertation cannot address all these factors, therefore, this thesis concentrates on certain aspects of speech recognition in humans:

- Speech recognition is investigated with normal-hearing listeners only, although implications on measurements with hearing-impaired listeners are discussed in section 5.3.
- Cognitive factors, such as training or cognitive load (e.g., Zekveld et al., 2011; Mattys and Wiget, 2011) are omitted in the discussion of the experimental results.
- Only previously published and established models for predicting human speech recognition are utilized in this dissertation. The models are not further developed, as one aim of this dissertation is to test the implemented analysis strategies. However, some suggestions for further model optimizations are given in sections 4.6.4 and C.2.2.

Chapter 2

The influence of high-frequency envelope information on low-frequency vowel identification in noise

Abstract[1]

Vowel identification in noise using consonant-vowel-consonant (CVC) logatomes was used to investigate a possible interplay of speech information from different frequency regions. It was hypothesized that the periodicity conveyed by the temporal envelope of a high frequency stimulus can enhance the use of the information carried by auditory channels in the low-frequency region that share the same periodicity. It is hypothesized that this acts like a strobe-like mechanism and would increase the signal-to-noise ratio for the voiced parts of the CVCs. In a first experiment, different high-frequency cues were provided to test this hypothesis, whereas a second experiment examined more closely the role of amplitude modulations and intact phase information within the high-frequency region ($4 - 8$ kHz). CVCs were either natural or vocoded speech (both limited to a low-pass cutoff-frequency of 2.5 kHz) and were presented in stationary 3-kHz low-pass filtered masking noise. The experimental results did not support the hypothesized use of periodicity information for aiding low-frequency perception.

[1] This chapter is a reformatted version of the manuscript "The influence of high-frequency envelope information on low-frequency vowel identification in noise", W. Schubotz, T. Brand, B. Kollmeier, and S.D. Ewert, published at PLOS ONE on january 5^{th}, 2016.

2.1 Introduction

Speech signals in general cover a wide range of frequencies and usually information across several frequency regions is grouped to form a single auditory object (Darwin, 1997). However, in everyday life speech is rarely perceived in quiet, but in a masking noise and thus, not all parts of the spectro-temporal representation of the speech signal can contribute equally to speech perception. According to speech perception models such as the speech intelligibility index (ANSI, 1997) or the glimpsing model (Cooke, 2006), those parts of the representation that have large positive signal-to-noise ratios (SNRs) are most useful for speech perception. Therefore, any mechanism that increases the SNR can generally be assumed to improve the perception of masked speech stimuli. Such mechanisms can be external (e.g., a directional microphone in a mobile device or hearing aid) or internal in the auditory system, e.g. selection of appropriate auditory channels that carry specific speech cues. The current study aims at clarifying whether stimulus information derived from a high-frequency auditory channel can be used to enhance the identification of low-frequency speech sounds, precisely vowels, in a masking noise.

The information from the different frequency regions is thereby generally represented by different aspects of the filter output. Narrow auditory filters can extract the specific frequency components of a signal very accurately at low frequencies, i.e. resolve individual components of complex tones, whereas broader auditory filters at higher center frequencies extract information from the envelope of a signal only. Therefore, the temporal representation of an analyzed signal can be very different for filters with different center frequencies (see Plack et al., 2006).

In case of the vowels tested in the current study, the human voice produces pulse trains with a periodicity that varies over time. In the frequency domain this corresponds to a complex tone with varying fundamental frequency F0 (inverse of the periodicity). Despite variations over time, periodicity and F0 can be regarded as quasi-stationary assuming a short-term analysis in the auditory system. When analyzed by narrow auditory filters in the low-frequency range, F0 is represented in a series of quasi-stationary frequency peaks (resolved harmonics). When analyzed by wider auditory filters in the high-frequency range, the periodicity is visible in the envelope of the filter output (unresolved harmonics). Therefore, this kind of periodicity information can be called F0-related information and occurs mostly in regions where the individual frequency components are unresolved (Meddis and O'Mard, 1997). Studies on the detection of pitch changes in complex tones (Moore and Moore, 2003; Oxenham et al., 2009) also suggest that periodicity information is encoded in the repetition rate of high-frequency temporal envelopes. For a wide-band signal such as speech, periodicity is thus correlated across different frequency regions, but thought to be extracted with different mechanisms and represented by different aspects of the filter output. However, some studies (e.g., Meddis and O'Mard, 1997) propose a single mechanism for the extraction of periodicity across different frequency regions.

The influence of F0 and resulting periodicity on speech perception is twofold.

On the one hand, F0 and periodicity information can be used for the segregation of speech, on the other hand they also facilitate a combination of information across different frequency regions. Studies such as Assmann and Summerfield (1990) and Culling and Darwin (1993) showed that, for example, the discrimination of two synthesized vowels and vowel identification is easier when the two presented stimuli do not share the same F0. Broadbent and Ladefoged (1957) showed that formants are grouped together if they share the same F0. Bregman et al. (1985) stated that in addition to periodicity, temporal aspects are also important. They showed that congruent amplitude modulations across several frequency regions are fused and support the discrimination of two complex tones. Brokx and Nooteboom (1981) and Bird and Darwin (1998) reported that the F0 is also important for the intelligibility of longer speech tokens (short sentences). Brokx and Nooteboom (1981) showed that intelligibility increased with pitch difference between target and interfering speech. In that study, listeners had to report the number of words they understood from (syntactically correct, but nonsense) target sentences when a constant difference in pitch between target and interfering speech was introduced by linear predictive coding (LPC). Bird and Darwin (1998) investigated the mechanism by which the auditory system exploits F0 differences in separating two sentence-length utterances. They found that a common F0 is used to group components within the two sentences when the F0 of the individual utterances are more than 5 semitones apart. They suggested that for smaller differences, speech intelligibility is governed solely by factors in the low-frequency region (such as separate formants in the first formant region or individual harmonic components that are attributed to either the masker or the target sentence). This is also found in Houtsma and Smurzynski (1990), where it is indicated that the discrimination of harmonic complex tones with different F0s relies primarily on low-frequency information, such as resolved lower harmonics.

In Josupeit et al. (2012), detection and discrimination thresholds of low-frequency complex tones (designed as "stylized formants") were found to improve significantly in the presence of an additional high-frequency cue band. The band provided information on temporal on- and offsets as well as periodicity of the low-frequency complex tone, but carried no other information whatsoever. Discrimination improved even if the low-frequency complex tone and cue band were not in harmonic relation. Studies on coherence masking protection (Gordon, 1997, 2000) found that certain cues (termed co-signal) in Gordon (1997) and Gordon (2000), although they alone did not provide direct information on the target signal, supported the perception of certain stimuli. In Gordon (1997) it was shown that high-frequency vowel energy can provide such cue. In that study, listeners had to distinguish between the vowels /ε/ and /i/ that had different first formant energies (more than a critical band apart), whereas the high- frequency energy was identical for both vowels. Discrimination between both vowels increased, according to Gordon (1997), because the first formant energy could be fused with the high-frequency vowel energy and enhanced the percept of the vowels. In Gordon (2000), listeners had to identify synthesized vowels that consisted of

a sine wave at a frequency corresponding to the first formant of the vowel /ɛ/ or /i/ and the co-signal, which was synthesized vowel energy corresponding to the second and third formant of the vowel /ɛ/. The co-signal was appropriate for either, /ɛ/ and /i/, and stimuli were only perceived as a certain vowel when sine wave and co-signal were presented together. It was found that identification thresholds decreased significantly due to the presence of the co-signal, suggesting that, although it was spectrally separated by more than a critical band from the sine wave, it contributed to the perceived sound. It was suggested in Gordon (2000) that this was caused by auditory grouping of sine wave and co-signal due to the exploitation of regularities in the temporal pattern of the two. This finding persisted when the synthesized vowel energy was replaced by a complex tone that had the same amplitude modulation for all its components. However, identification thresholds only decreased when the co-signal was temporally aligned with the rest of the stimulus (Gordon, 2000).

Based on results by Josupeit et al. (2012), Gordon (1997), and Gordon (2000) it can be hypothesized that periodicity information presented in a co-signal from a high-frequency region supports the perception of speech parts with the same periodicity in another frequency region. Such a hypothetical mechanism would be conceptually similar to the strobed integration stage as proposed by Patterson et al. (1995). This describes a temporal integration stage that is sensitive to periodicity and stabilizes those structures in the neural responses to stimuli that share the same periodicity. It could use F0-related temporal envelope information from a high-frequency region to support the perception of low-frequency components that share the same periodicity. Temporal peaks in the F0-related temporal envelope of the high-frequency part of the stimulus would define the "strobe points" that promote a certain periodicity. It is hypothesized that low-frequency channels with the same periodicity are selected and that the overall signal-to-noise ratio of those low-frequency channels is thus improved. This hypothesized mechanism would work best when the temporal envelope peaks and temporal fine structure information in the lower frequencies of the stimulus have a fixed phase relation across the frequency regions. This is the case for voiced parts of human speech, consisting of pulse trains filtered by the vocal tract transfer function.

The current study examined whether F0-related temporal envelope information derived from high-frequency ($4 - 8$ kHz) channels can facilitate the identification of vowels in a masking noise in the low-frequency region below 2.5 kHz. This is based on the idea of a strobe-like mechanism proposed in Patterson et al. (1995) and therefore constitutes a feasibility study. It is hypothesized that this mechanism works best in situations where speech is quasi-stationary (vowels), thus only a small portion of everyday speech is examined. The high-frequency periodicity information was provided in a high-frequency cue band with various configurations. The high-frequency cue band itself did not carry speech information that could be used when presented in isolation, but is thought to aid vowel identification.

In experiment 1, the experimental design of Josupeit et al. (2012) was extended by using speech stimuli that are closer to real speech and vowel

identification instead of a psychoacoustic discrimination task. This was done to test if a hypothesized strobed integration (Patterson et al., 1995) due to common periodicity (Bregman et al., 1985; Gordon, 1997; Bregman et al., 1990) is in principle possible for stimuli that are similar to the vowels in natural speech, since the research hypothesis cannot be tested with unmodified natural speech. Earlier studies (Bregman et al., 1985; Josupeit et al., 2012; Gordon, 1997) used artificial stimuli, i.e., complex tones or synthesized vowels, but the current study used vowels in consonant-vowel-consonant (CVC) logatomes to investigate the proposed mechanism and still be comparable to earlier studies. Vowels were either generated with linear predictive coding (LPC) or by low-pass filtering of intact speech material. Thus, the stimuli in the current study bridged the gap between purely artificial stimuli and natural speech. Modifications of the high-frequency cue band were tested to assess the role of temporal fine structure information from the high-frequency region as a possible co-signal, provided in addition to the low-frequency masked CVCs.
In experiment 2, certain high-frequency speech cues (i.e., amplitude modulations and phase information) were presented in addition to the low-pass filtered logatomes or in isolation to prove that these cues alone cannot lead to a substantial performance in vowel identification. If this was not the case, an improvement in vowel identification rates could be ascribed to those speech cues alone, instead of the periodicity information that is thought to be important for the strobed integration.

2.2 Methods

2.2.1 Ethics Statement

Written consent was obtained from each participant prior to the experiments. The experiments were approved by the local ethics committee of the University of Oldenburg.

2.2.2 Subjects

Seven subjects, aged $25 - 32$ years, participated in the first experiment. Six of them also participated in the second experiment. All listeners had an audiometric threshold of less than 20 dB HL or better at octave frequencies between 125 Hz and 8 kHz, except for one person who had 25 dB HL at 8 kHz. All listeners were naïve to the speech material and received an hourly compensation for their participation.

2.2.3 Apparatus & procedures

A five-alternative forced choice vowel identification task (see, Coughlin et al., 1998) was performed using a subset of the CVC logatomes of the Oldenburger Logatome Corpus (OLLO, Wesker et al., 2005). Forty CVCs with a combination of eight consonants and five long vowels were used. The sampling rate of the logatomes was 16 kHz. The subjects had to identify

the vowel in the CVC logatomes which were presented in random order over 40 trials. For every CVC, the five response alternatives were shown on the computer screen and had to be selected. Feedback was given to the listeners. The order of the response alternatives on the screen was randomized each time. Additionally, the order in which the experimental conditions were presented was randomized as well, excluding the case that the same condition appeared in successive trials. The experiment was done with three repetitions altogether. A shorter training test list with ten logatomes for each condition was presented to the subjects prior to each session of the actual measurement at a SNR of -14 dB. The signals were presented diotically at 65 dB SPL via Sennheiser HD 650 headphones in a double-walled, sound-attenuating booth. The stimuli were generated individually at runtime with Matlab (2011), using an alternative-forced-choice software package (Ewert, 2013). The headphones were calibrated on a Bruel&Kjær 4134 artificial ear.

2.2.4 Stimuli

The basis for the stimuli in both experiments were 40 CVCs (OLLO, Wesker et al., 2005), spoken by the same male speaker (S02M). The logatomes were selected with eight consonants ([b], [d], [f], [g], [k], [p], [t], [z]) and five long vowels ([a:], [e:], [i:], [o:], [u:]). Their mean fundamental frequency was 131 Hz. The stimuli were presented in an unmodulated masking noise (ICRA 1 noise from, Dreschler et al., 2001), low-pass filtered with an 8^{th}-order Butterworth filter with a cutoff-frequency of 3 kHz in order to mask the low-frequency parts of the CVCs. The ICRA 1 noise was derived from English text (see Dreschler et al., 2001) read by a female speaker that was filtered in three analysis bands (low-pass filter at 800 Hz, band-pass filter between 800-2400 Hz, and high-pass filter at 2400 Hz). Each band had a white spectrum and all three were added up to form the resulting noise. The added signal was then high-pass filtered at 100 Hz to produce a male speech spectrum of normal vocal effort as was desired in Dreschler et al. (2001) (find more details on the rationale and generation of the ICRA noises there). For the current study it was important that the noise had a spectrum that was similar to that of the target speech and masked frequencies below 3 kHz. A possible enhancement of periodicity in a low-frequency region would not be observable in quiet, but only in situations where lower frequencies are masked.
The measurements were performed for two fixed SNRs, -14 dB and -18 dB, calculated from the low-frequency part only. The logatomes had a length of about 500 ms (the segment representing the vowel was about 250 ms) and were placed in the middle of 1.5 seconds of masking noise. The exact stimuli setup however was different for both experiments (see Figs. 2.1 and 2.3).

2.3 Experiment 1

2.3.1 Detailed stimulus description

In the first experiment, the low-frequency part of the stimulus was either intact low-frequency speech (LFS) or a version of the logatome that was

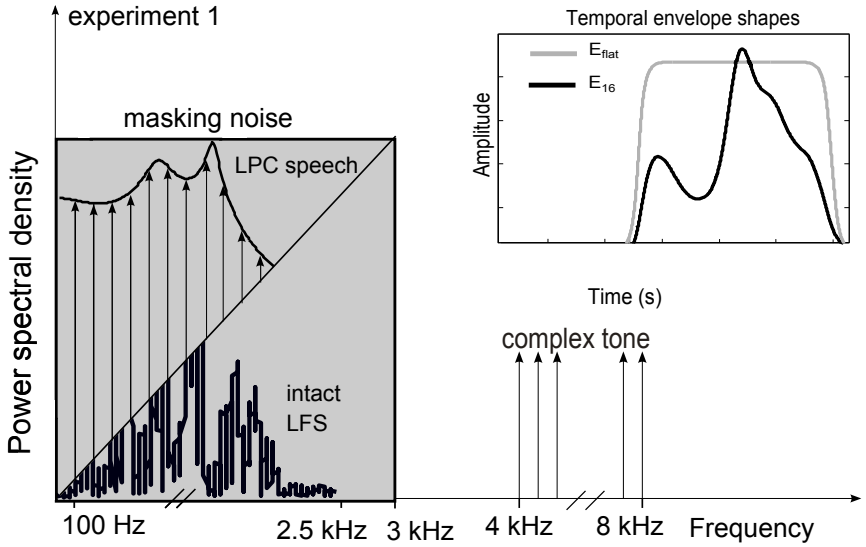


Figure 2.1: Fig. 2.1 shows the spectral properties (frequency regions of the complex tones, cutoff-frequency of the masker) of the low- and high-frequency part of the stimulus and the temporal shape of the high-frequency (HF) band envelopes of experiment 1. In this experiment, the low-frequency part of the stimulus is realized with either LPC-vocoded or intact, low-pass filtered speech (both limited to 2.5 kHz), as is schematically depicted in the left part of the figure. The flat envelope E_{flat} (gray line) encodes the on- and offset of the logatome, whereas the low-pass filtered envelope E_{16} (black line) ensures coherent amplitude modulations below 16 Hz in both frequency parts of the stimulus. Further high-frequency cues in experiment 1 are a 16 Hz low-pass filtered envelope (HFE_{16}), phase information provided by infinite clipping (HFIC), and intact high-frequency speech (HFS), all derived from the region of $4-8$ kHz from the intact logatome. All stimuli are presented in a low-pass filtered (cutoff-frequency is 3 kHz) stationary masking noise, indicated by the gray shaded area.

generated with linear predictive coding (LPC). LPC-vocoding was chosen to relate to findings from Bregman et al. (1985), Josupeit et al. (2012), and Bregman et al. (1990) where the stimuli were harmonic complex tones. The LFS was generated by low-pass filtering the intact speech with an 8^{th}-order Butterworth filter with a cutoff-frequency of 2.5 kHz. This was chosen to sufficiently cover the region of the second formants (see Pompino-Marschall, 2009), which is especially important for the differentiation between [i:] and [u:]. The LPC speech was restricted to the same frequency region as the LFS. For the LPC-vocoding, the spectral envelope of the intact logatome was approximated by an all-pole filter with 20 coefficients in 100 ms time windows with a Levinson algorithm (Markel and Gray, 1982). The envelope was imposed on a carrier that consisted of a harmonic complex tone ($F0 = 100$ Hz) with 25 components starting at 100 Hz that were all added in cosine phase. Thus, all natural F0 fluctuations of the speech stimuli were removed and a fixed F0 was set for the LPC synthesis, while the original spectral formant structure was maintained. Fixing the F0 led to a loss of voicing information (i.e., there was no distinction between voiced and unvoiced sounds), but this was irrelevant for the current study, since voicing information does not influence the identification of long vowels. Both low-frequency stimuli (LFS and LPC) were either presented alone in the masking background or in combination with an additional, simultaneous high-frequency cue band (HF band). The HF band consisted of a harmonic complex tone with 40 components ($F0 = 100$ Hz), starting at a frequency of 4 kHz. The cue band was multiplied in the time domain with two different temporal envelopes. The first condition provided a flat envelope (E_{flat}), where on- and offset ramps were aligned with those of the LFS and LPC to allow for grouping due to common temporal on- and offsets across the two frequency regions. The second condition was a temporal envelope derived from the LFS. The envelope was extracted via the Hilbert transform from the LFS and low-pass filtered with zero-delay to 16 Hz by forward-backward filtering with a 2^{nd}-order Butterworth filter. This low-pass filtered envelope (E_{16}) allowed a transfer of the slowly varying amplitude fluctuations in the speech envelope from the low-frequency to the high-frequency region. Thus, coherent amplitude fluctuations were provided in the frequency range below 2.5 kHz and in the frequency range from $4 - 8$ kHz. Both HF band conditions were presented with LFS and LPC, as shown in Fig. 2.1. Altogether, the experimental conditions were LFS, LFS-E_{flat}, LFS-E_{16}, LPC, LPC-E_{flat}, and LPC-E_{16}. The letters after the hyphen indicate that the HF bands were presented in addition to a certain low-frequency stimulus (LFS or LPC).

In combination with the LFS, three further HF band conditions were tested. While E_{flat} and E_{16} both transposed speech information originating from the low-frequency speech part of the logatome to the HF band, the additional conditions contained information from the intact high-frequency ($4 - 8$ kHz) speech part (HFS) of the logatome (indicated with the letters HF in the nomenclature of the experimental conditions). The intact HFS was generated by low-pass filtering the original speech token with an 8^{th}-order Butterworth filter (cutoff-frequency was 8 kHz) and subsequent high-pass filtering with

an 8^{th}-order Butterworth filter (cutoff-frequency was 4 kHz). For the first additional condition, termed LFS-HFE_{16}, the temporal envelope of the HFS was extracted via the Hilbert transform and then low-pass-filtered to 16 Hz by forward-backward filtering with a 2^{nd}-order Butterworth filter. Thus the envelope contained the slowly varying amplitude modulations that would naturally occur in the HFS. The HFE_{16} envelope was imposed onto the same complex tone as before. The second additional condition presented intact phase information (temporal fine structure) from the HFS together with the LFS. The phase information was extracted by determining the sign of the time signal (often referred to as infinite clipping), omitting amplitude fluctuations altogether. This condition was called infinite clipping (IC) condition, LFS-HFIC. The third additional condition was the intact HFS together with the LFS, called LFS-HFS. The different parts of the stimuli were set to the root mean square energy of the corresponding low- or high-frequency region of the original logatome to maintain the spectral energy distribution of the original speech token.

2.3.2 Results

Fig. 2.2 shows the mean vowel identification rates of experiment 1 and the corresponding standard deviations for both SNRs tested (-14 dB, left-hand side; -18 dB right-hand side). Panel a) presents those experimental conditions where LPC (open symbols) and LFS (filled symbols) were either presented alone (triangles) or in combination with E_{flat} (squares) or E_{16} (circles). Panel b) shows identification rates that were obtained with LFS and the various high-frequency cues indicated by the gray filled symbols. LFS, LFS-E_{flat}, and LFS-E_{16} are replotted from panel a) as black filled symbols.
When comparing LFS and LPC only in panel a), LFS showed higher identification rates. This was also the case when E_{flat} and E_{16} were presented in addition to LFS and LPC alone. The vowel identification rates were in general about five (for -18 dB) to ten percent (for -14 dB) higher when LFS was presented instead of LPC speech. At -14 dB SNR the mean identification rate for the LFS alone was 83%, while it was 66% at -18 dB SNR.
The statistical analysis was not performed on the identification rates in percent correct, but on rationalized arcsine transformed units (rau). This transformation produces values close to the original percentage scores, but solves the problem of a limited range of values. A limited range can be a problem for statistical analysis when percentages appear that are close to the upper or lower ends of the scale and violate the assumption of a normal distribution. The rau transformation was performed using the equations (3) and (7) provided in Studebaker (1985) for the individual data from the listeners in all experimental conditions. The rationalized arcsine transformation maps the percent correct values on an open scale that is linear and additive, takes into account the binomial distribution assumption, and produces scores that can be interpreted like percentage (see paragraph A.1 in the appendix).
A three-way repeated measures analysis of variance (ANOVA) was performed on the data from Fig. 2.2, panel a) with the main factors low-frequency part of the stimulus (LFS, LPC), HF band condition (no cue band, E_{flat}, E_{16}),

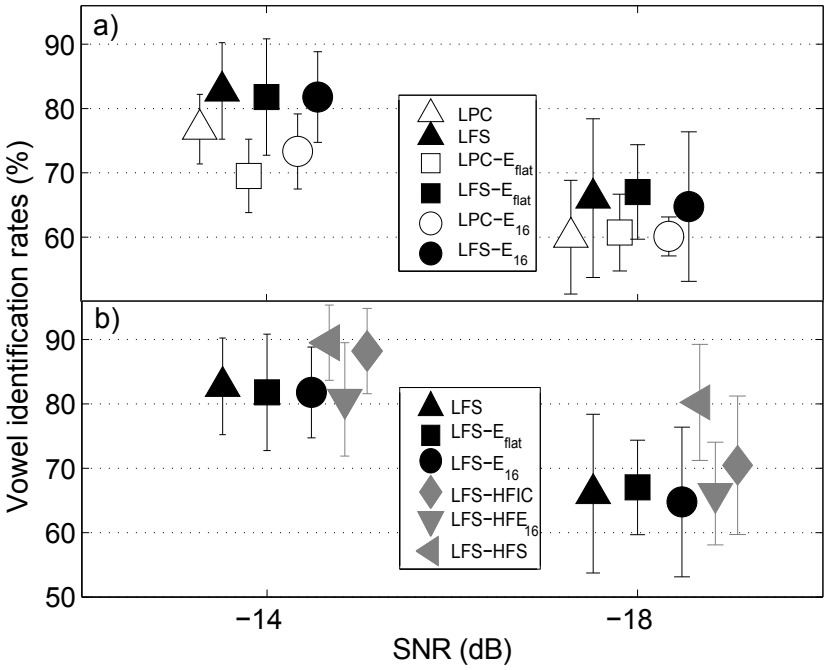


Figure 2.2: Mean vowel identification rates and corresponding standard deviations for the conditions in experiment 1. Panel a) shows the HF band conditions (E_{flat}, E_{16}) that were tested for both types of low-frequency design (LPC, LFS). Open symbols represent the rates that were measured when an LPC-vocoded logatome was presented in the low-frequency region of the stimulus. Filled symbols depict intact low-pass filtered logatomes in the low-frequency region of the stimulus. Stimuli where only low-frequency information was present are depicted with upward triangles. In panel b) the measured identification rates for all stimuli with LFS in the low-frequency part of the stimulus are shown. LFS-HFE_{16}, LFS-HFIC, and LFS-HFS are depicted with gray symbols, while LFS, LFS-E_{flat}, and LFS-E_{16} are replotted from panel a) with black symbols.

and SNR (-14 dB, -18 dB). The analysis showed a significant main effect of SNR [$F(1,6) = 1107.35$, $p < 0.001$] and low-frequency part of the stimulus $F(1,6) = 14.46$, $p < 0.01$]. Both values were Greenhouse-Geisser corrected. The HF band condition had no significant effect on vowel identification [$F(2,12) = 2.32$, $p = 0.14$]. Only the interaction of SNR and low-frequency part was significant [$F(1,6) = 18.36$, $p = 0.005$], all other interactions (SNR and HF band, low-frequency part and HF band, and the interaction of all three factors) were not significant.
In panel b), the additional HFS-based information (gray symbols) shows an increase in performance for the intact HF speech (LFS-HFS, gray left-pointing triangle) and HFIC (gray diamonds), while performance stays roughly the same for HFE_{16} (gray downward-pointing triangle). A two-way repeated measures ANOVA was conducted for the six conditions in Fig. 2.2, panel b), the main factors being SNR and HF band. The analysis showed a highly significant effect of the HF band [$F(5,30) = 32.33$, $p < 0.001$] and the SNR [$F(1,6) = 1673$, $p < 0.001$], the values for the SNR being Greenhouse-Geisser corrected. The interaction (Greenhouse-Geisser corrected) of both factors was not significant [$F(1.83, 10.98) = 2.34$, $p = 0.14$]. A post-hoc pairwise comparison (confidence level $\alpha = 0.05$) using Bonferroni correction was performed to investigate the simple effect of HF band. It showed that identification scores for LFS-HFS differed significantly from all other conditions. Moreover, LFS-HFIC was significantly different from LFS-E_{16} and LFS-HFE_{16}. Altogether, the presentation of F0-related information in addition to LFS improved vowel identification only in the LFS-HFIC and LFS-HFS condition. For LFS-HFS, the identification rates were generally about 10% higher than for all other LFS conditions.

2.4 Experiment 2

2.4.1 Rationale

Since the HFIC condition resulted in significantly improved vowel identification rates in experiment 1, the goal of experiment 2 was to examine possible explanations. Thus, the stimuli for experiment 2 were designed to specifically assess the role of phase information and amplitude modulations as a cue in the HF band. From experiment 1 it appeared that phase information in the HF region, conveyed by the HFIC, is as useful as intact HFS. A possible explanation is the used phase information in the HF region. Another possible explanation is reconstructed envelope cues, as described in Ghitza (2001). Ghitza (2001) showed that if manipulated speech with a flat envelope is provided as the input of an auditory filter, envelope fluctuations of the original speech can be partly recovered at the filter output when the input still contains the original phase information. The output of an auditory filter is then not smooth, but shows similar modulations as if the original signal had been analyzed in that particular filter. Thus, the HFIC condition from experiment 1, which preserved the phase information, could convey envelope modulations at the output of auditory filters, which are similar to

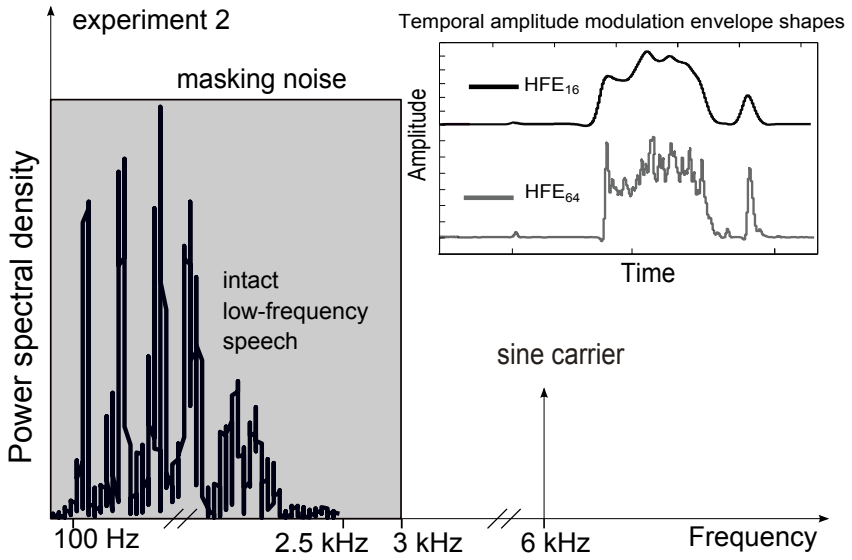


Figure 2.3: Fig. 2.3 shows the spectral setup of experiment 2. The HF band is replaced by a 6 kHz tone carrier, the low-frequency complex tone by intact low-pass filtered speech (LFS from experiment 1). The different low-pass filtered amplitude modulation envelopes (HFE_{16}, HFE_{64}) are shown in gray and black. Further high-frequency cues in experiment 2 are phase information provided by infinite clipping (HFIC from experiment 1), band-limited phase information provided by infinite clipping ($HFIC_{BP}$), and intact high-frequency speech (HFS). Again, all stimuli are presented in stationary noise that is low-pass filtered at 3 kHz.

intact HFS as a consequence of envelope reconstruction. Accordingly, not the high-frequency phase information itself, but reconstructed modulation cues could have improved the vowel identification in experiment 1.

2.4.2 Detailed stimulus description

For the second experiment the logatomes were up-sampled to 96 kHz during signal manipulation and then down-sampled and presented at 16 kHz, as this was the original sampling frequency. For this experiment the low-frequency part of the stimulus was LFS only, as shown in Fig. 2.3. It was examined if a change in vowel identification occurred when the infinite clipping information is band-limited after the phase extraction. Therefore, the HFIC condition in experiment 2 was slightly changed from the one in experiment 1: After the phase extraction from the HFS, a band-pass (BP) filter from $4 - 8$ kHz was applied to the HFIC, in order to limit the frequency region of the phase fluctuations. This was achieved with a zero-delay 4^{th}-order Butterworth filter. This experimental condition was termed LFS-$HFIC_{BP}$. To be able to compare this directly to the unfiltered condition, the HFIC condition was measured again in experiment 2 (LFS-HFIC).
The amplitude modulations from the high-frequency part of the logatome were

provided by applying different low-pass filtered temporal envelopes of HFS on a 6-kHz sine carrier. The resulting spectrum of the modulated 6-kHz carrier was thus centered in the $4-8$ kHz band. The use of a pure tone carrier ensured that the amplitude modulations at the output of an auditory filter most closely resembled the desired amplitude modulations and that the carrier phase did not carry any information. This 6-kHz sine carrier condition is indicated with HFE(S) in the following as it differed from the HFE setup of experiment 1, where the carrier was a complex tone. The amplitude modulations in Fig. 2.3 were generated by low-pass filtering the Hilbert envelope of the HFS to 16 Hz and 64 Hz (LFS-HFE(S)$_{16}$ and LFS-HFE(S)$_{64}$). These frequencies were chosen to allow for formant transitions within the CVC as those occur at modulation frequencies above 16 Hz. Moreover, Drullman et al. (1994) suggested that also energy modulations above 16 Hz are important in certain listening conditions. As in experiment 1, there was also the intact HFS presented as a high-frequency cue (LFS-HFS). All high-frequency cues were presented in addition to the LFS and also alone (without the LFS) in the masking noise. This was done to verify that improved vowel identification was caused by the presence of the additional high-frequency information (co-signal) that alone does not carry any valuable vowel information. As for experiment 1, the original energy distribution of the two frequency regions of the logatomes was maintained.

2.4.3 Results

Fig. 2.4 shows the mean vowel identification rates for experiment 2 across the listeners together with the standard deviations. The upper panel shows the results for a SNR of -14 dB, the lower panel for -18 dB. The identification rates for the HFIC conditions are depicted in the left part of the figure, identification rates for conditions with different cutoff-frequencies of the temporal HF envelopes in the middle, and identification rates for intact speech in the right part of the figure. Filled symbols represent those experimental conditions, where the HF band was presented in addition to LFS and the masking noise. Open symbols are for the respective HF bands alone in the masking noise.

As for experiment 1, the pattern of results was similar for both SNRs. For the LFS-based conditions (filled symbols) identification rates were close to 90% at an SNR of -14 dB and about ten percent lower (about 80%) for the SNR of -18 dB. When the four HF bands (amplitude modulations and intact phase information) were presented in isolation, the identification rates were about chance level for the amplitude modulations and slightly higher for the presentation of IC and IC$_{\mathrm{BP}}$ at both SNRs. The rates for the presentation of the intact HFS alone were above 60% for both SNRs and thus substantially above chance level. When LFS alone was presented, identification rates were even higher, 88% (for -14 dB) and 77% (for -18 dB). The highest identification rates were reached when LFS and HFS were presented in combination and this was found for both SNRs. All six experimental conditions containing LFS were analyzed with a two-way repeated-measures ANOVA with the main factors SNR and HF band. A significant main effect of SNR [$F(1, 5) = 196.98$, $p < 0.001$]

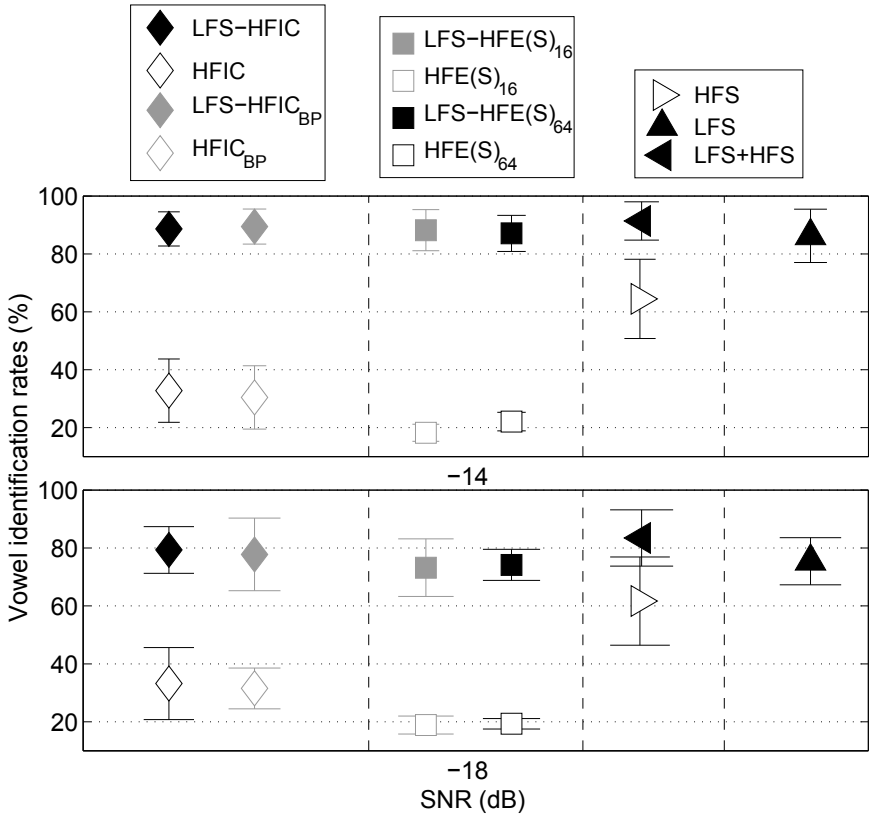


Figure 2.4: Mean identification rates and corresponding standard deviations for experiment 2. The upper panel shows the results for a SNR of -14 dB, the lower panel for a SNR of -18 dB. Filled symbols represent conditions where LFS and the high-frequency cues are presented together, open symbols conditions where the high-frequency cues are presented alone. In contrast to experiment 1, the high-frequency envelopes are imposed onto a 6-kHz sine carrier instead of a complex tone. The left part of the figure shows the identification rates for the infinite clipping cues, the middle panel the rates for the amplitude modulation cues, and the right panel the rates for intact speech cues. The rates for LFS alone are presented at the rightmost position as they serve as a reference for all other experimental conditions. As for experiment 1, all stimuli are presented in a stationary noise, low-pass filtered at 3 kHz.

and HF band [$F(1.96, 9.8) = 7.07$, $p < 0.001$] was found, but not of the interaction between the two [$F(5, 25) = 0.91$, $p = 0.49$]. The values for SNR and HF band were Greenhouse-Geisser corrected. A post-hoc pairwise comparison (confidence level $\alpha = 0.05$) using Bonferroni correction showed that identification rates for LFS-HFS were significantly higher than those for LFS-HFE(S)$_{16}$, LFS-HFE(S)$_{64}$, and LFS.

2.5 General discussion

2.5.1 Testing the strobe-like mechanism

Although a strobe-like mechanism is supported by several studies (Bregman et al., 1985; Patterson et al., 1995; Gordon, 1997, 2000; Josupeit et al., 2012) it is unclear to what extent this can be assumed for vowel identification in speech signals. The current data does not support a benefit for vowel identification in noise by means of periodicity that is conveyed via high-frequency envelope information: When LPC-vocoded speech and F0-related temporal envelope periodicity information are presented at the same time, there is no improvement in vowel identification rates. The HF band conditions E_{flat} and E_{16} do not increase vowel identification performance although both share the same periodicity with the LPC-vocoded logatome. Moreover, they provide synchronous on- and offsets, which should enable grouping of the two sounds according to Bregman and Pinker (1978). Furthermore, the E_{16} condition provides congruent amplitude fluctuations for both frequency regions, which should aid a grouping of the two sounds according to Bregman et al. (1985) and Bregman et al. (1990). In contrast to the assumptions, no increase in vowel identification rates was observed and therefore it cannot be verified that a strobe-like mechanism as described in Patterson et al. (1995) occurs for the stimuli in the current study. This does not completely rule out such a mechanism, but it cannot be observed when periodicity information is presented in high-frequency envelope information. The results obtained in studies with artificial stimuli (Broadbent and Ladefoged, 1957; Gordon, 1997, 2000) can therefore not easily be verified with the stimuli of the current study that are closer to real speech.

A possible explanation for the lack of improvement, opposed to earlier studies with non-speech stimuli, could be the stimulus duration. In Josupeit et al. (2012) it is stated that the absence of identification improvement for stimuli of about 500 ms or longer is caused by the long integration time in which within-channel information dominates across-frequency information. Stimuli in the current study are in total about 500 ms long, but the voiced part is about 250 ms long and thus a combination of across-frequency cues should in principle be possible. Stimulus duration alone is therefore not expected to be the main reason for the lack of identification improvement. Another possible explanation could be a part of the signal manipulation that could have caused a slight change in the formant structure of the vowels and thus vowel confusion: The spectral envelope used in the LPC-vocoding is derived from the original logatome with F0 = 131 Hz and is imposed on a complex

tone of F0 = 100 Hz, which constitutes a typical fundamental frequency for male speakers (Loizou, 2013). This could have caused a slight shift of formants in the LPC speech in the low-frequency range, but is unlikely to be the reason for a complete lack of identification improvement, because the formants should be analyzed in the same auditory filter as before.
Besides investigating the percent correct values, confusion matrices can be additionally assessed for each experimental condition (shown in Figs. A.2-A.6 in paragraph A.2 of the appendix). The confusion matrices suggest that the lack of vowel identification improvement is most probably caused by classification errors of certain vowels. Largest identification rates occur on the main diagonal and confusions only in sub-matrices. For the better SNR the [e:] is mistaken for [i:] and vice versa, and [o:] is mistaken for [u:] and vice versa, while all other possible confusions are negligible. This pattern smears out for the worse SNR, but still, confusions are largest for the sub-matrices [e:,i:], and [o:,u:]. Recognition of the vowel [a:] is very good throughout all experimental conditions, presumably because [a:] has no similar counterpart within the vowels tested in this study. This is, because its position in the vocal triangle is far apart from the other vowels (Pätzold and Simpson, 1997).

2.5.2 The influence of intact low-frequency speech information

Examining identification rates when envelope information is presented in addition to LFS (instead of LPC-vocoded speech) does not investigate a possible strobing mechanism, but clarifies the role of the LFS itself. When LFS is used instead of LPC-vocoded speech, vowel identification rates increase significantly. This is in line with findings from Kong and Carlyon (2007), Turner et al. (2004), and Qin and Oxenham (2006) on simulated combined acoustic and electrical hearing, stating that additional low-frequency information leads to significant improvements in word intelligibility performance and is greater as the cutoff-frequency increases. Moreover, Kong and Carlyon (2007) show that the presence of low-pass filtered intact speech at one ear enhances speech perception greatly, even if other low-frequency cues are presented at the other ear. This suggests a substantial influence of intact low-frequency speech to speech perception as is found in experiment 1 and experiment 2. The stimuli in the current study (intact low-frequency speech in addition to a complex tone) are similar to stimuli used in Kong and Carlyon (2007) and Qin and Oxenham (2006), although the low-frequency region for those studies has a lower cutoff-frequency. Thus, findings from the current study can in principle be compared to those on combined hearing.
In experiment 1, no increase in vowel identification rates is observed when the E_{flat}, E_{16} or HFE_{16} cue bands are presented in addition to LFS. The HF cue band with its fixed F0 conveys "false" F0 information in the HF region, producing a mismatch of periodicity for the two frequency regions. According to Bird and Darwin (1998), "false" F0 in the high-frequency range can easily be rejected and thus LFS alone should suffice for the vowel identification observed. This is confirmed as the identification rates are similar compared

to LFS alone. When E_{16} is presented vowel identification should increase due to grouping of congruent amplitude modulations (Bregman et al., 1985, 1990), but this is not observed. Again, low-frequency information (such as F0 fluctuations or formant transitions) seems to suffice for vowel identification in stimuli that are similar to natural speech. This is in line with findings from Carlyon (1996) and Deeks and Carlyon (2004), stating that normal-hearing listeners rely primarily on resolved lower order harmonics when they segregate two concurring sounds. Moreover, the pattern of vowel confusions also suggests that most information on the vowels is already present in the LFS. Confusion matrices for experiment 2 (shown in Figs. A.4 - A.6 in the appendix A.2) show that if LFS information is present in the stimuli, vowel identification is generally good. As for experiment 1, confusions appear only in sub-matrices for [e:,i:] and [o:,u:] and the vowel [a:] is robust towards confusion. If only a HF cue is presented, confusions are randomly distributed across the entire confusion matrix, confirming that the HF cue alone does not provide any substantial information on the vowel. The only exception is the vowel [i:] that shows larger identification rates than all other vowels for the condition HFIC and $HFIC_{BP}$. Throughout experiment 2 there are hardly any confusions for [i:] and [u:] when LFS is present in addition to a HF cue, which means that information on the second formant is present to allow a distinction between both.

Data from both experiments show that HFS significantly improves vowel identification when combined with LFS. In experiment 2, HFS alone yields rates of about 60% correct, which is lower than the rates for LFS alone, but substantially above chance level. LFS alone leads to identification rates of about 86% correct for -14 dB (75% for -18 dB) and is increased by $6-8\%$ when the combination of LFS and HFS is presented to the listeners. Combined intact speech results also in significantly higher identification results for experiment 1. This finding is similar to Warren et al. (1995), where the influence of spectral slits on sentence intelligibility is investigated. The study states that even if intelligibility is reduced for single, narrow frequency bands, it rises tremendously when these are combined. The effect found in the current study is not as large as in Warren et al. (1995), but still noteworthy. The maximally expected improvement under the assumption of optimal combination of independent information in LFS and HFS can be calculated using the combined error rates (for details, see Allen, 1994). The error rates for HFS and LFS alone are 0.356 and 0.138, and 0.383 and 0.246 at -14 dB and -18 dB, respectively. The maximally expected identification rates for HFS and LFS in combination, as calculated from multiplication of the error rates for HFS and LFS alone, are 95.1% for -14 dB and 90.6% for -18 dB. The measured rates are 91.38% and 83.47%, which is less than the maximally expected rates, but still substantial, regarding the overall high identification rates.

2.5.3 Possible use of phase information from the high-frequency region

Somewhat unexpectedly, the presentation of intact phase information in the HF band (HFIC) improves vowel identification significantly for experiment 1. There are two possible explanations for this, the reconstruction of envelope fluctuations according to Ghitza (2001) and the use of temporal fine structure information from the high-frequency region above 4 kHz. If the improvement is caused by envelope reconstruction, conditions in experiment 2, where the corresponding amplitude modulations are provided as a HF cue, should substantially improve vowel identification. However, this is not the case. The lack of identification improvement weakens the hypothesis that the LFS-HFIC condition is helpful because of reconstructed envelope cues. To rule out this possibility, subsequent measurements should be conducted in which amplitude modulations are presented over a smaller range of vocoder bands (e.g. one or more 1-ERB filters), instead of one broad filter range as done in the current study. This would allow a closer assessment of the frequency range in which recovered modulations are eventually helpful.
Results of experiment 1 leave the use of temporal fine structure as a possible explanation for the improvement of vowel identification. In contrast to literature like Johnson (1980) or Palmer and Russell (1986), reporting that phase-locking of the auditory nerve fibers limits the direct extraction to frequencies of $1-2$ kHz, studies, such as Oxenham et al. (2011), Moore and Sęk (2009), and Moore and Ernst (2012) report that the auditory system could have access to fine structure cues above 3 kHz. So far, it is under debate at which frequency a transition occurs from a direct extraction of the phase information (possibly at lower frequencies) to a place mechanism (possibly at higher frequencies). But these studies (Oxenham et al., 2011; Moore and Sęk, 2009; Moore and Ernst, 2012) show that a direct extraction is robust up to 6 kHz and indeed possible for even higher frequencies up to 8 kHz. This indicates that correlated phase information across frequencies, provided in the HFIC condition, could cause the improved vowel identification rates in the current study, even if the phase information is band-limited as for the HFIC_{BP} condition.

2.5.4 Limitations of the current study

Considering the overall identification rates in experiment 2, there is a general trend towards higher identification rates for conditions that were also measured in experiment 1 (LFS, LFS-IC, and LFS-HFS). This might be caused by training effects of the participants or by general ceiling effects, due to the amount of LFS information that is present in the stimuli. Listeners were naïve to the target material for experiment 1, however, six of the seven listeners also participated in experiment 2, which might have led to a training effect in the second experiment. On the other hand, experiment 2 was a follow-up study that took place half a year after the first experiment and thus, it is questionable to what extent listeners could rely on knowledge from the first experiment. Taken together it is not obvious that the results from

experiment 2 would differ much when a new set of listeners was recruited. Thus, further studies should be performed at slightly lower SNRs to reduce the ceiling effects. Generally, vowel identification rates are high for all conditions that include intact LFS information. A possible reason could be that the cutoff-frequency of 2.5 kHz for the low-frequency region of the stimuli is chosen too high and that therefore, vowel identification in both experiments is ruled mostly by the LFS alone. But this certain cutoff-frequency is chosen to allow a distinction between [i:] and [u:], which depends on the second formant of both vowels. Moreover, the low-frequency part of the logatome is masked by the background noise, so low-frequency information is not easily accessible.
Another possible reason for the lack of identification improvement might be the presentation of the stimuli in a stationary background noise. Studies, such as Qin and Oxenham (2003) show that normal-hearing listeners benefit from temporal envelope information only when the masker is an interfering talker and provides temporal gaps. For the current study, however, a presentation in a stationary background is chosen, to be comparable to studies like Josupeit et al. (2012) and Gordon (1997) that use a stationary background noise, and to prevent the vowel from being unmasked: the logatomes of the current study are so short that the entire logatome could randomly fall in a gap of a fluctuating masker, strongly reducing the low-frequency masking effect and probably strongly increasing variability in the data. But, it can be hypothesized that vowel identification with a similar setup should increase when measurement are done in a fluctuating masker that provide silent intervals in which the information from both frequency regions can be optimally combined.

2.6 Conclusions

1. Vowel identification in CVC-logatomes in a stationary masking noise is improved for low-pass filtered speech when compared to LPC-vocoded speech limited to the same frequency range.

2. Findings on the improvement of identification of "stylized formants" in Josupeit et al. (2012) cannot be reproduced directly for signals that are closer to real speech than complex tones or synthesized vowels. The presentation of a high-frequency band with common periodicity, on- and offsets, and temporal envelope shape in addition to a complex tone has no effect on vowel identification.

3. The results do not support a hypothesized strobe-like mechanism that uses common periodicity information across frequency bands. With the current data it cannot be verified that F0-related temporal envelope information, providing such periodicity information, aids the enhancement of frequency channels with the same periodicity in a low-frequency region. This does not rule out the existence of such mechanism, but it cannot be verified with stimuli chosen in the current study.

4. A significant improvement is observed in experiment 1 when the high-frequency band contains the intact phase information (HFIC condition) of the speech signal in that frequency band. The presentation of amplitude modulation cues in experiment 2 does not indicate that this improvement is caused by recovered amplitude modulation cues according to Ghitza (2001). This leaves the use of temporal fine structure in the high-frequency region as a possible explanation for the vowel identification improvement in experiment 1.

5. Significant identification improvement is observed in experiment 1 when intact high-frequency speech is presented as a high-frequency cue in addition to low-frequency speech. In experiment 2 the intact high-frequency speech leads to a significant improvement in identification rates compared to other high-frequency cues. However, there is no improvement for most high-frequency cues in experiment 2, indicating that vowel identification is possibly ruled by the information in the low-frequency region of the stimuli.

Acknowledgments

We would like to thank the members of the Medical Physics group for constant support and fruitful discussions. Furthermore, we thank two anonymous reviewers for their helpful comments on earlier versions of this manuscript. This study was supported by the Deutsche Forschungsgemeinschaft (DFG) within the project B1 of the SFB/TRR31 “Das aktive Gehör.”

Chapter 3

Binaural masking release in symmetric listening conditions with spectro-temporally modulated maskers

Abstract[1]

Speech-reception thresholds (SRTs) decrease as target and maskers are spatially separated (spatial release from masking, SRM) even if two maskers are symmetrically placed around the listeners head and long-term interaural level and time differences (ILD, ITD) are absent. In this case, speech intelligibility (SI) cannot be explained by an improved long-term signal-to-noise-ratio (SNR) caused by the head shadow at one better-ear alone, but could be facilitated by short-term spectro-temporal segments ("glimpses") in each ear that provide favorable SNRs. The current study assessed SRTs for a frontal target in a symmetric masker setup. The spectro-temporal properties of the maskers were systematically varied, ranging from stationary noise to single talkers. Maskers were modified by head-related transfer functions providing different binaural cues, by presenting only glimpses derived with a fast-switching better-ear mechanism, and by generating a masker with an "infinite ILD". It was investigated to which extent the observed SRM can be explained by the individual modifications and data were compared to predictions of a binaural SI model. Results suggest that SI is influenced by the coherence of masker modulations and the semantic content. Predictions demonstrate the importance of a short-term analysis and suggest that listeners cannot optimally derive glimpses from a masker.

[1]This chapter is a reformatted version of the manuscript "Binaural masking release in symmetric listening conditions with spectro-temporally modulated maskers", W. Schubotz, T. Brand, and S.D. Ewert, submitted for publication to the Journal of the Acoustical Society of America and currently under revision.

3.1 Introduction

In binaural listening, interaural differences of the sound arriving at the two ears can be used for decoding the listening situation. The listener's head causes a shadowing effect, which results in an interaural level difference (ILD). Furthermore, a delay of the sound occurs at the ear opposite to the sound source, which is called interaural time difference or interaural phase difference (ITD or IPD). Both ILDs and IPDs provide important cues for source localization, distinction of multiple sources, and separation of a target sound from other, spatially distributed background sounds (maskers). Generally, the separation of two auditory signals is easier when they are not placed at the same location, as this leads to different ILDs and IPDs (Cherry, 1953; Bronkhorst and Plomp, 1992). In the case of speech intelligibility (SI), a spatial separation between the target speech and a masker decreases speech reception thresholds (SRTs; see Licklider, 1948; Hawley et al., 2004). The effect of decreasing SRTs due to a spatial separation is termed spatial release from masking (SRM) and is discussed, for example, by Plomp and Mimpen (1981), Bronkhorst (2000), and Hawley et al. (2004). SRM is influenced by many factors including number of interfering talkers, spatial configuration, room acoustics, and similarity of target and masker (Plomp and Mimpen, 1981; Marrone et al., 2008; Brungart et al., 2001).

In a spatially asymmetric setup where the target is positioned in front of and the masker at one side of the listener's head, a SRM of 8 dB (Bronkhorst and Plomp, 1988) or larger (Bronkhorst, 2000) can be found with a stationary masker. An asymmetric setup generally leads to a long-term "better-ear" as the ear opposite to the masker has a better signal-to-noise ratio (SNR) throughout the entire stimulus presentation as a consequence of the head shadow effect. This long-term better-ear listening can explain much of the results found on binaural SI with an asymmetric masker (Edmonds and Culling, 2006; Kidd Jr et al., 2008). In contrast, when maskers are symmetrically placed at either side of a frontal target, there is no long-term better-ear. Nevertheless, a SRM can be observed in such situations: Jones and Litovsky (2011) reported a considerably smaller SRM of $2-3$ dB for a stationary masker, but for speech-like maskers the SRM can still be quite large. Brungart and Iyer (2012) found a SRM of about 6 dB and Marrone et al. (2008) up to 8 dB (12 dB) for frontal presentation of the target and interfering talkers placed at $\pm 15°$ ($\pm 90°$). Hawley et al. (2004) observed advantages in speech intelligibility up to 12 dB when two interferers were presented along with a frontal target signal.

In a symmetric masker setup there are only short spectro-temporal segments, so-called "glimpses" (Cooke, 2006), which provide a favorable SNR in either ear, instead of a long-term better-ear as found in asymmetric masker conditions. Brungart and Iyer (2012) investigated the usage of such glimpses and showed that intelligibility for a diotic signal with simulated "better-ear glimpsing" (via an ideal monaural better-ear mask, IMBM) is similar to intelligibility observed for binaural presentation. Brungart and Iyer (2012) concluded that listeners are able to use better-ear glimpses, even if glimpses

fluctuate rapidly across frequency and the two ears. This suggests that SI with two spatially symmetric maskers can be explained by an optimal glimpsing strategy, e.g., by selecting that ear with the larger SNR in time-frequency segments and allowing for a rapid change between the two ears. In contrast, Glyde et al. (2013) showed that such a glimpsing strategy cannot fully account for binaural performance in conditions with high informational masking (IM). Similarly, a recent study by Best et al. (2015) suggested that, depending on the sentence material, better-ear glimpsing as assessed by the IMBM stimuli is not sufficient to explain observed SRM. Moreover, Best et al. (2015) suggested that some listeners use differences in perceived location in the stimuli to segregate the target from the interfering talkers. Therefore they use information that was not preserved in the diotic IMBM.
Recently, Lingner et al. (2015) assessed spatially symmetric masker conditions with a larger number of interfering talkers and compared their data to predictions of the binaural speech intelligibility model (BSIM, Beutelmann et al., 2010) which combines a short-time binaural equalization-cancellation (EC) model (Durlach, 1963; Wan et al., 2014) with a short-term speech intelligibility index, similar to the extended speech intelligibility index (ESII) by Rhebergen et al. (2006). Lingner et al. (2015) concluded that better-ear glimpsing predicts SI qualitatively better than a purely monaural model approach (long-term better-ear); however, it does not fully account for the empirically observed SRM. Their results point to an important role of IPD/ITD for multiple interfering talkers, which is neglected by a purely SNR-based (thus ILD-related) better-ear glimpsing strategy. In summary, SRM is strongly affected by interaural differences related to the spatial configuration of target and symmetric maskers as well as the masker itself. But the relative role of ILD-based better-ear glimpses and IPD/ITD processing in connection with spectro-temporal and semantic properties of the maskers still remain unclear. Therefore, the current study examines SI in a symmetric masker setup with maskers that systematically vary in their spectro-temporal properties (see also Schubotz et al. (2015) in chapter 4 for monaural release from masking), ranging from a stationary masker to single, interfering talkers. The maskers differ, among other aspects, in their spectral coherence across the frequency spectrum, as this largely influences signal detection (Piechowiak et al., 2007; Dau et al., 2013). Since the target stimuli used in this study consist of entire sentences, the investigation of spectral coherence is not identical to the studies on "classical" co-modulation masking release (e.g., Hall et al., 1984; van de Par and Kohlrausch, 1998; Festen, 1993), where the detectability of a target tone, surrounded by one or more flanking noise bands with co-modulated waveforms, is examined. Nevertheless, spectral coherence is expected to be an influential factor on binaural speech intelligibility. Besides, the applied maskers also differ in their amount of informational masking. SRTs are measured for co-located maskers presented from a frontal position (0°) and for spatially separated maskers at +60° and −60° in the horizontal plane, while the target speech is always presented from the front (0°). Virtual acoustics using anechoic head-related transfer functions (HRTFs, Kayser et al., 2009) is used, as this avoids the effect of head movements and provides the

opportunity to manipulate HRTFs in such a way that only ILD or IPD cues are preserved. In addition, SRTs are measured for an unnatural condition with independent masker sequences at either ear ("infinite ILD") and for a diotic presentation based on an IMBM to simulate an optimal better-ear glimpsing strategy, while removing spatial cues during playback. Observed SRTs are compared to BSIM predictions, which allows an analysis of the better-ear glimpsing strategy and an analysis based on ILD and IPD cues separately. Moreover, contributions of individual model stages, such as a long- or short-term SNR analyses or the EC-stage, to the overall predicted SI are analyzed.

3.2 Methods

3.2.1 Listeners

Eleven listeners (six male, five female) aged $16 - 29$ years (mean 23.2 years) participated in the measurements on speech intelligibility. All had audiometric thresholds of 20 dB HL or better at octave frequencies between 125 Hz and 8 kHz, except for one person who had a threshold of 25 dB at 750 Hz. The listeners were naïve to the target material and received an hourly compensation for their participation.

3.2.2 Apparatus & procedures

SRTs were measured using the German Oldenburger Satztest OLSA (Wagener et al., 1999) with an adaptive procedure (Brand and Kollmeier, 2002) to determine the SNR at which 50% of the presented words in a sentence were understood correctly. Measurements were performed as an open test, i.e., the listeners were instructed to orally report the understood words to the experiment leader and were allowed to guess. No feedback was provided. The level of the target sentences was varied during the measurements while the level of the masker was fixed at 65 dB SPL (a single stationary masker at 0° had a level of 65 dB SPL in the acoustic coupler). The stimuli were presented via Sennheiser HD 580 headphones, with a flat frequency weighting, that were calibrated with a Bruel&Kjær artificial ear (4133). A list of 20 target sentences was used to estimate the SRT in each masker condition. Measurements for the six different spectro-temporal maskers of the current study were interleaved. The order of the presentation of the different HRTF conditions was Latin-Square balanced to prevent learning effects. To familiarize with the speech material and task, two lists of 20 sentences each were presented prior to the actual measurements to the listeners in a procedure that converged at 80% SI. The masker in the training setup was a cafeteria noise placed at $\pm 60°$, which is an unrealistic masker setting, but was used to retain the masker types of the current study for the actual SRT measurements. All target sentences were embedded in a short (5 seconds), random sequence of the different maskers. All stimuli were presented binaurally and the measurements took place in a double-walled, sound-attenuating booth. The sampling frequency of the stimuli was 44.1 kHz.

3.2.3 Stimuli

Target speech material

The OLSA provided a large number of test sentences that had a correct grammatical structure but no semantical predictability, as all sentences were constructed from a total of 50 words with ten words for each word type (name-verb-numeral-adjective-object). The sentences were spoken by a male talker with only a light accent and had a mean fundamental frequency of 110 Hz. The target speech was always presented from the front and therefore convolved with a head-related transfer function (HRTF) that represented an angle of 0° in the horizontal plane. The HRTFs were the in-ear microphone recordings from the database of Kayser et al. (2009) and were recorded in an anechoic environment with the Bruel&Kjær head and torso simulator (HATS, type 4128C). The distance to the speaker was 80 cm.

Maskers

Spectro-temporal masker types Six different masker types were used in this study, two speech-like maskers and four maskers that were based on a stationary speech-shaped noise (SSN) as previously used in a monaural study on SI (Schubotz et al., 2015). Spectrograms of 2.5 s long sequences of all masker types (for a single masker sequence located at 0°) are shown in Fig. 3.1. The original masker signals were between 20 s and 60 s long.
The two speech-like maskers were a male version ($\text{ISTS}_{\text{male}}$) of the International Speech Test Signal (ISTS; Holube et al., 2010) and a male single talker from Hochmuth et al. (2014). The original ISTS consists of intact continuous speech, uttered by six different female talkers in different languages. Recordings of these talkers were cut into short time fragments and recomposed to form the final ISTS signal. Thus, ISTS offers all major characteristics of speech, can be recognized by humans as being composed out of real speech, but is largely not intelligible (Holube et al., 2010). The current $\text{ISTS}_{\text{male}}$ was generated using the STRAIGHT software package (Kawahara et al., 2008). The fundamental frequency of the ISTS was lowered to match that of the male OLSA speaker (F0 = 110 Hz) and the vocal tract length of the original female speakers was extended by 25% within STRAIGHT. The single talker (ST) masker was composed of continuous OLSA sentences uttered by a male speaker that was not the original OLSA speaker.
All four SSN-based maskers were derived from the $\text{ISTS}_{\text{male}}$ stimulus by a Fast Fourier Transformation and randomization of the phases. Thus, $\text{ISTS}_{\text{male}}$ and all SSN-based maskers had the same long-term amplitude spectrum. These maskers were used:

i) Next to the basic stationary SSN, three amplitude modulated SSN maskers were generated.

ii) An 8-Hz sinusoidal amplitude modulation (SAM) with 100% modulation depth was applied to the SSN. This condition was termed SAM-SSN.

iii) The SSN was multiplied with the Hilbert envelope of a broad-band speech signal, introducing temporal gaps that reflect the modulations of intact speech. The underlying broad-band speech signal was composed of a sequence of ten randomly selected OLSA sentences from the target material. Temporal gaps between and within sentences were shortened to approximately 150 ms. The Hilbert envelope was low-pass filtered to 64 Hz with a 4^{th}-order Butterworth filter. This masking condition was termed broadband (BB-) SSN. The amplitude modulations in the SAM- and BB-SSN were applied to the entire frequency range of the maskers, yielding modulations that were coherent across all auditory channels (co-modulation). However, only the SAM-SSN condition provided temporally regular modulations.

iv) Finally, to provide modulations that were not co-modulated across the masker spectrum, an across-frequency shifted SSN masker (AFS-SSN) was created. This was done by filtering the SSN into 32 auditory channels within a frequency range of 50 Hz−12 kHz with a 4^{th}-order Gammatone filter bank with a spacing of 1-ERB (equivalent rectangular bandwidth; Glasberg and Moore, 1990). Afterwards, four adjacent channels were modulated with the same envelope, which was a random part from the same low-pass filtered Hilbert envelope used for the BB-SSN. Overall eight different randomly time-shifted modulations were applied to eight groups of four adjacent channels. As a consequence, coherent modulations in the AFS-SSN were introduced only in those parts of the masker that belonged to the four adjacent auditory filters. Amplitude modulations were otherwise incoherent across the entire masker spectrum. All masker types were normalized to remove level differences that might have been created during the spectro-temporal manipulation.

Spatial masker configurations Two spatial masker configurations were used: in the reference condition, two masker sequences were co-located with the target at 0° (termed co-located maskers), and in the spatially separated symmetric condition (termed spatially separated maskers) one masker sequence was presented at a positive azimuth angle (+60°) and one at a negative azimuth angle (−60°), with the target direction (0°) as a reference. As for the target, the according anechoic HRTFs from Kayser et al. (2009) were applied for the headphone virtualization. The two masker sequences of 5 sec (either both at 0° or ±60°) were always statistically independent realizations of the same masker type, randomly cut from the longer masker signal, but separated by at least 2 seconds to avoid temporal overlap. For the SAM-SSN, the two masker sequences were selected to always have a fixed 90° phase-shift of the SAM. This was done to avoid large variations in the SRT data due to the high probability of in-phase and anti-phase envelopes. The convolution with the HRTFs lead to slightly different masker levels at 0° and ±60°.

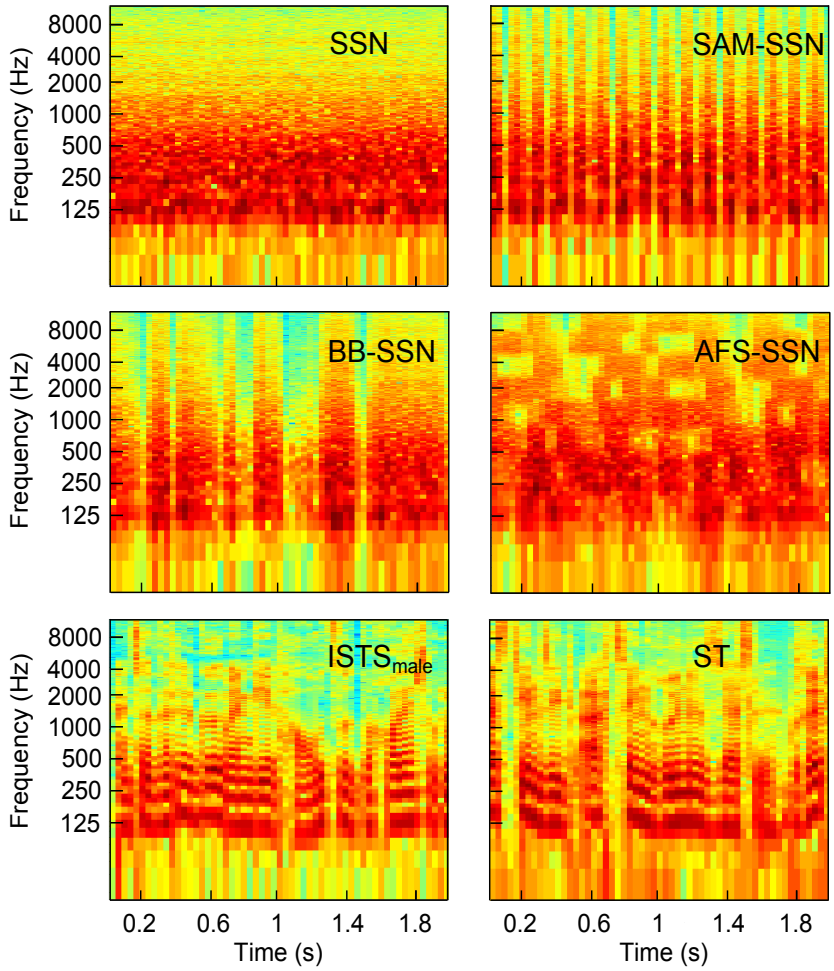


Figure 3.1: Spectrograms for the six masker types used in the current study. The upper four panels show the SSN-based maskers, the lower two panels the speech-like maskers. The upper left panel shows the spectrogram of the stationary speech-shaped noise (SSN) that was derived from $\text{ISTS}_{\text{male}}$, the male version of the ISTS (Holube et al., 2010), by a FFT and randomization of the phase information. $\text{ISTS}_{\text{male}}$ was generated with the STRAIGHT algorithm (Kawahara et al., 2008) by lowering the mean fundamental frequency of the original ISTS speech material and extending the vocal tract of the original speakers as to match the F0 from the target material (OLSA, Wagener et al., 1999). The upper right panel shows the SAM-SSN, which is the SSN fully amplitude-modulated with an 8-Hz sinusoid. The middle left panel shows the BB-SSN, which is the SSN modulated with a broad-band speech envelope of a single talker. The middle right panel shows the (across-frequency shifted) AFS-SSN that results from imposing different parts of the broad-band speech envelope (also used for the BB-SSN) onto four adjacent filters with which the SSN was analyzed. The SSN was analyzed with 32 channels in total, resulting in eight different envelope sequences altogether. The lower left panel shows the spectrogram of the $\text{ISTS}_{\text{male}}$. The lower right panel shows the spectrogram of a single male talker (ST).

HRTF conditions For all HRTF conditions, the original head-related impulse responses (HRIRs) from Kayser et al. (2009) were shortened in the time domain to reduce pre-echoes from the HRTF recordings. The original HRIRs were multiplied with a window that corresponded to a rectangular window until 100 samples after the HRIR's maximum followed by a cosine-square offset ramp of 135 samples. The entire HRIR was then zero-padded to a length of 4422 samples. After this windowing the HRIR was Fast Fourier-Transformed (FFT), manipulated in the frequency domain and transformed back to the time domain by inverse FFT. The resulting HRIR was circularly time-shifted to have the maximum in the center and was multiplied with a Tukey window of 4422 samples to remove potential artifacts of the processing. The resulting change of spectrum was moderate: at 110 Hz (mean fundamental frequency of the target material) the level was decreased by only 1.6 dB.
For the first HRTF condition, termed $\mathrm{HRTF}_{\mathrm{full}}$, the shortened HRTFs were used to provide the original interaural level and phase differences, thus ILDs and IPDs were provided within the masker as they naturally occurred in the spatial configurations of 0° and ±60°.
For the $\mathrm{ILD}_{\mathrm{only}}$ condition, the original IPD was eliminated (i.e., the phase was set to zero in the frequency domain), but the original amplitude spectrum of both ears was left intact. Thus, the original ILDs were the only interaural information that was left in the $\mathrm{ILD}_{\mathrm{only}}$ condition.
For the HRTF condition where only the original IPD was preserved in the masker, the amplitude spectrum of the original HRTFs was replaced by the mean magnitude spectrum of both ears at either 0° or ±60°. This resulted in two different HRTF conditions with identical IPDs, but different amplitude spectra. The two conditions were termed $\mathrm{IPD}_{\mathrm{mag0}}$ and $\mathrm{IPD}_{\mathrm{mag60}}$, whereas the numbers refer to the direction from which the mean magnitude spectrum was derived. The differences in the magnitude spectra led to a difference in spectral coloration of the two conditions. All above-mentioned HRTF conditions were generated prior to the measurements. They resulted in masker signals that showed a certain correlation at both ears as the two statistically different masker sequences were mixed in the signals of the left and right ear.
In the final HRTF condition, an "infinite ILD" was applied such that each of the two masker sequences was mapped to only one ear. In this case, the left and right channels of the 0°-HRTF were used for the two independent masker sequences in the respective ear. The two independent masker sequences were cut out randomly from the two HRTF channel signals. This resulted in an artificial listening condition with completely uncorrelated masker sequences in each ear, referred to as independent $\mathrm{HRTF}_{\mathrm{full}}$ condition. This condition was generated during the measurements.

Ideal monaural better-ear masker (IMBM) The above-mentioned HRTF conditions always resulted in a dichotic binaural signal for the spatially symmetric masker conditions. For a diotic presentation, in which binaural cues were removed, an IMBM was calculated from both mixed ear signals as done by Brungart and Iyer (2012). The IMBM mimicked an optimal better-ear

glimpsing strategy and provided only those spectro-temporal segments of the maskers from either ear that would lead to a favorable SNR. According to the procedure in Brungart and Iyer (2012), an IMBM was created for both spatial configurations of the $\text{HRTF}_{\text{full}}$ condition, as well as the independent $\text{HRTF}_{\text{full}}$ condition. The two ear signals, originating from the two masker sequences (disregarding the target), were analyzed with a 4^{th}-order Gammatone filter bank including delay compensation (Hohmann, 2002) in 37 channels in the range of 50 Hz − 16 kHz (one filter per ERB). Then the signals were separated into 20 ms Hann-windowed time frames with 50% overlap, which resulted in a time-frequency representation of the signal. This was slightly different to the IMBM generation in Brungart and Iyer (2012), where a 128-point Gammatone filterbank was used that had 2048-pt FIR filters equally spaced (every 0.2 ERB) on an ERB scale from $80 - 5000$ Hz. For each time-frequency frame it was then calculated whether the left or the right ear signal contained less masker energy. The frames with lower masker energy were added to create a mono signal, the IMBM. The IMBM processing was performed for the complete binaural masker sequences of 60 seconds prior to the SRT measurements. During the experiment random parts were cut out from the IMBM and presented diotically to the listeners. To avoid any effect of remaining artifacts due to the Gammatone filterbank analysis and resynthesis, the target material was also analyzed and resynthesized in the same way prior to the measurement (skipping the better-ear selection). These modified target sentences were presented whenever the IMBM processed masker was used. There was also an independent IMBM, which was generated prior to the measurements by taking the independent $\text{HRTF}_{\text{full}}$ masker and applying the IMBM processing. The result was again a diotic signal. The two independent (independent $\text{HRTF}_{\text{full}}$ and independent IMBM) maskers were used to investigate how well listeners could perceive the target speech when the glimpsing process was either realized in the signal generation (independent IMBM) or by exploiting binaural differences (independent $\text{HRTF}_{\text{full}}$ masker).

3.2.4 Models

Speech intelligibility predictions were performed using one long-term SNR-based approach, three different BSIM versions, and one stimulus enhancement prior to the BSIM analysis. Further model versions are presented in appendix C.

Long-term SNR at eardrum

The SNR_{long} analysis was based only on the calculation of the long-term SNR (i.e., the SNR that is present when the entire masker sequence is used in the signal analysis) at the eardrum of both ears of the listener. This was done to estimate the amount of masking that is caused by the long-term energy of the different masker types. Since all situations tested in this study were symmetrical, no significant differences between the long-term SNRs at the two ears were expected to occur. Moreover, given that the six maskers were designed to have a similar overall power spectrum it was

expected that $\mathrm{SNR}_{\mathrm{long}}$ predictions would be similar across the six masker types. Predictions were calculated for the left and right ear individually, but were almost identical. Thus, results from the left ear only are shown.

Binaural speech intelligibility model

All other SI model predictions in the current study were based on the short-time binaural speech intelligibility model (BSIM) by Beutelmann et al. (2010), which incorporates an equalization-cancellation (EC) stage (Durlach, 1963) and a short-time version of the speech intelligibility index (SII, ANSI, 1997). The model input consisted of the masker signal and the target signal separately. As target, a noise that was spectrally matched to the target speech was used in accordance to the ESII by Rhebergen et al. (2006). The signals in both ears were passed through a Gammatone filterbank with 30 frequency bands (1-ERB spacing) between 146 Hz and 8364 kHz and each frequency band was processed individually by the EC-stage in order to maximize the SNR in each analysis band. The processing was performed in consecutive 23-ms (1024 samples) time frames with Hann windows and 50% overlap. The channel output of the EC-stage was compared to the filter channel output of each ear and the output providing the best SNR (either the output of the EC stage or the left-ear or right-ear signal) was selected for further analysis. In selecting the time-frequency frames from either ear, it was in principle possible to model a glimpsing strategy with BSIM. The resulting time-frame and frequency-dependent SNRs served as an input for the SII that included the speech-in-noise (SPIN) band-weighting of frequencies in the range 146 Hz − 8 kHz (ANSI (1997), Table B.1, rightmost column). The SII was calculated in each time frame and averaged over all time frames of the masker sequence to calculate the final resulting SII. All model parameters were identical to those described in Beutelmann et al. (2010). To derive an SRT estimate, the SII value corresponding to the observed SRT for the co-located SSN masker in the $\mathrm{HRTF}_{\mathrm{full}}$ condition was calculated for the long-term version of the model (see below). This SII value served as reference value ($\mathrm{SII}_{\mathrm{ref}} = 0.2095$). For all other combinations of masker type, spatial configuration, HRTF condition, and model version, the SNR was adjusted to match the model SII output of the reference value. Calculations were performed for ten realization of each masker type in each spatial configuration and HRTF condition. The ten realizations were chosen randomly from the longer masker signals, as was done in the listening tests. The masker sequences used for the model predictions were 3 s long (the duration of the OLSA sentences ranged from $2.3 - 2.7$ s). The predicted SRTs reported in the current study were averages across those ten SRTs.

Three modified BSIM versions were used in the current study and enabled access to different binaural features:

For the $\mathrm{BSIM}_{\mathrm{long}}$ model version, the setup according to Beutelmann et al. (2010) with disabled short-term processing was used. Thus the predictions were based on the entire masker sequence of 3 seconds. $\mathrm{BSIM}_{\mathrm{long}}$ incorporated the EC stage and the speech band-importance function of the SII, which were both not present in $\mathrm{SNR}_{\mathrm{long}}$.

In the BSIM model version, BSIM was used as described by Beutelmann et al. (2010), including the short-time processing.
In $\text{BSIM}_{\text{begl}}$ the EC stage was disabled and the model was thus limited to better-ear glimpsing. Consequently, there was no SNR improvement based on the equalization of IPDs in the masker. But the model could select the better ear in each 23-ms time frame and each auditory filter independently with respect to the best possible SNR.
In the model version ADD, a simplistic binaural processing was realized by adding the left and right ear signals to create a monaural signal before model processing. This was presented to one (monaural channel) of BSIM, thus effectively using the short-time SII backend. In this case interaurally correlated stimulus parts (mainly from the target speech) were enhanced and their SNR improved by 3 dB over uncorrelated stimulus parts (which were only present in the maskers).

3.3 Experimental results

3.3.1 Speech reception thresholds

Figure 3.2 shows the mean SRTs along with the inter-individual standard deviations for co-located (filled symbols) and spatially separated maskers (open symbols) for the various spectro-temporal masker types as indicated on the abscissa. The four panels denote the three HRTF conditions and the IMBM processing.
Panel 3.2a) shows the mean SRTs for the $\text{HRTF}_{\text{full}}$ condition. Considering the SSN-based maskers, the highest SRTs (-8.5 dB and -8.4 dB) were observed for SSN and AFS-SSN, whereas SRTs were about $3-4$ dB lower for SAM- and BB-SSN. The SRT for the $\text{ISTS}_{\text{male}}$ was in a similar range as SRTs of the SSN-based maskers, while the SRT for the ST was about 5 dB higher. SRTs for both speech-like maskers had considerably larger inter-individual standard deviations than those obtained in the SSN-based maskers. For spatially separated maskers (open symbols), SRTs were generally shifted downwards by $5-7$ dB, but the overall pattern was similar to that obtained for the co-located maskers. The largest difference in comparing the SRTs patterns for both spatial configurations occurred for the speech-like maskers, because speech-like masker SRTs were below those of the SSN-based maskers in case of a spatially separated masker, but not for a co-located masker. Again, SRTs for $\text{ISTS}_{\text{male}}$ and ST showed larger inter-individual standard deviations than those of the SSN-based maskers. The SRT for the ST masker was about 3 dB higher than that of the $\text{ISTS}_{\text{male}}$.
Panel 3.2b) shows SRTs for the ILD_{only} condition, which were very similar to those of panel 3.2a), especially for co-located maskers (filled symbols). When target and maskers were spatially separated, SRTs for SSN-based maskers were about $1-2$ dB, SRTs for speech-like maskers about 3 dB ($\text{ISTS}_{\text{male}}$) and 5 dB (ST) higher than in panel 2a). Thus, SI was in general slightly worse for the spatially separated configuration when only ILD information was presented within the masker, compared to the presentation of ILD and

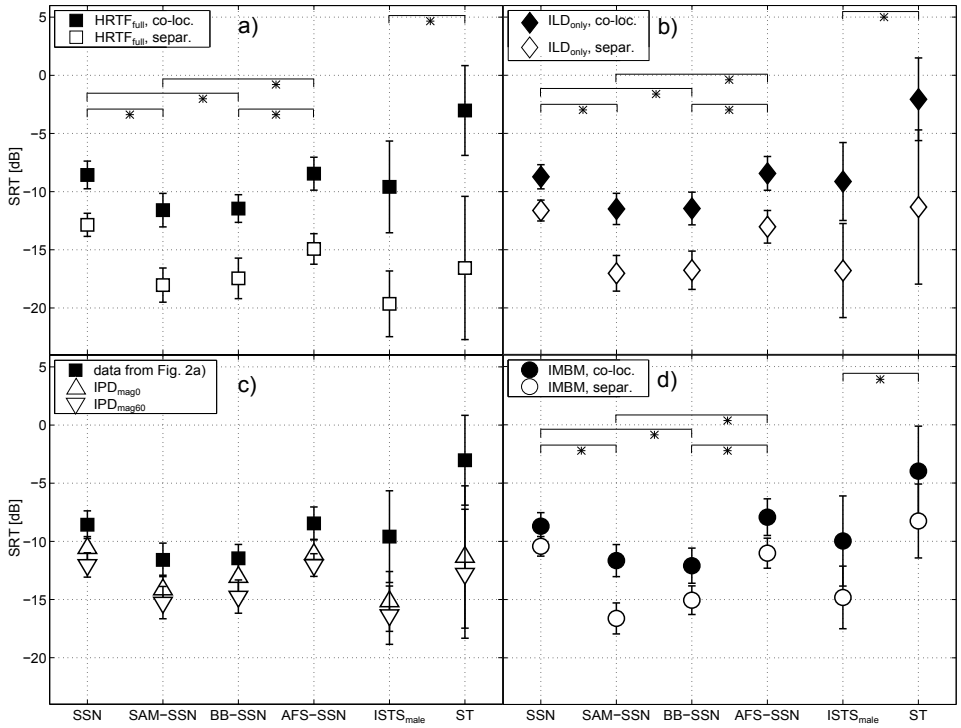


Figure 3.2: Mean SRTs across the eleven listeners together with the corresponding standard deviations. Closed symbols represent the co-located masker position, whereas open symbols represent the ±60° spatially separated masker position. The different masker types are noted at the abscissa, the observed SRTs at the ordinate. Panel a) shows SRTs for the $\mathrm{HRTF_{full}}$, panel b) for the $\mathrm{ILD_{only}}$, and panel c) for the two IPD conditions (see section 3.2). Data from panel a) is re-plotted in panel c) as a reference condition. Panel d) shows the SRTs for the IMBM. Statistically different SRTs within the four SSN-based and two speech-like maskers are marked and are derived from simple pairwise comparisons of the co-located SRTs in each panel. All differences are statistically significant at a level of $p < 0.001$ and marked with *. Statistically different SRTs across SSN-based and speech-like maskers are presented in the text.

IPD cues together ($\text{HRTF}_{\text{full}}$).
Panel 3.2c) shows SRTs observed with the IPD condition. The co-located masker configuration was omitted for this condition, since inspection of the HRTFs indicated only negligible changes to the co-located $\text{HRTF}_{\text{full}}$ condition. Accordingly, SRTs for the co-located $\text{HRTF}_{\text{full}}$ masker from Fig. 3.2a) were re-plotted as a reference in black squares. The general pattern of SRTs for the spatially separated IPD maskers (open triangles, upwards and downwards) was again very similar to the data from panel 3.2a). The difference between the $\text{HRTF}_{\text{full}}$ co-located condition (filled squares) and the IPD_{mag0} condition (upward triangles) was caused by the IPD information alone (given that the magnitude spectrum of the 0° direction was used). SRTs for the IPD_{mag0} masker were about $3-5$ dB higher (SSN-based maskers) and $5-7$ dB higher (speech-like maskers) compared to the spatially separated masker SRTs in panel 3.2a). The difference between the upward and the downward triangles reflects the additional effect of the spectral coloration due to the different amplitude spectra of IPD_{mag0} and $\text{IPD}_{\text{mag60}}$. This caused a further decrease of SRTs by $1-2$ dB.
Panel 3.2d) shows SRTs for the IMBM processing. SRTs for the IMBM derived from co-located maskers (closed symbols) were again similar to those from panel 3.2a), but SRTs for the IMBM derived from a spatially separated masker (open symbols) were higher than the corresponding SRTs from panel 3.2a). This increase in SRTs was $2-4$ dB for the modulated SSN-based maskers, about 5 dB for $\text{ISTS}_{\text{male}}$, and 8 dB for the ST.
To assess the effect of the different HRTF conditions and the IMBM processing on SRTs in the co-located masker setup (all filled symbols in Fig. 3.2), a two-way repeated-measures analysis of variance (ANOVA) was performed. The two main factors were HRTF condition ($\text{HRTF}_{\text{full}}$, ILD_{only}, and IMBM) and masker types (indicated on the abscissa of Fig. 3.2). The ANOVA indicated no significant main effect of HRTF condition [$F(2, 20) = 0.40$, $p = 0.67$], but a significant effect of the masker type [$F(1.56, 15.6) = 75.33$, $p < 0.001$]. The interaction between both factors was not significant [$F(2.89, 28.9) = 1.44$, $p = 0.25$]. The degrees of freedom were Greenhouse-Geisser-corrected for the effects of masker type and the interaction. The ANOVA results showed that the SRT pattern for the co-located masker configuration was not dependent on the applied HRTF condition, but only on the different spectro-temporal masker types. A post-hoc test regarding the main effects was performed by pairwise comparison using Bonferroni correction (confidence level $\alpha = 0.001$) and showed significant differences between the SRTs. All significant differences within the SSN-based and speech-like masker are significant at a level of $p<0.001$ and indicated with * in the corresponding panels of Fig. 3.2. Moreover, there were significant differences ($p<0.001$) between SRTs from the ST and all other SSN-based maskers for each HRTF condition. In contrast, SRTs obtained in the $\text{ISTS}_{\text{male}}$ masker did not differ from those of the SSN-based maskers in any of the HRTF conditions.

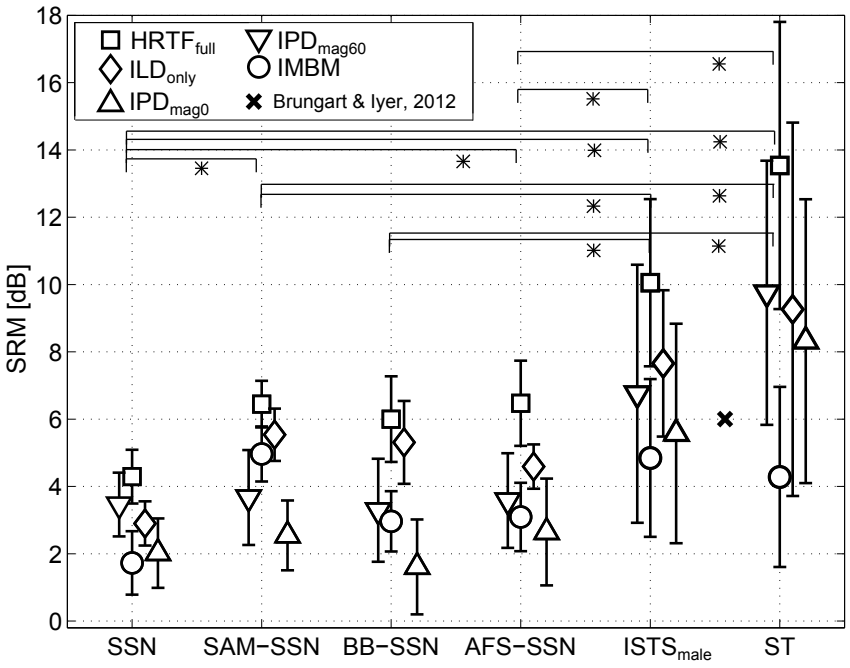


Figure 3.3: Resulting SRM, calculated as the difference between SRTs of the co-located and spatially separated masker position in Fig. 3.2, along with the standard deviations. Larger SRM implies a lower SRT and thus better SI for a spatially separated masker configuration. The different masker types are noted at the abscissa of the figure. The symbols correspond to those in Fig. 3.2, where the associated SRTs are shown. For the two IPD conditions, the SRM was calculated by subtracting the IPD_{mag0} and $\text{IPD}_{\text{mag60}}$ SRT from the co-located $\text{HRTF}_{\text{full}}$ SRT. Results of the statistical analysis are noted in Fig. 3.3 with * (main factor masker type) and in the text (main factor HRTF condition). For reference, the SRM determined in the study by Brungart and Iyer (2012) is shown with a cross.

3.3.2 Spatial release from masking

The resulting SRM is shown in Fig. 3.3 and was determined by the difference between the SRTs for the respective co-located masker configuration and the respective spatially separated masker configuration. The symbols correspond to the HRTF conditions in Fig. 3.2. SRM for the IPD conditions was obtained by subtracting the IPD_{mag0} and $\text{IPD}_{\text{mag60}}$ SRTs from the co-located $\text{HRTF}_{\text{full}}$ SRT (reference in Fig. 3.2c). Additionally, the cross marks the SRM from Fig. 9 in Brungart and Iyer (2012).
Generally, SRM was smallest for the SSN ($2-4$ dB) and larger (up to 6 dB) for the modulated SSN-based maskers. SRM for SAM-, BB-, and AFS-SSN was very similar for each particular HRTF condition. SRM was largest ($5-14$ dB) for the two speech-like maskers and standard deviations were also especially large for these two maskers. Comparing the HRTF conditions and the IMBM processing, the largest overall SRM was observed for $\text{HRTF}_{\text{full}}$, followed by ILD_{only}, and $\text{IPD}_{\text{mag60}}$. SRM was smallest for the IPD_{mag0} condition, except when speech-like maskers were considered. In this case, the smallest SRM occurred for the IMBM. SRM for the IMBM processing was generally smaller than the SRM for the other HRTF conditions, about 2 dB for SSN, 5 dB for SAM-SSN, and about 3 dB for BB- and AFS-SSN. The SRM in the IMBM was about 5 dB for the two speech-like maskers and thus almost identical for $\text{ISTS}_{\text{male}}$ and ST. This was in contrast to the other HRTF conditions, where differences occurred in SRM of the two speech-like maskers.
A two-way repeated-measures ANOVA with the two main factors HRTF condition/IMBM processing and masker type showed significant effects of HRTF condition [$F(2.04, 20.4) = 21.56$, $p < 0.001$] and masker type [$F(1.67, 16.7) = 36.07$, $p < 0.001$], as well as a significant interaction [$F(20, 200) = 4.82$, $p < 0.001$]. The values for the two main factors were Greenhouse-Geisser-corrected and post-hoc tests were performed using pairwise comparison with Bonferroni correction (confidence level $\alpha = 0.01$). All reported differences in SRM were significant at a level of $p < 0.01$ and are indicated (regarding the masker type) with * in Fig. 3.3. Regarding the factor HRTF condition/IMBM processing, the pair-wise comparison revealed significantly different SRM for $\text{HRTF}_{\text{full}}$ and IPD_{mag0} ($\text{IPD}_{\text{mag60}}$, IMBM), for ILD_{only} and IMBM, as well as for IPD_{mag0} and $\text{IPD}_{\text{mag60}}$.

3.3.3 SRTs and masking release for independent maskers in the two ears

SRTs for the independent maskers in both ears are shown in black in Fig. 3.4 and SRTs for the co-located configuration of the maskers (from Fig. 3.3) are re-plotted as a reference with gray symbols. The squares indicate a dichotic presentation using the $\text{HRTF}_{\text{full}}$ condition, the circles indicate a diotic presentation using the IMBM.
The independent $\text{HRTF}_{\text{full}}$ masker (black squares) showed a roughly similar pattern of SRTs as the co-located $\text{HRTF}_{\text{full}}$ masker (gray squares). Again, the standard deviations were largest for the speech-like maskers and particular large for the ST in the independent $\text{HRTF}_{\text{full}}$ condition

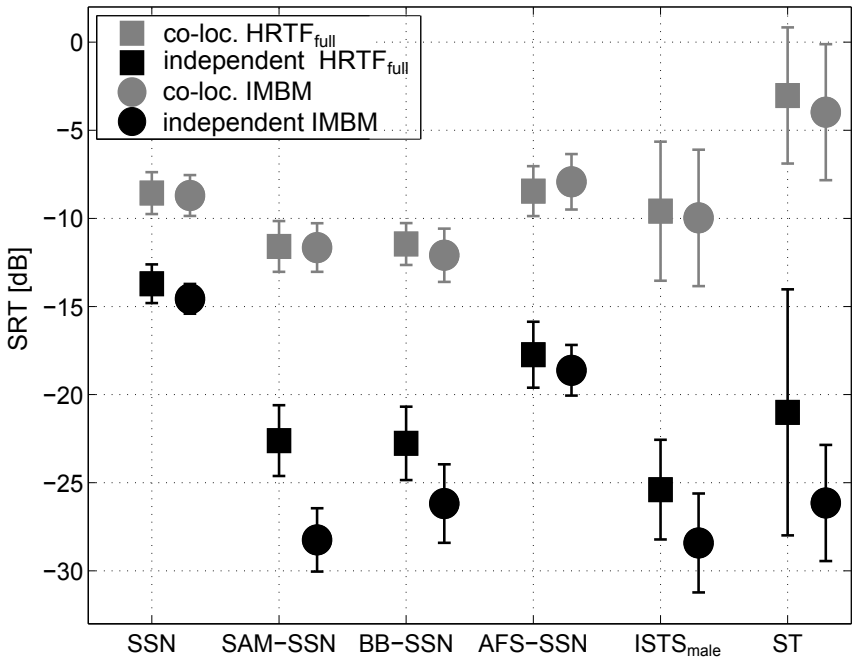


Figure 3.4: Mean SRTs for the presentation of independent masker sequences in the two ears including standard deviations. SRTs from the co-located $\text{HRTF}_{\text{full}}$ masker and IMBM are re-plotted in gray from Fig. 3.2. SRTs from the independent maskers are depicted in black with the same symbols (squares for $\text{HRTF}_{\text{full}}$, circles for IMBM).

(right-most black square). Generally, SRTs were considerably lower for the independent than for the co-located maskers and differences between the masker types were more pronounced (compare gray and black symbols). There was, for example, a difference of 5 dB between the SSN and AFS-SSN SRTs for the independent maskers, which was not present in the co-located masker configuration. Moreover, the independent IMBM SRTs were generally lower (about 1 dB for SSN and AFS-SSN, $3-6$ dB otherwise) than the independent $\text{HRTF}_{\text{full}}$ SRTs (compare both black symbols), while this was not the case for the co-located SRTs (compare both gray symbols). Comparing SRTs from the independent maskers in Fig. 3.4 with those from the spatially symmetric masker configurations (Fig. 3.2), it is apparent that an independence of masker sequences leads to lower SRTs than a spatial separation of target and masker. The differences in SRTs between independent and spatially separated maskers (both $\text{HRTF}_{\text{full}}$ and IMBM) are smallest for the SSN ($3-4$ dB) and largest (up to 20 dB for the IMBM) for ST.

The resulting masking release (MR) for the independent masker sequences was calculated as the difference between SRTs from the respective co-located and independent maskers (gray and black symbols in Fig. 3.4) and was thus not caused by the spatial separation of target and masker. Instead, MR was caused by the independence of the two masker sequences. The MR is

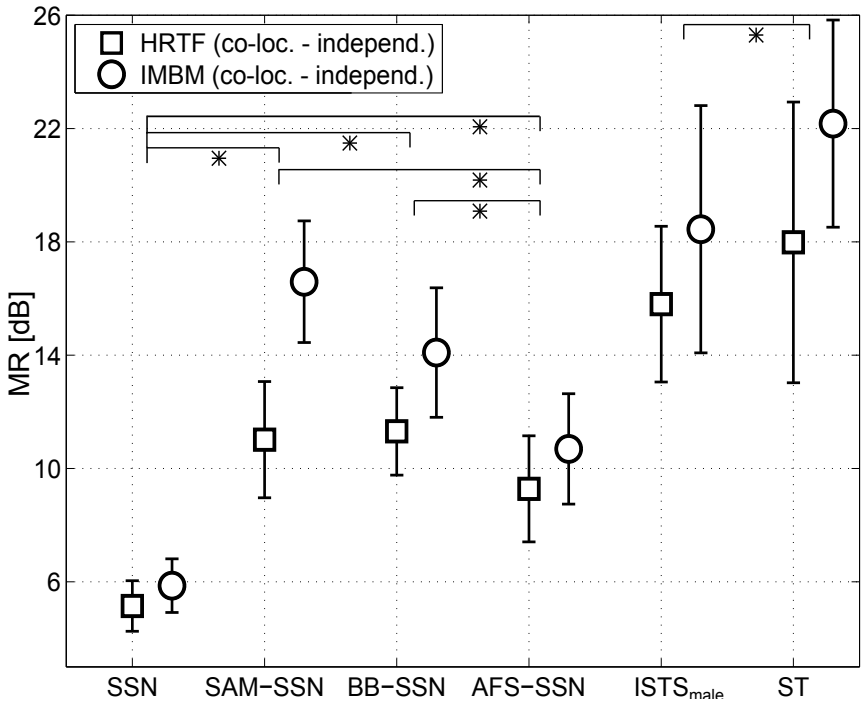


Figure 3.5: Masking release (MR) for the SRTs of Fig. 3.4 with the resulting standard deviations. The symbols are those used in Fig. 3.4, where the corresponding SRTs are shown. Squares denote the MR for the $\text{HRTF}_{\text{full}}$, circles the MR for the IMBM condition. Statistical differences (of main factor masker type) in MR within the SSN-based and speech-like maskers are marked by * and are significant at a level of $p < 0.01$. Significant differences in MR across SSN-based and speech-like maskers are presented in the text.

shown in Fig. 3.5 with symbols corresponding to the HRTF conditions in Fig. 3.4. The MR showed a similar pattern as the SRM in Fig. 3.3, but was generally larger. The smallest MR was found for the SSN masker (6 dB for independent $\text{HRTF}_{\text{full}}$ and independent IMBM) and larger MR was found for the modulated SSN-based maskers ($9-12$ dB for the independent $\text{HRTF}_{\text{full}}$ condition, $11-17$ dB for the independent IMBM). The largest MR was found for the speech-like maskers and accounted to $16-18$ dB for the independent $\text{HRTF}_{\text{full}}$ condition and $18-22$ dB for the independent IMBM. Standard deviations for the MR were largest for the speech-like maskers and the actual MR was larger for the ST than for the $\text{ISTS}_{\text{male}}$. The MR for the independent IMBM was generally larger than the one for the independent $\text{HRTF}_{\text{full}}$ condition, suggesting that listeners could not fully utilize a better-ear glimpsing strategy for independent maskers in both ears. Instead, listeners possibly faced a processing limitation in the extraction of glimpses from the independent $\text{HRTF}_{\text{full}}$ masker. They could not access glimpses in such an optimal way as was provided by the independent IMBM.
A two-way, repeated-measures ANOVA was performed on the MR with the main factors HRTF condition and masker type. There was a significant main effect of HRTF condition [$F(1, 10) = 8.36$, $p < 0.05$] and masker type [$F(5,50) = 210.16$, $p<0.001$], but the interaction was not significant [$F(1.57, 15.7) = 3.71$, $p = 0.57$]. The last value was Greenhouse-Geisser-corrected. Post-hoc pairwise comparison with Bonferroni correction (confidence level $\alpha = 0.01$) showed that all reported differences were significant at a level of $p < 0.01$. The significant differences for masker type within the SSN-based and speech-like maskers are indicated by * in Fig. 3.5. Additional significant differences across SSN-based and speech-like maskers were found for SSN and $\text{ISTS}_{\text{male}}$ and ST, for SAM-SSN and $\text{ISTS}_{\text{male}}$ and ST, for BB-SSN and $\text{ISTS}_{\text{male}}$ and ST, as well as for AFS-SSN and $\text{ISTS}_{\text{male}}$ and ST. Pairwise comparison regarding the MR for independent $\text{HRTF}_{\text{full}}$ and independent IMBM only showed a significant difference for SAM-SSN and BB-SSN.

3.4 Model predictions and comparison to data

To assess potential binaural mechanisms that could explain the observed data, SRTs from Fig. 3.2 were compared to model predictions for the long-term SNR at the eardrum and various BSIM versions. $\text{BSIM}_{\text{long}}$ was matched to correctly predict the SRT for the co-located SSN masker in the $\text{HRTF}_{\text{full}}$ condition, while all other predictions for all model versions were derived with the so-defined reference SII. This section shows model data only for few representative listening conditions. The full set of model predictions (all HRTF conditions and spatial configurations) is shown in chapter C in the appendix.

3.4.1 Long-term and short-term analysis

Figure 3.6 shows SRT predictions for BSIM (left-pointing triangles) for all three HRTF conditions and the IMBM processing in the same format as in

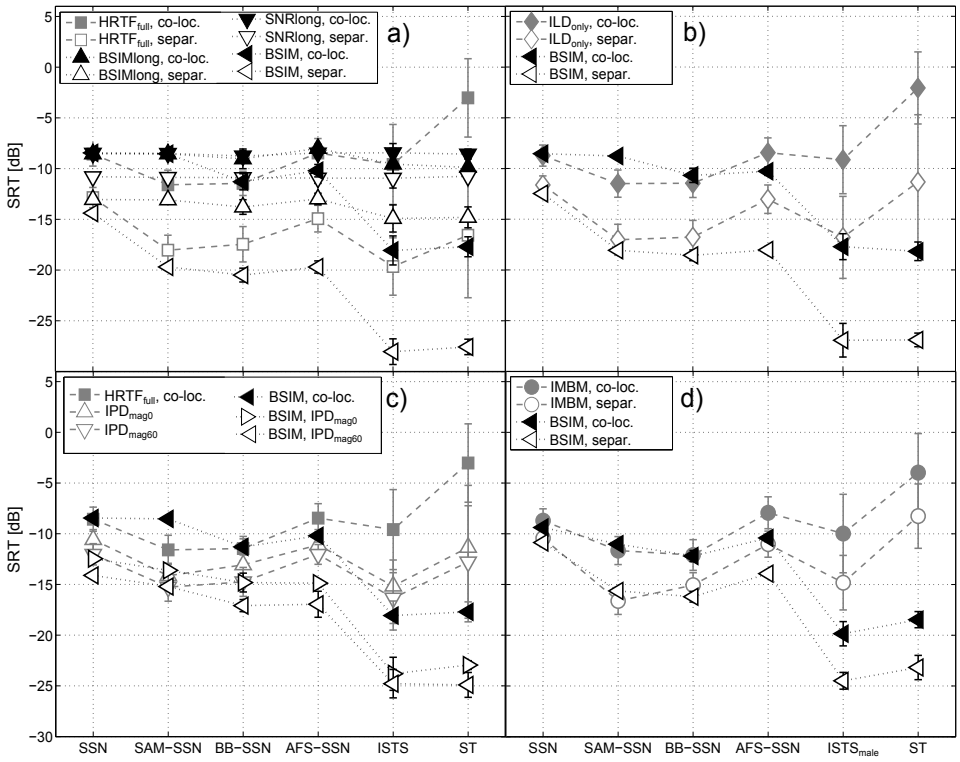


Figure 3.6: Speech intelligibility predictions for BSIM (left-pointing triangles) and the two long-term analysis models $\text{BSIM}_{\text{long}}$ (upward triangles, panel a) and SNR_{long} (downward triangles, panel a). SRTs from the listening experiment are re-plotted in gray, model predictions in black. Closed symbols denote the co-located, open symbols the spatially separated masker configuration. The individual data points are connected with lines to guide the eye. Panel c) also shows BSIM predictions for the reference data (co-located $\text{HRTF}_{\text{full}}$ masker) as black left-pointing triangles. SRT predictions are shown along with the corresponding standard deviations and are averages of predictions for ten different masker sequences of 3 s each.

Fig. 3.2. Individual data points are connected by lines to guide the eye, and filled and open symbols denote the case of co-located or spatially separated maskers, respectively. Observed SRTs from Fig. 3.2 are re-plotted in gray. In addition to predictions by BSIM, panel 3.6a) shows predictions by $\text{BSIM}_{\text{long}}$ (upward triangles) and by SNR_{long} (downward triangles).
As expected, both long-term models could not benefit from temporal gaps in the modulated masker types and showed almost identical SRTs across the different masker types, particularly for the co-located configuration. The simulations thus demonstrated that the original design goal of identical power spectra for all masker types was reached. For the spatially separated masker configuration, there was a more pronounced difference between the two long-term models. The SRTs for SNR_{long} were 3 dB lower compared to the co-located configuration, indicating an effect of spectral filtering by the HRTFs for $\pm 60°$. SRTs for $\text{BSIM}_{\text{long}}$ were about 5 dB lower, which was

most likely a consequence of the speech-in-noise (SPIN) band-importance function. It should be noted that $\text{BSIM}_{\text{long}}$ predicted the SRT for the spatially separated SSN masker perfectly. This indicates that spectral cues, in combination with the SPIN speech band-importance function in the SII, can fully account for the observed SRM in that condition. Predictions for the two long-term models were not shown in the other panels, as results were nearly identical to those of the $\text{HRTF}_{\text{full}}$ masker.

In contrast to the long-term models, predictions with the (short-term) BSIM showed a better agreement with observed SRTs. The short-time analysis enabled a decrease in predicted SRTs for the modulated SSN-based maskers, mimicking so-called "listening in the dips" (Festen, 1993; Bronkhorst, 2000) that occurred for both the co-located and spatially separated masker configuration (open and closed left-pointing triangles). BSIM also predicted a substantial decrease of SRTs for the speech-like maskers, but underestimated the observed SRTs by up to 10 dB. This could indicate that informational masking affected the SRTs in speech-like maskers, but was not captured by the model. This assumption is supported by the fact that predicted SRTs did not differ for $\text{ISTS}_{\text{male}}$ and ST, but that observed SRTs showed significant differences between these two conditions (see gray symbols).

BSIM predictions for ILD_{only} in panel 3.6b) were similar to those in panel 3.6a). Predicted SRTs decreased for the modulated SSN-based masker, and SRTs were slightly underestimated for modulated SSN-based maskers when the masker was spatially separated from the target. SRTs obtained with speech-like maskers were underestimated by $7-10$ dB for both spatial masker configurations.

BSIM predictions for the two IPD conditions are shown in panel 3.6c) for the spatially separated maskers (left and right-pointing open triangles) and for the co-located $\text{HRTF}_{\text{full}}$ (reference) condition (filled black triangles, replot from 3.6a). BSIM predicted lower SRTs for the spatially separated maskers than observed in the listener's data for all six maskers, except for the SAM-SSN. The underestimation of SRTs for the speech-like maskers was about the same as for the other HRTF conditions (panels 3.6a-b, 3.6d). The effect of spectral coloration due to the different mean amplitude spectra was well accounted for in the BSIM predictions of IPD_{mag0} and $\text{IPD}_{\text{mag60}}$. Model predictions differed by 2 dB for these two conditions, as did the observed SRTs from the listening experiment.

A better overall agreement between BSIM predictions and observed SRTs was found for the IMBM in panel 3.6d): SRTs for the SSN, SAM-, and BB-SSN matched very well for the co-located and the spatially separated masker configuration. This indicates that the better-ear glimpses in the IMBM could be fully utilized by the short-time analysis stage in the model. AFS-SSN SRTs were underestimated by about 2 dB and speech-like masker SRTs were underestimated by about 10 dB, as was the case for the other HRTF conditions. It has to be noted, however, that BSIM predictions showed a difference in SRTs for the $\text{ISTS}_{\text{male}}$ and ST, which was also seen in the listener's data (gray symbols).

Summarizing, a short-term analysis was required to predict binaural SRTs

in maskers with temporal gaps and to allow for a prediction of SRTs in the spatially separated masker configuration. Moreover, SRTs in the IMBM were well met, while SRTs for the IPD condition were generally underestimated. SRTs for the speech-like maskers were constantly underestimated, which was most probably caused by informational masking that was not accounted for in the different model versions.

3.4.2 Better-ear glimpsing and interaural summation

SRT predictions for $\text{BSIM}_{\text{begl}}$ (asterix), ADD (plus sign) and BSIM (as a reference with left-pointing triangles) are shown in black in Fig. 3.7, while SRTs from the listening experiment are again re-plotted in gray in the same style as in the previous figures. It should be noted that, for clarity, only those SRT predictions are shown that change in comparison to BSIM. Since interaural cues are negligible for a co-located masker, predictions were the same for BSIM, $\text{BSIM}_{\text{begl}}$ and ADD for this spatial configuration (regardless of HRTF condition) and are thus not shown. Moreover, all model predictions for the IMBM masker were identical to those of BSIM, shown in Fig 3.6d), because the stimuli were diotic and the models thus performed a monaural analysis. Panel 3.7d) was therefore removed from the figure. Finally, predicted SRTs for IPD_{mag0} and $\text{IPD}_{\text{mag60}}$ were always parallel, while SRTs for $\text{IPD}_{\text{mag60}}$ were always 2 dB lower, due to spectral coloration. Therefore, only predictions for the $\text{IPD}_{\text{mag60}}$ masker are shown.
Generally, SRTs predicted by $\text{BSIM}_{\text{begl}}$ (asterix), using only the better-ear glimpsing stage, were shifted about 2 dB for $\text{HRTF}_{\text{full}}$ (panel 3.7a) and 4 dB for the IPD condition (panel 3.7c) to higher SNRs compared to BSIM, but maintained the same overall pattern. $\text{BSIM}_{\text{begl}}$ predictions led to a better agreement between observed and predicted SRTs for the $\text{HRTF}_{\text{full}}$ condition, to a slight change in SRTs in the ILD_{only} condition, and to an overestimation of SRTs for SAM-SSN and BB-SSN in the $\text{IPD}_{\text{mag60}}$ condition. The SRTs for the SSN and AFS-SSN in the $\text{IPD}_{\text{mag60}}$ were met quite well with $\text{BSIM}_{\text{begl}}$. It is apparent that predicted SRTs for $\text{BSIM}_{\text{begl}}$ were very similar for $\text{HRTF}_{\text{full}}$ and ILD_{only}. This was expected, given that the additional IPDs in the $\text{HRTF}_{\text{full}}$ masker cannot be exploited in $\text{BSIM}_{\text{begl}}$ and that the ILD information is identical in $\text{HRTF}_{\text{full}}$ and ILD_{only}. Likewise, ILDs and consequently glimpsing cues, are absent in the $\text{IPD}_{\text{mag60}}$ condition. When comparing SRTs predicted by $\text{BSIM}_{\text{begl}}$ and BSIM for the $\text{IPD}_{\text{mag60}}$ masker, SRTs predicted by $\text{BSIM}_{\text{begl}}$ then move to higher SNRs (lower SI), as this model version cannot utilize IPDs in the EC stage. Altogether, $\text{BSIM}_{\text{begl}}$ can well account for observed SRTs when ILD information is presented in the masker, but overestimates SRTs when IPD information is present. This suggests that IPD information is used by the listeners in binaural listening conditions and needs to be incorporated in the model analysis.
Predictions by ADD examined to which extent a simple binaural summation, enhancing correlated signal parts, could explain the observed data. ADD predictions showed the same pattern of SRTs as BSIM for all HRTF conditions, but predicted SRTs were about 5 dB higher for $\text{HRTF}_{\text{full}}$ and ILD_{only}, except for SSN. In contrast, SRTs were nearly identical for ADD and BSIM in the

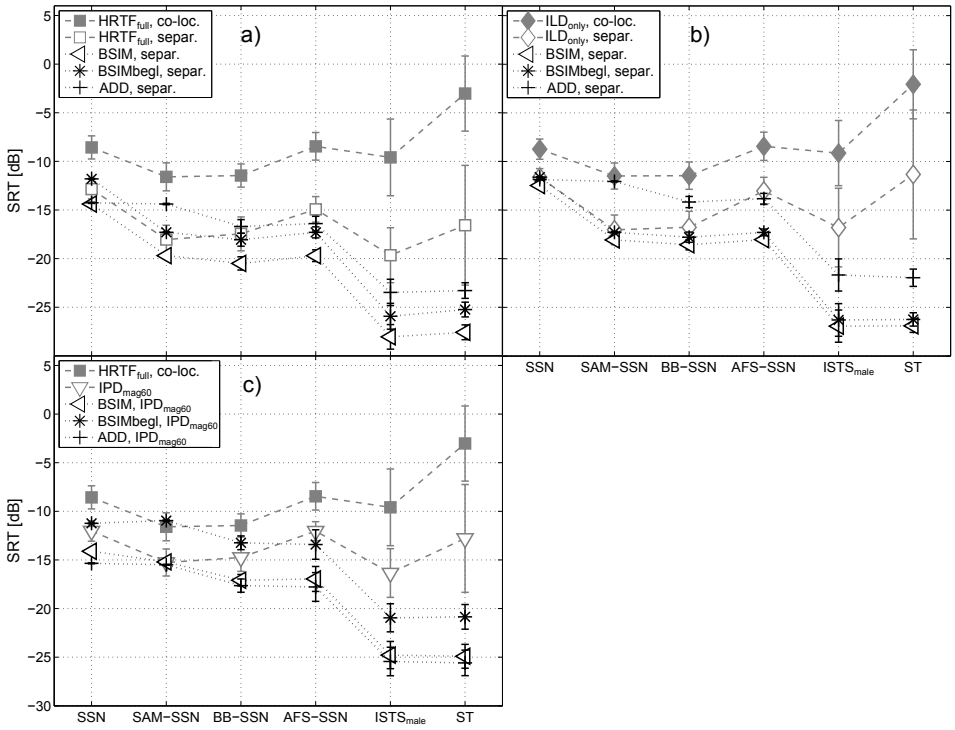


Figure 3.7: Predicted SRTs and standard deviations with the $\text{BSIM}_{\text{begl}}$ and ADD model versions and BSIM predictions as a reference. Data from the listening experiment are re-plotted in gray with the corresponding symbols, model SRTs are depicted in black left-pointing triangles (BSIM), asterix' ($\text{BSIM}_{\text{begl}}$), and plus signs (ADD). The different panels show predictions for the case of $\text{HRTF}_{\text{full}}$ (panel a), ILD_{only} (panel b), and the $\text{IPD}_{\text{mag60}}$ (panel c) condition and correspond to the order in which the observed SRTs were presented (Fig. 3.2). Model predictions are shown for the spatially separated masker configuration only, because predictions for $\text{BSIM}_{\text{begl}}$ and ADD did not change for the co-located masker SRTs compared to BSIM predictions. Model predictions for the IMBM are not shown, because predictions are identical for BSIM, $\text{BSIM}_{\text{begl}}$, and ADD for this certain masker.

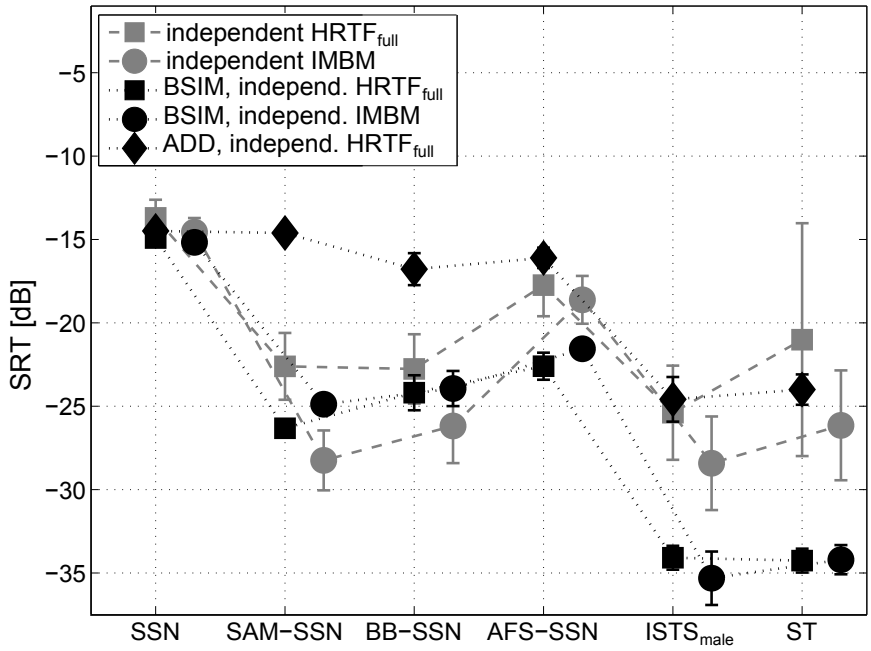


Figure 3.8: Predicted SRTs and standard deviations for the case of independent masker signals in both ears. The figure shows the predictions by BSIM and ADD for the independent HRTF$_{full}$ and independent IMBM condition with black symbols and the corresponding SRTs from the listening experiment (from Fig. 3.4) with gray symbols. The predictions for ADD in the IMBM case are omitted, because they show the same pattern as the predictions of BSIM for this certain condition.

IPD$_{mag60}$ condition (panel 3.7c). This suggests that ADD can, to a large degree, account for the use of IPD information as done in the EC stage in BSIM, but provides a much simpler approach to binaural processing of IPD. However, the use of ILD information as presented in HRTF$_{full}$ and ILD$_{only}$ cannot be accounted for by the ADD approach.

3.4.3 Independent masker signals in both ears

Figure 3.8 shows observed SRTs (re-plotted from Fig. 3.4 with gray symbols) and predicted SRTs by BSIM and ADD (black symbols) for the independent masker sequences in both ears. ADD predictions for the IMBM processing are not shown as the masker is diotic and the simulations are therefore the same as for BSIM. Predictions from BSIM$_{begl}$ are omitted, since they did not differ from those of BSIM. Due to the independence of the masker signals in both ears there were no IPDs that could be utilized in the EC stage of BSIM, consequently predictions depended only on better-ear glimpsing which was identical in BSIM and BSIM$_{begl}$.
BSIM predictions generally resembled the observed SRT pattern and were very similar for the independent HRTF$_{full}$ and IMBM. This similarity could be expected, since the IMBM processing already performed more or less the

same glimpsing analysis that was done in the better-ear stage in BSIM (or likewise $\text{BSIM}_{\text{begl}}$). Predicted SRTs corresponded to the observed SRTs for SSN, but were underestimated for SAM- and BB-SSN in the case of the independent $\text{HRTF}_{\text{full}}$ masker. This means that BSIM was able to extract glimpses better than human listeners when two independent masker sequences were provided. Thus, the glimpsing strategy in BSIM was more efficient than the glimpsing performed by the listeners. In contrast, predicted SRTs were overestimated for SAM- and BB-SSN in the case of the independent IMBM masker, suggesting that the listeners benefitted more than could be explained by the model. This indicates that the (monaural) processing in BSIM is suboptimal in this case. For the $\text{ISTS}_{\text{male}}$ and ST maskers, predicted SRTs were underestimated by almost 10 dB compared to the observed SRTs, pointing again to a possible informational masking effect that is not captured in BSIM.
If the MR is calculated from the simulations and compared to the observed MR (see Fig. 3.5), it is interesting to note that for $\text{ISTS}_{\text{male}}$ and ST, the predicted MR of $16-17$ dB is comparable to the observed MR found in the independent $\text{HRTF}_{\text{full}}$ masker. This is different for the independent IMBM. In that case, the predicted MR is about 3 dB lower for $\text{ISTS}_{\text{male}}$ (5 dB lower for ST) than the observed MR in Fig. 3.5. This indicates that BSIM cannot fully exploit the speech information provided in the independent IMBM.
In contrast to BSIM, SRT predictions by ADD differed substantially from the observed SRTs for SAM-SSN and BB-SSN. Predicted SRTs were generally overestimated by $5-7$ dB for coherently modulated maskers. The overall agreement between SRTs predicted by ADD and observed SRTs was better for the AFS-SSN and the speech-like maskers.

3.5 General discussion

3.5.1 Role of spectro-temporal masker type

Observed SRTs are similar for the SSN and AFS-SSN, presumably because for both maskers there is masker energy present in one of the frequency bands at every instant in time. The AFS-SSN bridges the gap between a modulated and a stationary background noise and suggests that there is a transition between the two when modulations become incoherent. SRTs in the current study decrease as coherent modulations across frequency are introduced to the masker. But SRTs are not affected by the temporal regularity of the modulations (compare SAM- and BB-SSN SRTs). In studies on monaural speech intelligibility, such masking release is often explained by "dip listening" (Festen and Plomp, 1990) and can be related to co-modulation masking release (CMR; Hall et al., 1984; van de Par and Kohlrausch, 1998; Festen, 1993), which occurs when listeners are able to pick up speech in the temporal troughs of a masker. In Hawley et al. (2004), SRTs differ by $2-3$ dB between stationary noise and modulated-noise interferers due to dip listening. In the current study it is found that SRTs for the co-located maskers are about 3 dB lower and SRTs for the spatially separated maskers about 5 dB lower when

modulations are introduced to the masker and findings are thus comparable to Hawley et al. (2004). SRTs decrease although the current masker consists of two sequences and troughs are less likely to occur than for a single-interferer (e.g., Bronkhorst and Plomp, 1992). Altogether, SRTs measured in the current study show a similar behavior as monaural SRTs in Schubotz et al. (2015) using the same target and masker material (see study in chapter 4). However, there are certain deviations from the masker setup in that study. In Schubotz et al. (2015), the masker consists of only one masker sequence, so temporal gaps occur at a slightly higher rate and are longer than those in the two sequence masker of the current study. Thus, the general decrease for the modulated SSN-based maskers is smaller (about 5 dB) in the current study, compared to the masking release found in the monaural study (about 10 dB). This effect is especially apparent for the SAM-SSN, as there is no fixed phase shift of the modulations for the monaural masker, and thus temporal gaps can be fully utilized by the listeners and improve SI. Schubotz et al. (2015) find a significant effect of coherence of the modulations (AFS- versus BB-SSN) on SRTs and a significant effect of the temporal regularity of modulations (SAM- versus BB-SSN). In the current study, the effect of across-frequency coherence is significant and suggests that co-modulation is an influential factor also for binaural SI. In contrast, no influence of temporal regularity is found, which might be related to the superposition of two maskers in the current study. Schubotz et al. (2015) find that SRTs do not significantly differ for SSN and AFS-SSN, as it is also seen in the current study.
The highest SRTs in the current study (even positive for some listeners) occur for the ST masker, and these SRTs are always higher than those of the $\text{ISTS}_{\text{male}}$ masker for all HRTF conditions. Moreover, standard deviations are always larger for the ST than for the $\text{ISTS}_{\text{male}}$ masker. It can be hypothesized that those high SRTs are caused by informational masking (IM), which appears in addition to other masking effects, such as energetic or amplitude modulation masking (Durlach et al., 2003a; Dubbelboer and Houtgast, 2008). This is supported by the reasoning that IM occurs when the masker is similar to the target (Micheyl et al., 2000). This is the case for the ST, given that the masker sentences are from the same speech corpus (OLSA, but with a different speaker) as the target sentences. This presumably enlarges the lexical interference (Hoen et al., 2007; Scott et al., 2004) between target and masker and leads to higher SRTs, as is observed in the data. From Hoen et al. (2007)and Micheyl et al. (2000) it can be concluded that the $\text{ISTS}_{\text{male}}$, although it consists of single interfering talkers, provides less IM than the ST masker. This is supported by the observed SRTs for the co-located maskers that are similar for the SSN-based maskers (supposed to contain no IM) and the $\text{ISTS}_{\text{male}}$.

3.5.2 Spatial release from masking

SRTs for the spatially separated maskers are well below those of a co-located masker, showing a substantial SRM for all masker types and HRTF conditions tested in the current study. Although the general pattern of SRTs is similar for co-located and separated maskers, there are distinct differences in SRM.

In the current study, a SRM of $2-4$ dB for the stationary masker, of $2-7$ dB for the modulated SSN-based maskers, and of $4-14$ dB for the speech-like maskers is found, while the exact numbers vary for the different HRTF conditions. The observed SRM is in line with Marrone et al. (2008), who find a SRM of 12 dB for speech maskers at $\pm 90°$, and Jones and Litovsky (2011), who find $2-3$ dB for modulated and stationary maskers and 9 dB for speech-like maskers at $\pm 90°$. The SRM is reduced to $7-8$ dB in Jones and Litovsky (2011) for speech-like maskers when the masker sequences are positioned at $\pm 45°$. The largest overall SRM in the current study occurs for the $\text{HRTF}_{\text{full}}$ condition, which supports the idea that binaural release from masking is influenced by both, IPD and ILD cues. These two aspects are thought to be additive according to Hawley et al. (1999). For the conditions where either only ILD or IPD information is preserved, the binaural cues are limited and the resulting SRM is indeed smaller. According to Hawley et al. (1999), the head shadow effect only plays a minor role in symmetric masker situations, leaving IPDs as the major influence. The current study, in contrast, suggests that SRM originating from IPD information is generally smaller than SRM from the ILD information (especially for modulated SSN-based maskers). This is supported by work from Culling and Summerfield (1995), showing that listeners are unable to use ITD cues for the segregation of concurrent synthetic vowels. The binaural gain found in that study is mainly caused by the head shadow effect.

Moreover, the current study suggests that the overall SRM found in the $\text{HRTF}_{\text{full}}$ conditions is composed of the three factors IPD, ILD and spectral coloration caused by the different spatial positions of the masker at $0°$ and $\pm 60°$. The release caused by spectral coloration can be estimated from the statistically significant SRT difference for the IPD_{mag0} and $\text{IPD}_{\text{mag60}}$ condition and generally amounts to $1.5-2$ dB. Thus, the SRM in the ILD_{only} condition is actually composed of the interaural component (ILD) as well as the monaural effect of spectral coloration. Subtracting the $1.5-2$ dB due to coloration from the overall SRM for the SSN in the ILD_{only} condition leaves a release of about $1-1.5$ dB that is truly caused by ILDs (see Fig. 3.3). However, for other HRTF conditions, the SRM due to ILDs can be a lot larger. All three components contribute to the SRM for the $\text{HRTF}_{\text{full}}$ condition, which is about 5 dB for the SSN. When adding up the individual contributions (the amount of release due to the IPD information is seen in the IPD_{mag0} and is 2 dB), the SRM for the SSN in the $\text{HRTF}_{\text{full}}$ condition is met, but this is not the case for other masker types and HRTF conditions. This supports the hypothesis of additivity of the effects (Hawley et al., 1999); however, the effects do not necessarily add up linearly.

The overall largest SRM in the current study is found for the ST masker, where IM likely plays a dominant role (see section 3.3.2 and section 3.5.1). This supports the idea that IM is counteracted by spatial separation of target and masker. Jones and Litovsky (2011) show that angular separation has in general a larger influence on SRM for speech maskers than for noise maskers. Besides, they mention that SRM does not necessarily grow with increasing angular separation between target and masker, but is indeed

masker-dependent as found in the current study. Brungart et al. (2001) and Freyman et al. (2008) also show that SRM with speech maskers is larger than with noise masker. They claim that this is especially the case in masking conditions with large IM, i.e. when target and masker are of the same gender or the speech material similar, as is the case for the ST masker. Findings in Kidd Jr et al. (1998) show that large SRM occurs especially for purely informational maskers that are moved in location and Arbogast et al. (2002) report a SRM up to 18 dB for "primarily informational maskers" that are moved from 0° to ±90°. These findings explains well the difference in SRM for the $\text{ISTS}_{\text{male}}$ and ST masker: The current study shows a difference of 5 dB or more between the SRTs of SSN (supposed to provide no IM) and ST (supposed to provide much IM) for co-location of target and masker, which almost vanishes when the masker is spatially separated from the target. Thus, SRM is largest for the ST (and larger than for the $\text{ISTS}_{\text{male}}$), suggesting that indeed the influence of IM is larger for the ST masker, as was already mentioned with regard to the co-located masker SRTs.

3.5.3 Role of better-ear glimpsing and IPD

Comparing co-located SRTs from the $\text{HRTF}_{\text{full}}$ condition with those from the IMBM, it is obvious that SRTs are very similar, which acts as a proof of concept that the stimulus generation with the IMBM introduces only little to no artifacts and does not change the overall frequency spectrum.
However, the situation is different for spatially separated maskers: In this case, SRTs from the IMBM are generally higher than those from the $\text{HRTF}_{\text{full}}$ condition, which is most likely caused by the loss of IPD information in this condition. The resulting SRM is thus generally smaller for the IMBM, especially in the case of speech-like maskers (6 dB compared to 14 dB for $\text{HRTF}_{\text{full}}$), suggesting that an optimal better-ear glimpsing strategy cannot fully explain SRM observed in situations with truly binaural cues. This finding supports the idea that both components, IPD and ILD, are needed for a binaural release from masking (Hawley et al., 1999). The results from the current study are in contrast to those of Brungart and Iyer (2012), who find that better-ear glimpses alone can lead to the same SRTs (SRM respectively) as if full binaural information were available. The current study shows differences up to 10 dB for SRM between the $\text{HRTF}_{\text{full}}$ and IMBM conditions. Moreover, the SRM found with the IMBM is about $1-2$ dB smaller than in Brungart and Iyer (2012), as seen in Fig. 3.3. This is most probably caused by the different target and masker material that was used in Brungart and Iyer (2012). In that study, the target material consisted of sentences from the modified rhyme test (House et al., 1965) and listeners had to mark the target word from a list of rhyming counterparts instead of repeating whole sentences. This reasoning is underpinned by Best et al. (2015), who determined SI with the same and different target and masker material as used in Brungart and Iyer (2012). Best et al. (2015) also find that listeners perform the same, no matter if binaural cues or better-ear glimpses are presented to them, when they use the sentences from the modified rhyme test. In contrast, for a different target and masker material (both being

sentences from the Coordinate Response Measure corpus; Bolia et al., 2000), Best et al. (2015) find differences in SI for the two conditions. They claim that the difference in performance is caused by the amount of IM that is present in the masker. When the masker is low in IM, SI is comparable for better-ear glimpsing and true binaural presentation. In contrast, when the masker is high in IM, listeners fall into two categories: "better" listeners make use of the perceived location of the masker and show lower SRTs in the case of binaural presentation, while "poorer" listeners have no advantage of binaural cues. The findings of the current study support the results of Best et al. (2015) in showing that listeners benefit from a spatial separation of target and masker, which is represented by binaural cues, but absent in the diotic IMBM. Such reasoning is also supported by Lingner et al. (2015), who claim that in a complex cocktail-party-like situation, better-ear glimpsing alone does not lead to the same results as if IPD and ILD information is presented together. Schoenmaker and van de Par (2013) state that ILD cues can aid the listener to make use of the glimpses at the better ear and that ILD cues alone cannot compensate information that is carried in the IPD cues.

3.5.4 Independent maskers in both ears

SRTs in the artificial listening situation with independent masker sequences in the left and right ear, are in general lower than those of the partly correlated maskers (in each ear), resulting from the spatially separated or co-located configuration. The SRT difference between spatially separated and independent maskers is small for the SSN (about 1 dB) and increases for the modulated SSN-based maskers. The difference is particularly large for the speech-like maskers, because they show the highest degree of sparseness in the spectro-temporal domain. Thus, independent masker sequences can lead to very high SNRs in the better ear, in contrast to the spatially separated condition where the SNR in the better ear is limited by "crosstalk" of the masker at the other ear. Moreover, the independent maskers do not produce a natural spatial impression of the masking source, but an artificial maximal separation with each masker apparently positioned at one ear. This might cause the masker not to be perceived as a natural auditory object at a certain spatial position and it might therefore be easier to separate the masker from the target speech. Consequently, this would lead to lower SRTs. Hence, it can be hypothesized that the independence of the masker sequences can also act as a separation cue. Similar to the (partly correlated) $\mathrm{HRTF}_{\mathrm{full}}$ condition (see Fig. 3.2), the MR for the ST is particularly large in the independent $\mathrm{HRTF}_{\mathrm{full}}$ condition. A hypothesized masking release due to the independence of the masker sequences can consequently be expected to be especially large for maskers with a large amount of IM. This is conceptually similar to a release from IM when speech maskers are spatially moved away from the target (Glyde et al., 2013; Arbogast et al., 2002).
For the ST masker in Fig. 3.4, the SRT difference between the independent $\mathrm{HRTF}_{\mathrm{full}}$ and independent IMBM is particularly large (5 dB), showing that listeners perform considerably better if an optimal-glimpsing strategy is ap-

plied already in the IMBM. However, the larger standard deviations for the independent $\text{HRTF}_{\text{full}}$ masker indicates that some listeners perform equally well in the independent $\text{HRTF}_{\text{full}}$ and independent IMBM masker. This suggests that these listeners can perform an optimal-glimpsing strategy by themselves. For the other masker types in Fig. 3.4, the standard deviations are not as large and SRTs of the independent IMBM are generally lower than those of the independent $\text{HRTF}_{\text{full}}$. This indicates that listeners are not able to fully extract the glimpses that arise in the independent $\text{HRTF}_{\text{full}}$ masker.

3.5.5 Informational masking in the model predictions

Schubotz et al. (2015) investigate the effects of the maskers used in the current study and discuss their different masking characteristics, such as energetic, amplitude modulation, and informational masking, with respect to speech intelligibility in monaural listening conditions. From SI predictions by various monaural speech prediction models, Schubotz et al. (2015) conclude that the effect of IM is not incorporated in any of the applied models. Using the extended speech intelligibility index (ESII, Rhebergen et al., 2006), the amount of unexplained IM can be specified to be as large as 10 dB for speech-like maskers. For all BSIM predictions (and the other model versions) in the current study that basically use an ESII-like short-term SII as backend, SRTs for the two speech-like maskers are underestimated. This suggests that IM is not captured by the model versions and could explain the differences between predicted and observed SRTs. The masking caused by IM would then amount to 10 dB for the $\text{ISTS}_{\text{male}}$ and 15 dB for the ST and is comparable to the findings on monaural SI. This discrepancy can be slightly smaller for certain model versions (e.g., $\text{BSIM}_{\text{begl}}$). Another indication that IM is not captured by any BSIM version is the fact that the model predicts the same SRTs for the $\text{ISTS}_{\text{male}}$ and ST for the three HRTF conditions (Fig. 3.7). Thus, the model does not account for the differences in IM for the two speech-like maskers (Micheyl et al., 2000; Hoen et al., 2007; Scott et al., 2004), which is in contrast to the listener's data.

3.6 Conclusions

The current study measured speech intelligibility in various symmetric masker conditions with maskers ranging from stationary speech-shaped noise to single, interfering talkers. Moreover, HRTFs were modified to selectively present different binaural cues to the listeners as well as an artificial "infinite ILD", which was introduced to investigate the influence of independent masker sequences in both ears. A possible optimal "better-ear glimpsing" strategy was investigated by providing glimpses via an ideal monaural better-ear mask. All observed SRTs and the different spectro-temporal maskers used in this study can be downloaded from `http://www.uni-oldenburg.de/mediphysik-akustik/mediphysik/downloads`, providing systematic reference material for further studies with binaural speech intelligibility measurements and speech prediction models. By examining the observed SRTs

and comparing them to model predictions, this study leads to the following conclusions:

1. The different spectro-temporal masker types show that coherent modulations in a masker significantly decrease SRTs in spatial listening conditions. This confirms findings from monaural measurements with the same target and masker material (Schubotz et al., 2015). In contrast to Schubotz et al. (2015), the current results do not show any influence of the regularity of the modulations on observed SRTs.

2. Informational masking can be counteracted to some extent by spatially separating the target from the masker signal. SRTs for speech-like maskers show the largest decrease in SRTs when the maskers turn from co-location to a spatially separated position. Thus, binaural cues are used to solve the problem of object separation that appears especially important for informational maskers.

3. Based on the model results, SRM can be attributed to the three factors ILD, IPD, and spectral coloration (long-term power spectrum changes) caused by the spatial separation of target and masker. For stationary speech-shaped noise (SSN), each factor contributes about equally (2 dB) to the overall SRM. But the effects do not add up linearly, as the SRM for the SSN masker in the $\mathrm{HRTF}_{\mathrm{full}}$ condition is only about 5 dB. A better-ear glimpsing strategy can utilize differences caused by ILD and spectral coloration, but fails to explain the additional benefit, which human listeners gain by exploiting IPDs.

4. BSIM is in general applicable to predict SRTs observed in listening experiments with symmetric maskers. An analysis within short time frames allows for a correct prediction of the influence of (coherent) modulations across the masker spectrum and suggests a short-term analysis by the auditory system. However, for speech-like maskers, the masking is generally underestimated, which is possibly caused by the influence of informational masking. In comparing observed SRTs with those predicted by BSIM, it is generally possible to estimate the influence of informational masking. This accounts to 10 dB in the current study.

5. Independent masker sequences in both ears demonstrate the limits of an optimal better-ear glimpsing strategy in humans. SRTs are generally lower for the case of the independent IMBM than for the case of the independent $\mathrm{HRTF}_{\mathrm{full}}$ condition, indicating that glimpses cannot be optimally extracted by the listeners. This might be caused by processing limitations within the human auditory system that result in a maximal SRM of about 12 dB in situations, where informational masking is thought to be low or absent.

3.7 Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich “Das aktive Gehör” (DFG SFB/TRR31).

Chapter 4

Monaural speech intelligibility and detection in maskers with varying amount of spectro-temporal speech features

Abstract[1]

Speech intelligibility (SI) is strongly affected by the presence of maskers. Depending on the spectro-temporal structure of the masker and its similarity to the target speech, different types of masking can occur which are typically referred to as energetic, amplitude modulation, and informational masking. In this study, SI and speech detection was measured in maskers that vary systematically in the time-frequency domain from steady-state noise to a single interfering talker. Male and female target speech was used in combination with maskers based on the same or different gender. Empirical data was compared to predictions of the speech intelligibility index (SII), extended speech intelligibility index (ESII), multi-resolution speech-based envelope-power-spectrum model (mr-sEPSM), and the short-term objective intelligibility measure (STOI). Comparison shows that most masking can be explained by short-term energetic masking (ESII), but that the other types of masking influence SI as well. This is not captured in all models, but still qualitative and quantitative contributions of the individual masking effects can be determined in the current study (e.g., amplitude modulation masking is assigned to be $3-4$ dB). The study provides a systematic investigation of masker features and a database which can be used for further evaluation of speech prediction models.

[1]This chapter is a reformatted version of the manuscript "Monaural speech intelligibility and detection in maskers with various amount of spectro-temporal speech features", W. Schubotz, T. Brand, B. Kollmeier, and S.D. Ewert, submitted for publication to the Journal of the Acoustical Society of America and currently under revision.

4.1 Introduction

Speech is one of the most important ways of human communication. However, speech signals in everyday life are often perceived within a background noise (masker) and speech intelligibility can be severely hampered. In such listening conditions, a number of binaural and monaural signal properties can affect speech intelligibility. While binaural cues depend on the spatial distribution and the stimulus properties of the interfering sounds, monaural cues only depend on the latter. These monaural masker properties can include frequency content, amplitude modulations, and duration of the temporal gaps within the masker (Brungart, 2001). Thus, monaural speech intelligibility in a background noise depends largely on the spectro-temporal structure of the masker, whereas the degree to which speech perception is influenced can hardly be attributed to one single masking effect (Brungart, 2001; Durlach et al., 2003a; Stone et al., 2012). Considering the masking of speech in a background noise (monaural or diotic), several masking effects have been described in the literature, such as energetic masking, amplitude modulation masking, and informational masking. They have been used to describe masked thresholds or to motivate models of speech reception, although their relative role and possible partial redundancies are not entirely clarified yet. Further insight might be provided by systematically assessing the effect of different spectro-temporal masker features on speech intelligibility and speech detection.

Energetic masking (EM) refers to spectro-temporal regions where the noise energy is larger than the target energy (Barker and Cooke, 2007). In this case, the response within the auditory periphery is mainly caused by the masker signal (Stone et al., 2012; Moore and Vickers, 1997; Delgutte, 1990). Thus, EM can be described by the speech-to-noise (or speech-to-masker) ratio at the output of auditory filters (Durlach et al., 2003a). Classically, stationary noise is thought to be only an energetic masker (e.g., Arbogast et al., 2005; Barker and Cooke, 2007), but there are conceptional problems to this assumption. Stone et al. (2012) argued that a stationary noise mainly acts as a modulation masker due to its intrinsic modulations. Following this, only sinusoids that are far enough apart in the spectral domain to avoid beating act as pure energetic maskers. Nevertheless, the classical concept of energetic masking can successfully describe speech reception in stationary maskers, when predictions are based on per-band long-term signal-to-noise ratios (SNRs). Two examples of this analysis are the articulation index (AI, ANSI, 1969) and the speech intelligibility index (SII, ANSI, 1997).

Amplitude modulation masking (AMM) occurs when masker modulations are present (most often in fluctuating maskers) and interact with those from the target signal (Dubbelboer and Houtgast, 2008). Houtgast (1989) and Ewert and Dau (2000) showed in psychophysical experiments that amplitude modulations in the target are masked in an envelope-frequency selective manner, which can be described by the concept of modulation filters. Concerning speech perception, Dubbelboer and Houtgast (2008) proposed a description of AMM by the signal-to-noise ratio in the modulation domain

at the output of auditory modulation filters. In contrast to AMM, coherent across-frequency amplitude modulations (co-modulation) in the masker can reveal entire parts of the target speech. In this case, "dip listening" (Lorenzi et al., 2006; Bronkhorst, 2000) comes into play which is most prominent for low modulation rates (usually below 8 Hz). The observed masking release in fluctuating maskers can thus conceptually be compared to the psychophysical phenomenon of co-modulation masking release (CMR; e.g., Hall et al., 1984), where a release from masking for a pure tone in noise is caused by the coherence of modulations in adjacent frequency bands.

Informational masking (IM) usually refers to masking that does not occur in cochlear processing in the auditory periphery, but in more central regions of the auditory system (e.g., Micheyl et al., 2000; Arbogast et al., 2002; Durlach et al., 2003b). Pollack (1975) described informational masking as the uncertainty in the trial-to-trial variation in the noise waveform in psychoacoustic measurements, whereas for Brungart et al. (2001) the term holds for interfering talkers and speech-on-speech masking when the masker is a "similar-sounding distractor" (e.g., same gender). IM can also be prompted by factors such as speaker spectrum, sentence structure, and semantic content of the target signal, although this is no direct differentiation from the other masking aspects. Generally, IM is also thought to be present when masker and target are similar in terms of temporal coherence and harmonic structure (Micheyl et al., 2000). Lutfi (1990) even proposed a calculation for informational masking, based on the statistical structure of waveforms in a tone detection experiment and found the amount to be about 22% within maskers that are thought to be energetic maskers only. Durlach et al. (2003a), Durlach et al. (2003b), and Lutfi et al. (2013) claimed that two aspects rule informational masking: uncertainty of the masker and similarity between target and masker. These aspects were elaborated on in Lutfi et al. (2013), but they also have an overlap with the definition of EM and AMM by Stone et al. (2012). An alternative definition for IM is to attribute those masking effects to IM that cannot be described by speech intelligibility models that consider EM and AMM. Informational masking is often brought up if the magnitude of intelligibility thresholds cannot be explained by EM and AMM alone. Generally, IM is less clear-cut than the other two masking aspects, but it must be clearly separated from general inattention toward the task (Durlach et al., 2003a).

The concepts of EM and AMM have been successfully used in speech intelligibility models to predict speech reception thresholds (SRTs) in various masking conditions. The AI and SII use band importance functions for the different analysis channels and thus provide a weighted measure of energetic masking. The extended speech intelligibility index (ESII) by Rhebergen et al. (2006) uses the same concept in short-time frames and can therefore cope with fluctuating maskers, when listeners are able to "listen in the dips". The influence of amplitude modulations was first considered in the speech transmission index (STI, Steeneken and Houtgast, 1980), where the reduction of the modulation depth of clean speech due to noise was measured. From that it was assumed that a reduction of the target modulations leads to a

decrease in speech intelligibility. More recently, a sub-band SNR analysis was also used in the envelope domain (Dubbelboer and Houtgast, 2008; Jørgensen and Dau, 2011; Jørgensen et al., 2013), which implemented the concept of AMM in speech intelligibility predictions.
So far, the relative role of the above masking characteristics in arbitrary maskers with different spectro-temporal features is still unclear. The comparison of data from listening experiments and SI model predictions could therefore help to shed light on the different masking characteristics, since models analyze certain characteristics only. Furthermore, a comparison of data from listening experiments and model predictions can test the validity of the assumed processing stages implemented in the speech prediction models. Another open question is the relation of SI and the masked thresholds of the detection of a speech signal itself. While SI models conceptually relate SI to masked thresholds, it is unclear how SI and masked detection thresholds are related as a function of spectro-temporal masker features in data from listening experiments.
Therefore, the aim of the current study is to assess speech intelligibility and speech detection as a function of the spectro-temporal structure of the masker. A systematic variation of masker properties in the time-frequency domain for same and different gender talkers is used to help understand the relative contribution of the described masking characteristics (EM, AMM, and IM). Monaural SRTs and speech detection thresholds (SDTs) were measured with the same subjects for eight maskers, ranging from stationary noise to single interfering talkers, to systematically assess the role of the different masking effects. The maskers were designed in such a way that specific features in the time-frequency domain were changed separately, while keeping the long-term power spectrum identical. Four maskers were based on stationary speech-shaped noise, introducing different degrees of temporal fluctuations. The other four maskers were intact or noise-vocoded speech to examine the influence of pitch contours, the influence of meaningful versus nonsense speech maskers, and the similarity of target and masker spectra. Male and female target speakers were used in combination with maskers derived from male and female speech. The measured SI and speech detection data were compared to predicted SI using the speech intelligibility index (SII), the extended SII (ESII), the multi-resolution speech-based envelope power spectrum model (mr-sEPSM) by Jørgensen et al. (2013), and the short time objective intelligibility measure (STOI) by Taal et al. (2010).

4.2 Methods

4.2.1 Subjects

Eight listeners, aged 23−34 years, participated in the measurements on speech intelligibility and speech detection. They all had audiometric thresholds of less than 20 dB HL or better at octave frequencies between 125 Hz and 8 kHz. The listeners were naïve to the target material and received an hourly compensation for their participation.

4.2.2 Apparatus & procedures

Speech intelligibility was measured using the Oldenburger Satztest (OLSA, Wagener et al., 1999, Table E.1) with an adaptive procedure to vary the SNR (Brand and Kollmeier, 2002). The SNRs at which 50% and 80% of the presented words in a sentence were understood correctly were determined as speech reception thresholds SRT50 and SRT80, respectively. The noise level was fixed at 65 dB SPL throughout the entire experiment. The measurement was performed as an open test, where listeners repeated the perceived words to the experiment supervisor and were allowed to guess. No feedback was provided. A list of 20 sentences was used for the measurements of the SRTs for each masker. Two lists of 20 sentences each were presented in a cafeteria noise prior to the actual measurements to familiarize the listeners with the task and the speech material. SDTs were assessed using an 1-up-2-down two-interval, two-alternative forced choice method to determine the SNR with 70.7% correct responses on the psychometric function (Levitt and Rabiner, 1967). There were two intervals presented to the listener, one containing a random OLSA sentence embedded in the masker, the other containing the masker only. The noise token varied for each trial, but was the same for the two intervals (half-frozen noise). The listeners had to choose the interval that contained the sentence.
The order in which the eight masker conditions were presented to the listeners was Latin-Square balanced in both experiments (speech intelligibility and speech detection) to control for possible learning effects. The duration of the masker was chosen such that the noise started one second before the target sentence started. SRTs and SDTs were measured once for the combination where target material and masker had similar spectra (male/male, female/female), and twice (test, re-test) for the combination male target and female masker spectrum. The stimuli were presented monaurally to the right ear via Sennheiser HD 580 headphones, which were calibrated with pure tones on a Bruel&Kjær artificial ear, and equalized. The measurements took place in a double-walled, sound-attenuating booth. The sampling frequency of the stimuli was 44.1 kHz. All measurements were performed using the AFC-package for MATLAB (Ewert, 2013).

4.2.3 Stimuli

The stimuli presented in the measurements consisted of OLSA target sentences embedded in a background masker. The OLSA sentence material consists of meaningless, but grammatically correct German sentences that follow a controlled sentence structure (noun, verb, numeral, adjective, object; see table E.1 in the appendix). They were spoken by a male talker with very mild accent. Three gender combinations of target and masker were used: male/female, female/female, and male/male. For the female target sentences, the female version of the OLSA was used (Wagener et al., 2014), where the talker had no accent. The frequency range of the target material was limited to 12 kHz, the sampling frequency was 44.1 kHz. Altogether, eight background maskers were generated such that their spectro-temporal characteristics changed while

maintaining their long-term spectrum. They covered a systematic change from speech-shaped, stationary noise to a single, interfering, talker. The eight maskers can be divided into two groups: four maskers were based on a stationary speech-shaped noise (see Fig. 4.1), the other four were speech-like maskers (see Fig. 4.2).

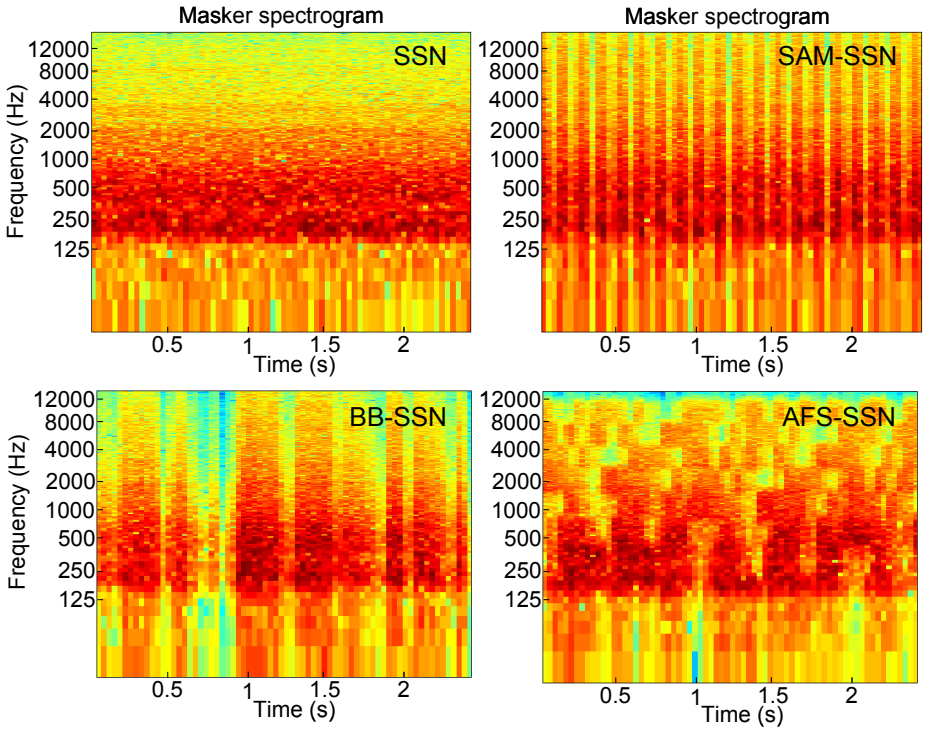


Figure 4.1: Spectrograms of the four SSN-based maskers that are used in the speech intelligibility and detection measurements. The basis is a stationary speech-shaped noise (SSN), which has the exact same long-term energy spectrum as the ISTS (Holube et al., 2010). For the SAM-SSN condition, the SSN is fully modulated with an 8-Hz sinusoid, which results in a regular and coherent modulations. For the BB-SSN condition the modulations are derived from intact broad-band speech (see section 4.2.3 for further details) and are thus irregular, but coherent. For the AFS-SSN, the SSN is split into 32 frequency channels to generate an across-frequency shifted SSN. Four adjacent channels are multiplied with the same sequence from a broad-band speech envelope. The sequence that is used for the next four adjacent channels is another part of the broad-band speech envelope, thus modulations patterns are shifted across frequencies. This results in irregular and incoherent modulations across the spectrum. For the maskers with a male speaker spectrum, the basis is a transformed ISTS that is generated with the STRAIGHT algorithm (Kawahara et al., 2008).

Speech-shaped noise based maskers

The basis was a stationary, speech-shaped noise (SSN, upper left panel of Fig. 4.1) that was derived from the International Speech Test Signal (ISTS, Holube et al., 2010) by a Fast Fourier Transformation, followed by randomization of the phase of the coefficients. For a second masker, the SSN was fully modulated with an 8-Hz sinusoid to introduce regular temporal gaps to the masker. This masking condition was termed SAM-SSN and is depicted in the upper right panel of Fig. 4.1. The SSN was also multiplied with the Hilbert envelope of a broad-band speech signal, introducing irregular temporal gaps that reflect the modulations of intact speech. The underlying broad-band speech signal was a sequence of ten randomly selected OLSA sentences from the male target material. Temporal gaps between and within sentences were shortened to approximately 150 ms. The Hilbert envelope was low-pass filtered to 64 Hz with a 4^{th}-order Butterworth filter. This masking condition was termed BB-SSN (lower left panel of Fig. 4.1). The resulting amplitude modulations in the SAM- and BB-SSN were coherent across all auditory channels (also referred to as across-channel co-modulation in the following). Incoherent amplitude modulations across the auditory channels were introduced in the across-frequency shifted SSN masker (AFS-SSN, lower right panel of Fig. 4.1). This masker was created by filtering the SSN into 32 auditory channels within a frequency range of 50 Hz − 12 kHz, using a 4^{th}-order Gammatone filter bank with 1-ERB (equivalent rectangular bandwidth) spacing of the auditory filters. Four adjacent channels were then modulated with the same envelope. The envelopes were random parts from the same low-pass filtered Hilbert envelope used for the BB-SSN condition. As a consequence, coherent modulations were introduced only in those parts of the masker spectrum that belong to the four adjacent auditory filters. Altogether, eight different randomly time-shifted modulations were applied to the 32 bands, yielding incoherent amplitude modulations across the entire masker spectrum.
Since the basis for these four maskers was the SSN, all had a long-term spectrum of the female ISTS speech. For those experimental conditions where a male masker spectrum was used, the basis was a transformed ISTS. For this, the STRAIGHT algorithm by Kawahara et al. (2008) was used to lower the fundamental frequency (F0) and to lengthen the vocal tract of the speakers of the original ISTS signal in such a way that the mean fundamental frequency of the transformed ISTS resembled that of the original male OLSA material ($F0 = 110$ Hz). The transformed ISTS was then used to derive a SSN with a male spectrum and from that the other three modulated maskers were derived as described earlier.

Speech-like maskers

Four speech-like maskers were generated, two consisted of intact speech and two consisted of noise-vocoded speech (right and left panels of Fig. 4.2). The intact speech maskers were the ISTS (Holube et al., 2010) and a single talker, taken from a study by Hochmuth et al. (2014). The ISTS is a mixture of

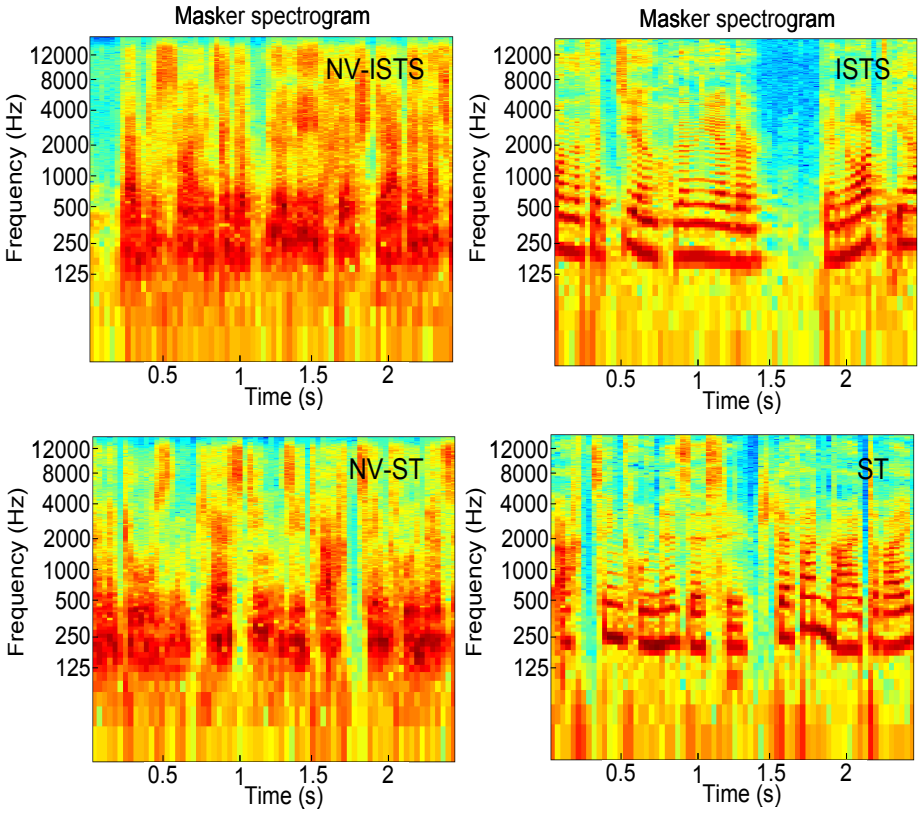


Figure 4.2: Spectrograms of the speech-like maskers ISTS, ST, and their noise-vocoded versions. Noise-vocoding was done with a 32-auditory channel vocoder (based on a 4^{th}-order Gammatone filter bank) within the frequency range of 50 Hz − 12 kHz. The carrier for each band was white noise. Before recombining the individual channels, they were filtered again with the same analysis filters to restrict the filter output to the corresponding frequency region. The spectral weighting of each filter was maintained. For the male maskers, the transformed ISTS and a male single talker were used. Note that the masker sequences in the four panels are not identical.

nonsense speech by six female talkers with different languages. The single talker (ST) masker was a sequence of ten OLSA sentences, spoken by a female talker. In the noise-vocoded masking conditions (NV-ISTS, NV-ST), the fine structure of the intact speech was removed, while maintaining the energy in the individual frequency bands. The vocoding was performed using a 32-channel vocoder (based on a 4^{th}-order Gammatone filter bank) with 1-ERB filtering in the range of 50 Hz − 12 kHz. The Hilbert envelope was extracted from each analysis channel and low-pass filtered to 64 Hz with a 4^{th}-order Butterworth filter. It was then used to modulate random white noise. Before recombining the individual channels, they were filtered with the same analysis filters in order to restrict the filter output to the appropriate frequency range. This was done for the combination of male target and female masker spectrum. For the same-gender combination of target and masker (female/female, male/male), the parameters of the noise-vocoding were slightly altered. There were only 16 channels with 2-ERB spacing and the Hilbert envelope was low-pass filtered to 32 Hz with a second-order Butterworth filter, to further remove possible temporal pitch information. All other parameters were unchanged. The overall power spectrum was the same for the original and the noise-vocoded maskers.
For the case of the male masker spectrum, the female ISTS and the single talker had to be substituted by a male version. Instead of the original ISTS, the transformed version was used and a male talker with no accent from Hochmuth et al. (2014) was taken. The noise-vocoding was performed with the altered parameters.

4.3 Speech intelligibility models

Four speech intelligibility models were used to predict the SRT50 for the same masker conditions as in the listening experiments. Two well established models were the speech intelligibility index (SII, ANSI, 1997) and the extended SII that was proposed by Rhebergen and Versfeld (2005) and Rhebergen et al. (2006). Two more recent models were the multi-resolution speech-based envelope power spectrum model (mr-sEPSM) by Jørgensen et al. (2013) and the short-time objective intelligibility measure (STOI), developed by Taal et al. (2010). For the SII (ESII), the implementations from the Speech Intelligibility Prediction Toolbox (SIP Toolbox, Kollmeier et al., 2011) developed at the Fraunhofer IDMT were used. For the other models, source code was used that was public available (STOI) or provided by the authors (mr-sEPSM).

4.3.1 Speech intelligibility index

The SII standard (ANSI, 1997) is a further development of the Articulation Index standard (AI, ANSI, 1969). In the current study, the signal was split into 21 critical bands and every band was multiplied with a band importance function that weights the contribution of the individual frequency band to the overall intelligibility, as implemented in the SIP toolbox. In the toolbox, the

speech-in-noise (SPIN) frequency weighting function from ANSI (ANSI, 1997, Table B.1, rightmost column) was used. The sum over all frequency bands yields the overall SII for the given speech material in the masker. The input variables for this model were the long-term speech and masker spectrum, thus temporal gaps in the stimulus were not taken into account. Therefore, a stationary noise that was spectrally matched to the OLSA sentence material was used as an input signal instead of the individual target sentences. For the masking noise, the long-term spectra of the eight individual maskers were used. Although the SII provides a measure for speech intelligibility in a given "communication system" (ANSI, 1997), it does not directly predict a speech intelligibility score. In order to predict the SRT50s, the SII that matched the observed mean SRT50 in the SSN masker, was taken as reference SII. Thus, the model outcomes were matched to correctly predict this certain SRT50. All predictions for the other maskers were derived from the reference SII. SII predictions were expected to be sensitive mainly to energetic masking and to be very similar across all maskers, as the long-term spectra of all SSN-based maskers were similar.

4.3.2 Extended speech intelligibility index

The ESII was proposed by Rhebergen and Versfeld (2005) and Rhebergen et al. (2006) as an extension to the SII model in order to account for speech intelligibility in fluctuating maskers. Normal-hearing listeners can well profit from temporal gaps (e.g., Brungart, 2001), but this is not captured in the SII. The extension contains a temporal analysis, where the target and masker signal are segmented into short time frames before further processing. The effect of forward masking is also included (for a detailed description see Rhebergen and Versfeld, 2005; Rhebergen et al., 2006). The model is recommended to use a stationary, speech-shaped noise as the target signal (Rhebergen et al., 2006), instead of real sentences. The current study used the same speech-shaped noise for the ESII, as for the SII. Thus, only the fluctuations of the masker were taken into account by the ESII. Both signals were analyzed within 21 critical bands (see Rhebergen and Versfeld, 2005) and partitioned into time frames, ranging from 35 ms for the lowest band to 9.4 ms for the highest band. Then, the conventional SII was calculated within each frame and the bands were weighted with the SPIN band-importance function from ANSI (1997). The short-term SII values were then averaged to yield the ESII model outputs. To derive actual speech intelligibility predictions from the model outputs, the same procedure and reference condition were used as for the SII predictions. The ESII model was run five times with different sequences from each masker and the predicted SRTs were averaged.
To compare the ESII predictions to a model configuration that takes the actual fluctuations of the target signal into account, an ESII extension by Meyer and Brand (2013) was used (referred to as ESIIsen). Here, the same parameters as for the ESII calculation were applied, the only difference was the input signal: Instead of the spectrally matched stationary noise, intact sentences (sen) from the OLSA corpus were used as input signals. The calculations were performed for twenty sentences, each presented in a

different masker sequence and the model outcomes were averaged.

4.3.3 Multi-resolution speech-based envelope power spectrum model

To assess the role of AMM, the multi-resolution speech-based envelope-power spectrum model (mr-sEPSM) model by Jørgensen et al. (2013) was used. The model is based on the sEPSM (Jørgensen and Dau, 2011), which performs an analysis in the modulation filter domain, but is designed for stationary interferers as it considers the stimulus' long-term envelope spectrum only. Both models are successors of the envelope-power spectrum model (EPSM, Ewert and Dau, 2000), which was developed to account for the aspect of AMM. The mr-sEPSM applies a short-time analysis and is thus designed to predict speech intelligibility in fluctuation maskers. Its core element is a modulation filter bank with nine modulation filters (1−256 Hz) that analyzes the output of auditory filters ranging from 63 Hz − 8 kHz. The auditory filter bank was composed of 22 4^{th}-order Gammatone filters with 1/3-octave spacing, the modulation filter bank was composed of a 3^{rd}-order lowpass filter and six overlapping 2^{nd}-order bandpass filters (see Jørgensen et al. (2013) for further details). The signal-to-noise ratio of the Hilbert envelope was calculated, averaged over time and combined in a root-mean-square manner over all auditory and modulation filters. The resulting $\mathrm{SNR}_{\mathrm{env}}$ value was fed into an ideal observer stage, which converted the $\mathrm{SNR}_{\mathrm{env}}$ to percent correct values (see equations 7 and 8 in Jørgensen et al., 2013). The ideal observer stage took four parameters into account: the value q was thought to be independent of the speech material and following Jørgensen et al. (2013), $q = 0.5$ was used in the current study. The other three parameters, m, σ_s, and k, represented the size of the vocabulary used in the observer stage, a value that was determined by the redundancy of the speech material, and an experimentally determined value that shifted the psychometric function. In the current study, the values were chosen such that the SRT50 in the SSN masker was met for each gender combination of target and masker spectrum. Table 4.1 shows the parameter values for the individual combinations.

The input for this model was the masker alone and a target sentence presented in the same masker sequence. The calculations were performed for a variety of SNRs to receive a psychometric function and to determine the SRT50 value for each of the eight maskers. The calculations were performed for twenty sentences, while each sentence was mixed with a different masker sequence, and the results were averaged. When processed within the mr-sEPSM model, masker and masked speech were down-sampled to 22050 Hz. Unlike in the listening experiments, the signal level was fixed at 65 dB for the model calculations, as done in Jørgensen et al. (2013). However, for the given range of levels, the results were the same as if the target level had been changed.

Gender combination	k	q	m	σ_s
male/female	0.351	0.5	50	0.6
female/female	0.4	0.5	50	0.6
male/male	0.655	0.5	50	0.8

Table 4.1: Parameter values used in the ideal observer stage of the mr-sEPSM.

4.3.4 Short-time objective intelligibility measure

The short-time objective intelligibility measure (STOI, Taal et al., 2010) is based on a short-time analysis and subsequent cross-correlation of the temporal envelopes of clean and degraded speech signals. STOI showed a high correlation with speech intelligibility for listening experiments by Kjems et al. (2009) and performed well when compared to other objective intelligibility measures (see Taal et al., 2010, for details). The main focus of the model lies on simplicity and straight forward calculations, therefore STOI does not contain a physiologically imposed band-importance function, as the SII and ESII. The main aspect is a decomposition of the input signals into discrete time-frequency bins. The duration of the temporal frames considered in the cross-correlation is 386 ms. The frequency analysis was performed with 15 one-third octave bands, covering a range of 50 Hz − 4.3 kHz. All signals were down-sampled to 10 kHz prior to the model analysis. The time-frequency decomposition was performed with 256-sample Hann windowed frames with 50% overlap. There was a monotonic relation between the model outcome (STOI units) and actual speech intelligibility scores for noisy and time-frequency weighted noisy speech. To compare the model outputs to experimental data, STOI units had to be mapped to the speech intelligibility data. This was done with a logistic function (see equation (8) in Taal et al., 2010) and fitting of the free parameters a and b with a non-linear least-square method. The parameter sets used for each combination of target and masker spectrum in the current study are shown in Tab. 4.2. Predictions were averaged over twenty target sentences, while each sentence was presented in a different masker sequence.
In their study, Taal et al. (2010) claimed that STOI worked well in case of additive noise. There was a "unprocessed" (UN) condition that resembled speech degraded by noise and the model predicted those listening situation quite well. For the UN, there was no preprocessing of the noisy speech (time-frequency weighting or single-channel noise reduction). Taal et al. (2010) stated that it would be interesting to test STOI on other types of speech degradation, in addition to the ones described in their paper. The stimuli in the current study resembled such degraded speech conditions. They provided several types of additive masking noise (more UN conditions) to test the STOI model in. Model predictions were obtained without any preprocessing of the target speech or masker. Unlike the other models, STOI made few assumptions on energetic and amplitude modulation masking.

Gender combination	a	b
male/female	−26.994	12.550
female/female	−28.938	13.140
male/male	−35.916	17.496

Table 4.2: The parameters a and b as chosen for STOI to fit the model outcome to the SRTs from the listening experiments. The parameters were chosen such that the psychometric function (see Taal et al. (2010), eq. 8) for the SSN masker matched the observed SRT50 in the SSN masker. The parameters were changed for the different gender combination of target and masker as indicated (target/masker).

4.4 Results

4.4.1 Experimental speech reception and speech detection thresholds

Fig. 4.3 shows the observed average SRTs and SDTs with the corresponding standard deviations. The masker conditions are denoted on the abscissa, the observed SRTs and SDTs at the ordinate. The individual thresholds are connected with lines to guide the eye. The different gender combinations of target and masker material are displayed in the different panels of Fig. 4.3. The combination of male target and female masker spectrum is shown in the top panel, the combination of female target and masker spectrum in the middle panel, and the combination of male target and masker spectrum in the bottom panel. Open squares represent the SRT50s while closed squares indicate the SRT80s. SDTs are depicted with triangles.

Male target and female masker spectrum

The upper panel of Fig. 4.3 shows a characteristic pattern of SRTs and SDTs across the eight maskers: The left-hand side of the panel shows data obtained with the SSN-based maskers, the right-hand side with speech-like maskers. The SSN yielded the highest SRT50 of -7.5 dB, followed by the SRT50 of the AFS-SSN condition (-9.2 dB). The SRTs for the modulated maskers with spectral coherence of the applied modulations were lower as those of the AFS-SSN and amounted to -14.1 dB and -17.1 dB for BB- and SAM-SSN, respectively. The largest masking release of 9.6 dB occurred between the SSN and SAM-SSN, which had regular and coherent modulations. All SRT50s for the speech-like maskers were in a similar range, whereas the NV-ISTS yielded the highest SRT50 (-17.5 dB) and the ST the lowest (-22.6 dB). Apparently, speech intelligibility was very similar, no matter if the masker consisted of a single or more talkers or if the original or noise-vocoded version of the speech-like maskers were presented. Comparing the SRT80s with the SRT50s across the upper panel in Fig. 4.3, a constant offset of about 4 dB (6 dB for speech-like maskers) between the two measures was observed. Otherwise, the pattern of the SRT80s was very similar to the SRT50s for the eight maskers: The

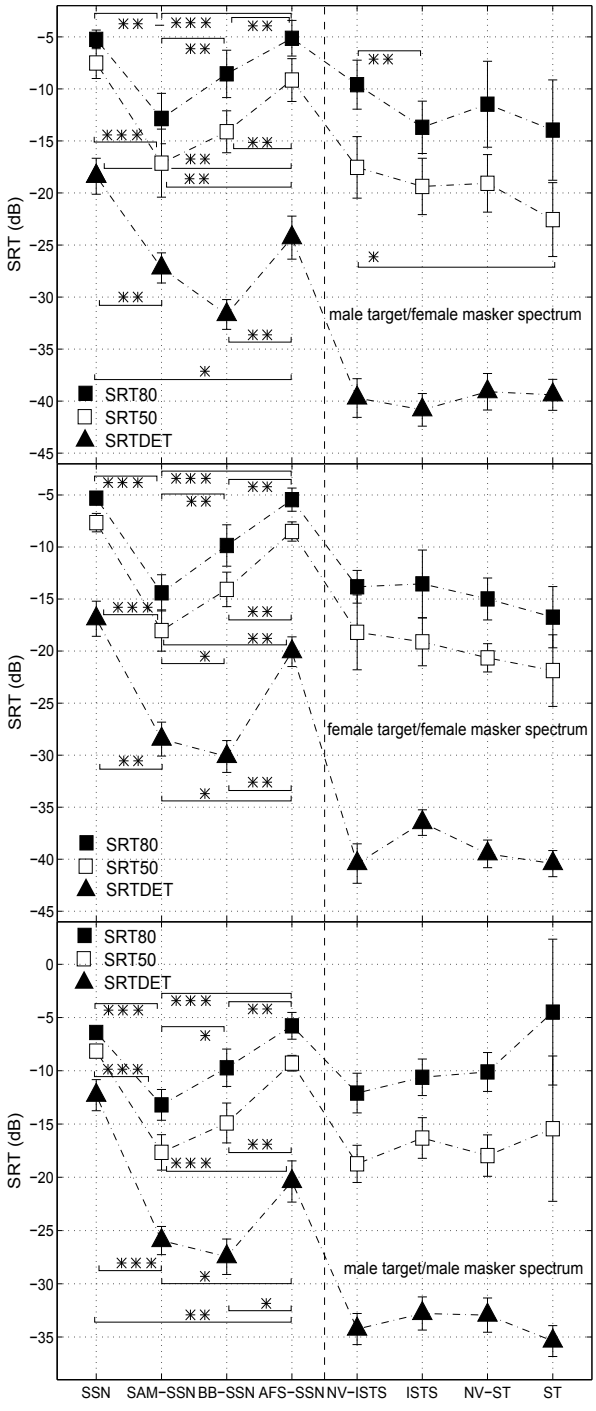


Figure 4.3: Mean speech reception (SRT) and detection thresholds (SDT) as measured for the eight masking conditions along with the standard deviations for SRT50 (open squares), SRT80 (closed squares), and SDT (triangles). The three different gender combinations of target and masker are shown in the different panels. The mean values are obtained from individual data of the same eight listeners. For the gender combination male target and masker, the male version of the ISTS, which was generated in the current study, is used. Significant threshold differences within the SSN-based and speech-like maskers are denoted with stars, where the number of stars indicates the confidence level (* is $p < 0.05$, ** is $p < 0.01$, and *** is $p < 0.001$). Significant threshold differences across SSN-based and speech-like maskers are shown in Table 4.3.

highest SRT80 was observed for the SSN and AFS-SSN (-5.2 dB and -5.1 dB) and there was a general masking release in SRT80s for the modulated maskers. The largest release occurred again between the SSN and SAM-SSN (7.6 dB). All SDTs were well below the SRTs: for the SSN and SAM-SSN maskers, the SDTs were about 10 dB lower, for the other BB- and AFS-SSN maskers they were about 15 dB lower, and for the speech-like maskers even about 20 dB lower than the SRT50s. Nevertheless, the overall pattern of SDTs was comparable to the SRTs. As for the SRTs, the highest thresholds were observed for the SSN and AFS-SSN. There was also a release from masking for the modulated SSN-based maskers in the SDT experiment, but unlike for the SRTs, the masking release did not increase with regularity, i.e., the SAM-SSN did not yield a lower SDT than the BB-SSN. The largest masking release for the SDTs occurred between SSN and BB-SSN and amounted to 13.3 dB. This was slightly larger than for the intelligibility measurements. Considering the SDTs for the four speech-like maskers, there was hardly any difference observed. As for the SRTs, the type of interfering talker (ISTS, ST) and the presence or absence of fundamental frequency information due to noise-vocoding did not significantly influence the SDTs.
To assess the effect of the maskers on SRT50, SRT80, and SDT, a one-way repeated-measures analysis of variances (ANOVA) was performed for each of the three measures. A highly significant main effect of the masking condition was found: [$F(7, 49) = 127.16$, $p < 0.001$] for SRT50s, [$F(2.73, 19.11) = 34.12$, $p < 0.001$] for SRT80s, and [$F(7, 49) = 82.02$, $p < 0.001$)] for SDTs. Post-hoc pairwise comparisons with Bonferroni correction for the SSN-based maskers showed that the thresholds for SSN and AFS-SSN differed significantly for SRT50s and SDTs. Significant differences also appeared between AFS- and BB-SSN (for all three measures) and AFS- and SAM-SSN (SRT50, SRT80). Differences due to the regularity of the modulations (irregular modulations of BB-SSN versus regular modulations of SAM-SSN) were only significant for SRT80s. The pairwise comparisons for the four speech-like maskers did not yield significant differences in thresholds for most of the measures. The only significant differences appeared for NV-ISTS versus ST (SRT50) and NV-ISTS versus ISTS (SRT80). Significant differences and the significance level for the thresholds within the SSN-based and speech-like maskers are indicated in Fig. 4.3 (differences within the SSN-based maskers at the left-hand side; differences within the speech-like maskers at the right-hand side). Significant differences are shown with stars, where $p < 0.05$ is marked with *, $p < 0.01$ with **, and $p < 0.001$ with ***. Significant differences in thresholds across the SSN-based and speech-like maskers are shown in Table 4.3.

Female target and female masker spectrum

The middle panel of Fig. 4.3 shows the intelligibility and detection thresholds for the combination of female target and masker in the same style as in the upper panel. Again, SRT50s for SSN and AFS-SSN were similar, the values were -7.7 dB and -8.5 dB. In general, SRTs decreased as modulations were introduced to the SSN masker, the largest masking release was 10.4 dB between SSN and SAM-SSN. All SRT50s for the speech-like maskers were

about -20 dB. Overall, the course of the SRT80s was the same as for the SRT50s, the offset between the two measures was again 4 dB for the SSN-based maskers and about 6 dB for the speech-like maskers. The highest SRT80s were -5.3 dB and -5.5 dB (SSN and AFS-SSN) and thus almost identical to the SRT80s in the upper panel. The largest release from masking occurred for the SSN and SAM-SSN and was 9.1 dB for the case of SRT80s. All speech-like masker SRT80s were in a similar range of -15 dB, which was also similar to the speech-like masker SRT80s in the upper panel.
The course of the SDTs was again similar to the SRTs. As for speech reception, the highest SDTs were obtained with the SSN and AFS-SSN masker (-16.9 dB and -20.1 dB), but the lowest SDT was found for BB-SSN. The masking release between the SSN and BB-SSN condition was 13.2 dB and thus almost identical to the value in the upper panel. SDTs for the speech-like maskers were again about 20 dB lower than the SRTs. A one-way repeated-measures ANOVA showed a highly significant main effect of masker type for SRT50s [$F(7, 49) = 65.88$, $p < 0.001$], for SRT80s [$F(7,49) = 53.54$, $p < 0.001$], and for SDTs [$F(7,49) = 66.93$, $p < 0.001$]. Post-hoc pairwise comparisons with Bonferroni correction showed similar results as for the upper panel. The SSN and AFS-SSN masker thresholds did not differ significantly for any of the three measures, but coherent across-frequency modulations led to significant differences for the individual masker types. AFS- and BB-SSN thresholds differed significantly from another for all three measures (SRT50s, SRT80s, SDTs), as did AFS- and SAM-SSN thresholds. The introduction of regular modulations (SAM-SSN) compared to irregular modulations (BB-SSN) was significant only for SRT50s and SRT80s. The pairwise comparisons between the four speech-like maskers showed that there were no significant differences in thresholds for any of the speech-like maskers at any of the three measures. All significant differences in thresholds are indicated by the stars in the middle panel of Fig. 4.3 and in Tab. 4.3.

Male target and male masker spectrum

The lower panel of Fig. 4.3 shows the SRTs and SDTs for the combination of male target and male masker spectrum. The course of the SRTs was similar to the other two panels. The highest SRT50s were obtained with the SSN and AFS-SSN (-8.2 dB and -9.3 dB). A masking release occurred for the introduction of coherent modulations (BB- and SAM-SSN) across the frequency spectrum and was largest between SSN and SAM-SSN (9.5 dB). This was almost exactly the same value as for the other panels in Fig. 4.3. Considering the speech-like maskers, SRT50s for NV-ISTS, ISTS, and NV-ST were similar. An exception, compared to the other panels, was the ST masker. For the combination of male target and male masker spectrum, this masker yielded SRT50s that were about 5 dB higher and had a larger standard deviation than all other speech-like masker thresholds in this and the other panels. Some subjects had severe problems with the reception of speech in the single talker masker, as will be discussed later. As for the other panels, there

setup	measure	SSN-based masker	speech-like maskers			
			NV-ISTS	ISTS	NV-ST	ST
male target female masker	SRT80	SSN	*	**	*	**
		SAM-SSN	*			
		BB-SSN		**		*
		AFS-SSN	**	**	*	**
	SRT50	SSN	**	***	***	***
		SAM-SSN				**
		BB-SSN	*	***	***	***
		AFS-SSN	**	***	***	***
	SDT	SSN	***	***	***	***
		SAM-SSN	**	***	**	**
		BB-SSN	**	**	*	*
		AFS-SSN	**	***	**	**
female target female masker	SRT80	SSN	***	**	***	**
		SAM-SSN				
		BB-SSN	*		*	
		AFS-SSN	***	**	***	**
	SRT50	SSN	***	***	***	**
		SAM-SSN				
		BB-SSN		**	**	**
		AFS-SSN	**	**	***	**
	SDT	SSN	**	**	**	***
		SAM-SSN	*	*	*	**
		BB-SSN	*		**	**
		AFS-SSN	**	**	**	**
male target male masker	SRT80	SSN	**	**		
		SAM-SSN				
		BB-SSN				
		AFS-SSN	**	**	*	
	SRT50	SSN	***	***	***	
		SAM-SSN				
		BB-SSN	*		*	
		AFS-SSN	***	**	**	
	SDT	SSN	***	**	***	***
		SAM-SSN	**		**	**
		BB-SSN	**		**	**
		AFS-SSN	**	**	**	**

Table 4.3: Statistically significant differences in speech reception and detection thresholds across the SSN-based and speech-like maskers. The upper part of the table displays the significances for the gender combination of male target and female masker. The middle and lower parts display the significant differences for the measurements with the same gender spectra. Significances at a level of $p < 0.05$ are shown as *, $p < 0.01$ as **, and $p < 0.001$ as ***.

was an offset of 4 dB between the SRT50s and SRT80s for the SSN-based maskers and 6 dB for the speech-like maskers (10 dB for the ST masker). The highest SRT80s were again found for SSN and AFS-SSN (-6.4 dB and -5.8 dB), although compared with the other two panels, the SRT80 for the SSN was about 1 dB lower in the case of male target and masker. The masking release between SSN and SAM-SSN was 6.8 dB, which was slightly smaller than in the other two panels of Fig. 4.3.
The SDTs were similar to the data presented in the other panels as well. The largest SDTs were observed for SSN (-12.3 dB) and AFS-SSN (-20.4 dB), the largest masking release occurred between SSN and BB-SSN and amounted to 15.2 dB. SDTs for all speech-like maskers were around -34 dB, with the ST masker showing the lowest SDT. This was in contrast to the SRTs for this panel, where the ST masker showed the highest SRTs. The SDTs for the speech-like maskers in the lower panel were in general about 5 dB higher than the SDTs for the other gender combinations of target and masker. One-way repeated-measures ANOVAs for each measure showed a significant effect of the maskers for SRT50 [$F(1.43, 10.01) = 18.14$, $p < 0.001$], for SRT80 [$F(1.31, 9.17) = 9.73$, $p = 0.009$], and for SDT [$F(7, 49) = 84.62$, $p < 0.001$]. Subsequent pairwise comparisons with Bonferroni correction showed that the differences in thresholds between SSN and AFS-SSN were significant only for the SDTs. As for the female spectra combination, SRTs between AFS- and BB-SSN and AFS- and SAM-SSN (as well as SDTs for these maskers) differed significantly from another. Differences between BB- and SAM-SSN were only significant for the SRT80s. Pairwise comparisons for the speech-like masker thresholds indicated no significant differences for any of the three measures and this was consistent with data from the other panels of Fig. 4.3. Due to the large standard deviations, the SRTs for the ST masker in the lower panel of Fig. 4.3 were not significantly different from all other SSN-based masker thresholds (for both SRT50s and SRT80s). This was in contrast to the results in the other panels of Fig. 4.3.

4.5 Model predictions

Fig. 4.4 shows the predicted SRT50s of the five speech intelligibility models for all three gender combinations of target and masker along, with the experimental data in the same style as in Fig. 4.3. Predictions for the SRT80s were omitted, because the predictions showed the same pattern as for the SRT50s. Experimental data are shown with open symbols, model data with filled symbols. The root-mean square errors (RMSEs) that occurred between the empirical data and the model predictions are shown in the legend. All model outcomes in Fig. 4.4 were adjusted as to match the SRT50 of the SSN masker in each gender combination. SII predictions are depicted with squares, ESII predictions with diamonds, ESIIsen predictions with downward triangles, mr-sEPSM predictions with circles, and STOI predictions with upward triangles. The individual masker types are denoted at the abscissa. Connecting lines between the SRTs are again used to guide the eye.
The upper panel in 4.4 shows the empirical data and model predictions for

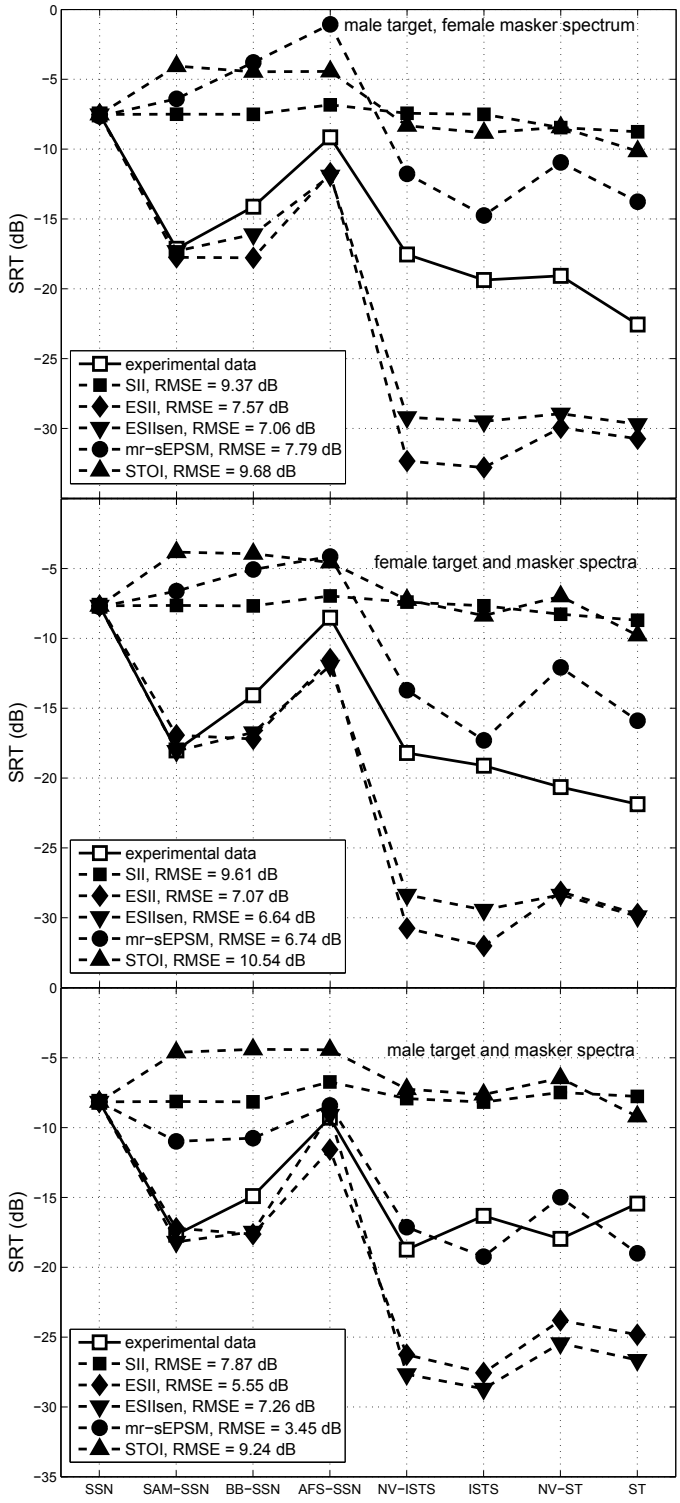


Figure 4.4: Predicted SRT50s of the five speech prediction models. The experimental data (SRT50s) is shown with open symbols, model predictions with closed symbols. The legend shows the root mean square errors (RMSEs) for each model. All model outcomes are adjusted as to match the SRT50 in the SSN masker in each particular combination of target and masker spectrum (displayed in each panel). For the combination of male target and masker spectrum, the male version of the ISTS is used. Model predictions are shown without errors, as these are discussed in section 4.5.

the combination of male target and female masker spectrum. Predictions by the SII were more or less independent of the maskers, ranging from -7.5 dB (SAM-SSN) to -8.4 dB (ST), as was expected, given that all SSN-based maskers and the ISTS shared the same long-term power spectrum. This similarity of predicted thresholds demonstrated that the applied signal manipulations did indeed not change the long-term spectrum of the maskers. As the ST and NV-ST had a similar long-term spectrum as the other six maskers, their predicted SRT50s were also similar to those of the other six maskers. Altogether, speech recognition was underestimated for all maskers (except the SSN).
In contrast, predictions of the ESII and ESIIsen varied strongly across the eight maskers and showed a good agreement with the empirical data for the SSN-based maskers. The prediction for the SAM-SSN masker (-17.7 dB) was almost identical to the experimental SRT50. The release from masking was overestimated by 3−4 dB for BB- and AFS-SSN. For the speech-like maskers, there was a considerable mismatch between empirical data and predictions, the predictions underestimated the experimental SRT50s by about 10 dB. In summary, ESII and ESIIsen predictions matched the listener's data better than the SII predictions, as can be seen by the respective RMSEs in Fig. 4.4. The ESIIsen predicted SRTs slightly better than the ESII in the case of different genders for target and masker and for female target and masker. ESIIsen showed the lowest RMSE for male target and female masker due to a better agreement with the experimental data for the speech-like maskers, especially the ISTS. Here, SRT50s were underestimated too, but predictions were generally about 3 dB better than those of the ESII.
Predictions of the mr-sEPSM yielded large RMSEs, which was mainly caused by the lack of predictive power in the modulated SSN-based maskers. The mr-sEPSM did show a decrease in thresholds that grew with coherence of the modulations across the frequency spectrum (AFS-SSN versus BB-SSN) and the regularity (BB-SSN versus SAM-SSN), as was observed in the experimental data, but the model did not show a release from masking compared to the SSN. Thus, SRTs were overestimated by more than 10 dB for the SSN-based maskers. This trend continued for the speech-like maskers, where the masking was overestimated by more than 5 dB. However, it should be noted that if the model predictions were shifted down by about 8 dB, the experimental data would be explained much better, except for the SSN (see the discussion in section 4.6.4).
The largest RMSE in the upper panel of Fig. 4.4 was found for the predictions of the STOI model. STOI neither showed a masking release for the modulated maskers, nor for the speech-like maskers. The coherence of the applied modulations did not influence the outcomes of the model at all, neither did the regularity of the modulations within the masker. Thus, predictions generally overestimated the experimental SRT50s by about 10 dB.

The other two panels of Fig. 4.4 show the model predictions together with the experimental data for the combinations where target and masker spectra are both female (middle panel), or male (lower panel). While the

female target talker spectrum was different from all female maskers, the male maskers were spectrally better matched to the male target talker. All male maskers were derived from the male version of the ISTS, which had the same mean fundamental frequency as the male target material. Thus, prediction differences between the middle and lower panel could, in principle, be caused by a better spectral match of target and masker for the lower panel. For both panels, the overall pattern of predictions was the same as for the upper panel. SII predictions hardly differed for the individual maskers and ESII and ESIIsen predictions for SAM-SSN were almost identical to the empirical SRT50 for the SAM-SSN. The release from masking for the other modulated SSN-based maskers was again slightly overestimated, while the release for the speech-like maskers was largely overestimated. The RMSEs for the ESII were smaller for the lower panel, than for the other two panels. RMSEs for the ESIIsen model were similar across all three panels. Predictions made by the mr-sEPSM showed the smallest RMSE for the combination of male target and male masker. Here, the release from masking was generally underestimated for the modulated SSN-based maskers, but the experimental AFS-SSN SRT50 was met very well. The prediction for the SRT50 of the NV-ISTS deviated by only 1.7 dB from the experimental value and all other SRT50 predictions for the speech-like maskers were similar to the experimental data. However, listener's thresholds were lower for the original, instead of and noise-vocoded speech-like maskers, but the mr-sEPSM predictions showed the opposite. The largest RMSEs for the middle and lower panel were obtained with the STOI model. The release from masking for the modulated SSN-based maskers was underestimated by about 10 dB and SRT50s of the speech-like maskers generally overestimated. This was true for both combinations with the same gender for target and masker.
All model data in Fig. 4.4 are shown without standard deviations. For the SII and ESII model, the errors were about ±4 dB for the SSN-based and ±10 dB for the speech-like masker. This was the case for all three gender combinations of target and masker. When real sentences were used as input (ESIIsen) the errors for the SSN-based maskers increased slightly to ± 6 dB, but remained the same as for the ESII for the speech-like maskers. Errors for the mr-sEPSM model were in the range of 0.2−23%, where the lowest errors occurred for very low SNRs and the largest errors in the range between SNRs of -5 and -15 dB. For the combinations with the same gender spectra, the maximal errors were slightly larger (up to 28%), whereas errors in general were smaller for the speech-like maskers (up to 20%). The assumed errors for the STOI predictions were ±0.03 and ±0.06 in STOI units for SSN-based and speech-like maskers, which corresponded to ± (1−2) dB and about ± (2−3) dB, respectively.

4.6 Discussion

The maskers used in the current study differ in their spectro-temporal features regarding regularity and across-frequency coherence. In the following, these aspects, the relation of SRTs and SDTs, as well as the relation of predictions

by the applied SI models and observed SRT50s are discussed.

4.6.1 Role of long-term spectrum and absolute threshold

The SRT50s in the current study are around -7.5 dB, which is in line with the threshold of -7.1 dB as reported in Wagener et al. (1999) for the OLSA in a spectrally matched masking noise. There is a maximum release of masking between the SSN and SAM-SSN maskers and slightly smaller release between SSN and BB-SSN, although both modulated maskers have coherent across-frequency modulations. The speech-like maskers generally show considerably lower thresholds without systematic variations among them. This pattern of results is comparable across all three gender combinations of target and masker, although they are not always spectrally matched (only in the case of male target and male SSN-based maskers and the male ISTS version). This is in contrast to studies such as Brungart (2001) that show best SI performance when target and masker are of different gender. A possible explanation for the finding in the current study is the difference between male and female voices in the frequency range below 100 Hz. Male voices have more energy in this range, but such low frequencies do not contribute much to speech reception, i.e. male and female voices are similar in terms of low frequency speech processing in the auditory system and can thus produce similar SRTs. To assess a possible effect of the absolute hearing threshold for the target speech in conditions with the speech-like maskers, SDTs were also measured in quiet with three of the eight listeners. The same setup as for the masked SDTs was used for this. In this case, one of the two intervals contained an entire target sentence and the other silence. These measurements were performed for the male and female target material and showed SDTs that were -58.8 dB ($\pm$ 1.55 dB) for the male and -59.3 dB ($\pm$ 1.28 dB) for the female target sentences. SDTs in silence are therefore well below the masked SRTs in Fig. 4.3 and rule out any possible flooring effect due to the audibility of the target sentences.

4.6.2 Role of spectro-temporal masker structure for SRTs

The SSN and AFS-SSN maskers show the highest SRTs and SDTs and these thresholds do not differ significantly from another for most gender combinations of target and masker. Although the AFS-SSN introduces speech-like amplitude modulations with considerable gaps in the frequency bands, it has the most incoherent modulations across the spectrum of all modulated maskers (see lower right panel of Fig. 4.1). Thus, the AFS-SSN thresholds demonstrate that coherent AMM is required for an effective masking release as shown, for example, in Dau et al. (2013). SRTs for SAM-SSN and BB-SSN are lower than for AFS-SSN, showing a statistically significant release from masking for the coherently modulated maskers. SAM-SSN has a fully regular and coherent amplitude modulation pattern and shows the lowest thresholds, which is in line with Stone et al. (2012), stating that regular modulations result in a greater release from masking than irregular modulations. AFS-SSN and BB-SSN maskers share the same characteristics of their modulation pat-

terns when individual frequency bands are considered, but BB-SSN has the same coherent modulation pattern across all frequencies, whereas this is not the case for the AFS-SSN. Thus, the AFS-SSN and BB-SSN are conceptually similar to the classical random and co-modulated maskers in psychoacoustic co-modulation masking release (CMR) paradigms as described in Hall et al. (1984). In psychophysical studies, detection thresholds for a pure-tone target are lower in the presence of a co-modulated (here BB-SSN) than a random masker (here AFS-SSN). Because of the pure-tone target, the CMR can then be partly attributed to within-channel modulation cues and partly, but to a lesser extend, to across-channel mechanism. These across-channel mechanism are estimated to cause 2−4 dB of masking release (Piechowiak et al., 2007; Dau et al., 2013). This is in the same range as the differences in SRTs in the current study, although the underlying mechanisms for SRTs and pure-tone detection are likely not the same. However, as similar numbers are observed in the SDTs, this might indicate that across-channel cues are important for the speech detection experiment.

The generally lower SRTs for BB- and SAM-SSN, compared to the SSN, can be explained by the concept of dip listening (e.g., Bronkhorst, 2000). According to this assumption, the SRTs for the BB-SSN should be lower than those of the SSN, because temporal gaps in the masker are long enough so that large parts of the target speech are unmasked. For the SAM-SSN, the gaps are shorter and occur at a higher (but regular) rate, so less portions of the target speech are accessible following the dip listening concept. Consequently, SRTs for the SAM-SSN should be higher than those of the BB-SSN. However, SRTs are lowest for SAM-SSN, showing that regularity of the modulations is an important factor in speech recognition. The simulations of the ESII model are interesting, as they (representing dip listening) do not show differences in SRTs or even lower SRTs for BB-SSN than for SAM-SSN. This suggests that another masking aspect beside short-term EM, as described by the ESII, occurs for the SAM- and BB-SSN maskers. A prominent candidate for this is the aspect of amplitude modulation masking.

Concerning AMM, it is interesting to measure SRTs with a 4-Hz modulated masker in a future study, to investigate this more closely. For the SAM-SSN, the rate of modulation is 8 Hz, which is higher than the typical speech-modulation rate of $4 - 5$ Hz (Dubbelboer and Houtgast, 2008). As modulation masking is assumed to be modulation frequency selective (e.g., Houtgast, 1989; Ewert and Dau, 2000), the BB-SSN, which shows the typical speech-like modulation rates, should cause more masking than the (8-Hz modulated) SAM-SSN masker. This is also observed in the data. Based on modulation masking, a 4-Hz SAM-SSN masker should then cause higher SRTs than the (8-Hz modulated) SAM-SSN, as the masker modulation rate is closer to the target modulation rate (Houtgast, 1989; Ewert and Dau, 2000). In contrast, following the concept of dip listening, SRTs for a 4-Hz modulated SAM-SSN should then decrease, as larger temporal gaps are provided. Testing this certain condition in an future study would provide a deeper insight on the interplay of the two masking aspects EM and AMM.

It is apparent that SRTs for the speech-like maskers are considerably lower

than for the SSN-based maskers for all three gender combinations of target and masker. Again, dip listening (Bronkhorst, 2000) is a possible explanation for this, as there are larger temporal gaps in which the target sentence can be perceived in quiet for the speech-like maskers than for SAM-SSN or AFS-SSN. This reasoning is supported by the ESII predictions that take dip listening into account and show a great decrease between SRTs for SSN-based and speech-like maskers. Thus, SRTs of the speech-like maskers can in general be lower than those of the modulated SSN-based maskers, although the modulation rates are similar. The decrease in SRTs in the speech-like maskers seems to be robust across the three gender combinations of target and masker, but SRTs are highest for a male interfering talker on male target speech (right side of lower panel in Fig. 4.3). Some listeners showed SRTs above 0 dB in this certain condition, causing the large standard errors in Fig. 4.3. Interestingly, this is not observed for the female interfering talker and female target (although the gender of target and maskers were also the same in this certain combination).
Generally, SRT80s and SRT50s show a parallel course for all gender combinations of target and masker. The offset between SRT80 and SRT50 is always about 4 dB for SSN-based and larger (about 6 dB) for speech-like maskers. The larger difference for the speech-like maskers might be related to dip listening, as there are longer temporal gaps in the speech-like maskers that can be utilized.

4.6.3 Relation of SRTs and SDTs

Comparing SDTs with SRTs, it is obvious that the SDTs are in general well below the SRTs, while sharing a similar threshold pattern. SDTs are highest for the SSN and AFS-SSN maskers, lower for the SAM-SSN and BB-SSN maskers, and lowest for the speech-like maskers. It appears plausible that SDTs are in general lower, given that the task is pure signal detection and speech reception is of no concern. The interesting point is the difference of SRTs-SDTs or the reception-detection (RD) gap, which quantifies the SNR required for 50% (or 80%) speech reception as a function of the masker type. It is obvious that the RD gap (calculated from the SRT50s) depends on the masker type. For the upper panel of Fig. 4.3, the RD gap amounts to $10-11$ dB for SSN and SAM-SSN, 17.6 dB for BB-SSN, 15.1 dB for AFS-SSN, and $20-22$ dB for the speech-like maskers. These numbers are about 1 dB smaller for female target and masker, but vary more ($2-6$ dB) for male target and masker. According to Arbogast et al. (2005), speech detection is ruled mainly by the aspect of EM. Following this hypothesis, the RD gap can be used to estimate the effect of AMM and IM in addition to EM. Assuming that the RD gap for SSN and SAM-SSN represents the difference between energetic masking of the entire target speech signal and its correct reception, the larger RD gap for BB- and AFS-SSN can be interpreted as to reflect the additional effect of AMM. This accordingly amounts to $5-7$ dB for these two modulated SSN-based maskers. The even larger RD gap for the speech-like maskers can then be interpreted to contain the $5-7$ dB effect of AMM and a further offset of about 15 dB, which can in principle be attributed to either effects

of coherent across-frequency amplitude modulations or IM. Since Brungart et al. (2001) and Micheyl et al. (2000) state that much IM is conveyed by interfering talkers, it can be assumed that the large RD gap in the case of speech-like maskers is to a large part caused by IM.
It is, however, arguable if speech detection is mostly ruled by energetic masking or if amplitude modulations by the masker (Stone et al., 2012; Dubbelboer and Houtgast, 2008) might also influence speech detection. In that case, the RD gap would still allow an estimation of the effects of modulation-frequency-selective AMM (caused by the spectral coherence of the modulated maskers) and IM.

4.6.4 Model predictions

Predicted SRTs were compared to the empirical SRT50s to gain deeper insight on the role of the different types of masking accounted for by the individual SI models. However, informational masking is not addressed in any of the models; thus, differences between empirical data and predictions can also hint to effects of informational masking.

SII and ESII (ESIIsen)

The SII works rather rudimentary in terms of a stimulus-specific analysis. Its predictions rely only on the long-term energy spectrum of the individual maskers, thus only long-term energetic masking is explained. For the current maskers, SII predicts a more or less constant SRT, which supports the initial design goal of identical long-term masker spectra. As expected, the SII simulations show clearly that long-term energetic masking is a poor predictor for speech reception if the masker is not stationary.
The ESII incorporates the concept of EM by a short-term analysis of the input stimulus and is therefore expected to yield more accurate predictions for fluctuating maskers. Thus, the ESII accounts for EM in short time frames and listening in the dips (Bronkhorst, 2000). It is interesting to directly compare the ESII predictions to SDTs, instead of the SRTs, as it can be hypothesized that SDTs can be well explained by short-time EM (Arbogast et al., 2005). SDTs and ESII simulations share generally the same pattern (see Figs. 4.3 and 4.4), except for an offset which can be attributed to the adjustment of the ESII predictions to match the SRT50 in the SSN masker. Both the ESII predictions and the SDTs share a difference of about 23 dB between the SSN and the speech-like masker thresholds. Even specific details in the threshold pattern are similar, i.e. the predicted thresholds for the BB-SSN are the same or slightly lower than those of the SAM-SSN, which is also seen for the SDTs in Fig. 4.3. Thus, the ESII seems to be a good model for the SDTs. But the pattern of the ESII predictions is in contrast to the pattern of the empirical SRTs in Fig. 4.3, where the lowest threshold is found for the SAM-SSN. This can be another indication that there is a masking effect in addition to EM, namely AMM. Then again, given that the ESII perfectly accounts for the SRTs of the SAM-SSN masker, it could be hypothesized that speech reception in the SAM-SSN maskers is explained by

short-time EM alone. But this is somewhat in contrast to the idea of AMM, therefore investigating speech reception in a 4-Hz modulated SAM-SSN would help to distinguish between the effects of EM and AMM. Moreover, in all other conditions, where ESII underestimates the SRTs but would account for SDTs (if adjusted to match the SDT for the SSN masker), masking other than short-time EM (that is AMM or IM) must be responsible for the higher SRTs. These additional masking effects amount to $3-4$ dB for the SSN-based maskers and are likely related to the irregularity of the temporal fluctuation and reduced across-frequency coherence. In case of the speech-like maskers, the additional masking amounts to 10 dB. This is in line with Rhebergen et al. (2006), who mentioned that ESII could underestimate masking effects when speech-like maskers are used. Rhebergen et al. (2006) suggested that informational masking can act in addition to (short-time) energetic masking. In comparison to the estimates of $5-7$ dB for AMM (derived from the SRTs of BB- and AFS-SSN) and about 15 dB for IM in the speech-like maskers (derived from the RD gap), the ESII predictions suggest a slightly lower estimate for both values, namely $3-4$ dB for AMM and 10 dB for IM, respectively.
If real sentences are used in the predictions (ESIIsen), the predicted SRTs differ slightly from the ESII predictions, except for the SAM- and BB-SSN SRTs. In contrast to the ESII, the ESIIsen predicts slightly higher thresholds for BB-SSN than for SAM-SSN, showing that an ESII concept with exploiting the temporal statistics of the target sentences yields predictions that are closer to the measurements with the human listeners. Further differences occur for the speech-like maskers, where the predicted SRTs are about $2-5$ dB higher for the ESIIsen, compared to the ESII. Thus, the gap between experimental SRTs and model predictions (based on short-term EM by the ESII) is reduced when the real statistics of the short-time SNRs are better estimated by using real target sentences in the ESIIsen. In conclusion, a short-time EM approach that takes into account the full statistics of the short-time SNRs is a better model for human speech reception than a long-term EM analysis.

mr-sEPSM

The mr-sEPSM is expected to use cues from the analysis of the frequency and modulation filter domain and therefore take EM and AMM aspects into account. However, it does not predict the SRT50s very well when initially matched to the SRT50 of the SSN masker. SRT50s are then overestimated by $5-10$ dB for the individual masker types. Despite the general overestimation of the SRTs, a masking release due to the coherence of the applied modulations is visible and the size of the release between the AFS- and BB-SSN and the AFS- and SAM-SSN is captured correctly (larger release between AFS- and SAM-SSN), although there is no specific analysis of across-channel coherence (see Jørgensen and Dau, 2014) in this model. The overall poor predictive power of the mr-sEPSM is surprising, because this model is especially designed to predict speech intelligibility in fluctuating maskers and has performed well in Jørgensen et al. (2013). Since predictions from that original study could be reproduced quite well with the available model version, the reason for the

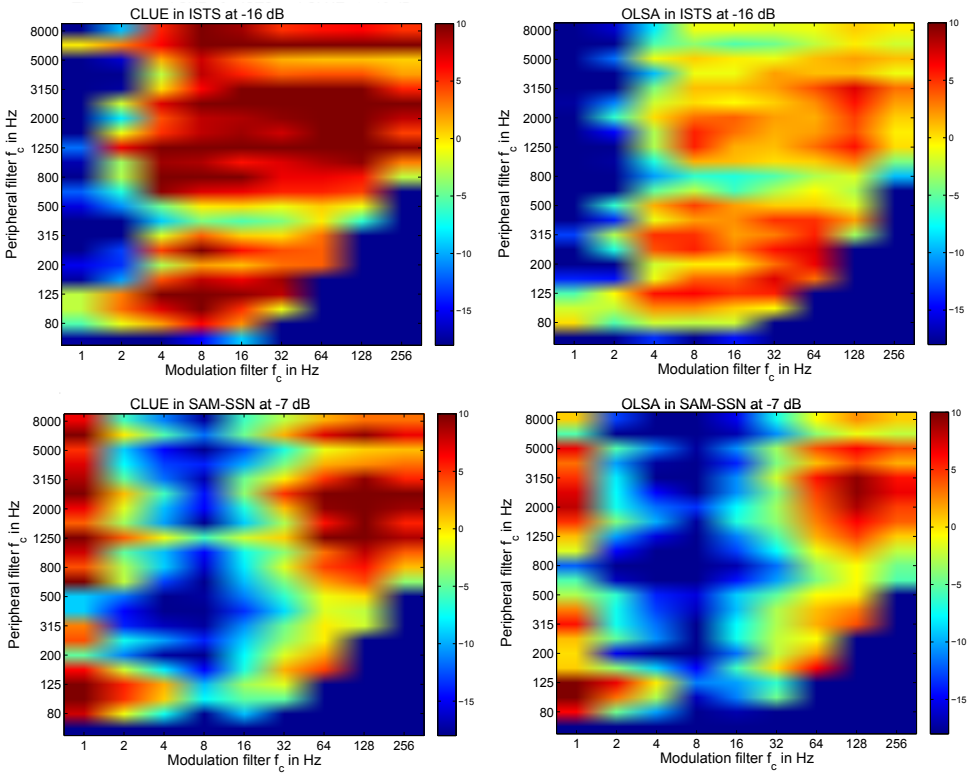


Figure 4.5: Time-averaged $\mathrm{SNR_{env}}$ outputs (in dB) of the mr-sEPSM across auditory and modulation filters for the CLUE and OLSA speech material in the ISTS and SAM-SSN maskers. Red parts correspond to large $\mathrm{SNR_{env}}$ values, blue parts to lower values in the auditory and modulation filters [a]. It is apparent that the CLUE material gains much of its $\mathrm{SNR_{env}}$ (and thus predictive power) from high auditory and high modulation filters, whereas the OLSA material has less energy in these bins. Consequently, the ideal observer stage in the model gains generally lower $\mathrm{SNR_{env}}$ for the OLSA speech material and could explain why predictions fail for the case of this certain speech material. For the OLSA speech material, the masking is generally overestimated, which is directly caused by the lower $\mathrm{SNR_{env}}$ for this target material.

[a] Disclaimer: If this figure looks funny, try the PDF-XChange Viewer. This is better than Adobe anyway and it's free!

	female masker				male masker			
	k	q	m	σ_s	k	q	m	σ_s
parameter set 1	0.351	0.5	50	0.6	0.655	0.5	50	0.8
parameter set 2	0.6	0.5	50	0.6	0.715	0.5	50	0.8

Table 4.4: This shows the different parameter sets for the mr-sEPSM when the SSN (parameter set 1) or AFS-SSN (parameter set 2) SRT50 is used for matching the model predictions to the data from the listening experiment. The values q and m are fixed throughout the predictions and only the other two parameters are adjusted.

prediction offset in the current study must be related to stimuli of the current study, e.g. the target material (OLSA versus CLUE). Although the same masking condition (SSN) is used for reference as in Jørgensen et al. (2013), the resulting parameters in the ideal observer stage are slightly different. The values q and m are fixed for all predictions, whereas the other two parameters are adjusted for each of the three gender combinations of target and masker (see Tab. 4.1).
Close investigation of the time-averaged $\mathrm{SNR_{env}}$ (compare to Fig. 5 in Jørgensen et al., 2013) shows that the OLSA speech material in the current study is very different from the CLUE material used in Jørgensen et al. (2013), regarding its energy content across the analysis frequencies. Fig. 4.5 shows the output of the $\mathrm{SNR_{env}}$ analysis for both speech materials in the SAM-SSN and ISTS masker and clearly, the CLUE material has more energy in the high auditory and modulation filters. If the model predictions, as stated in Jørgensen et al. (2013), are largely based on the SNR in those high auditory and modulation filters, the influence of the $\mathrm{SNR_{env}}$ power in those regions might be overrepresented for the CLUE material. When OLSA speech material is used (right panels of Fig. 4.5), the $\mathrm{SNR_{env}}$ power in the high auditory and modulation filters is smaller, explaining the general overestimation of masking for this certain target material.
Moreover, predictions of the mr-sEPSM are largely based on the conversion of the $\mathrm{SNR_{env}}$ values to percent correct values within the ideal observer stage. The predictions in Fig. 4.4 are gained by matching the model outputs to the SRT50 of the SSN masker. This is slightly different from Jørgensen et al. (2013), where the model outputs are adjusted to match a SSN that is spectrally matched to the CLUE material. Thus, there are slightly different parameter sets in that study and the current study. To assess the effect of different parameter sets, the choice of parameters was investigated more closely (see Tab. 4.4 and Fig. 4.6). This was done for two combinations of spectra (male target in female and male masker), since the mr-sEPSM model showed better predictions in the case of male target in male maskers in the first place. The values q and m are fixed for all predictions, as done in Jørgensen et al. (2013), whereas the other two parameters were adjusted for each of the two combinations of target and masker spectrum.
Parameter set 1 in Tab. 4.4 is the one that was used to match the experimental

SRT50 in the SSN masker (this is re-plotted from Fig. 4.4 in Fig. 4.6) and parameter set 2 was used to match the predicted AFS-SSN threshold to the SRT50 in the AFS-SSN. The corresponding predictions and RMSEs for the two parameter sets are shown in Fig. 4.6. The upper panel shows the predictions for the case of the female masker spectrum, the lower panel predictions for the male masker spectrum. The predictions with the smallest RMSE are gained with parameter set 2, when the SRT50 in the AFS-SSN is used for a reference. In this case, the SRT50 for the NV-ISTS is predicted exactly and SRT50s for the other speech-like maskers are closer to the data from the listening experiments. Modulated SSN-based masker SRT50s are still overestimated, but only by $3-4$ dB. For the parameter set 1 (SSN as reference) the overestimation is about $5-10$ dB for all masker types and thus generally larger.
For the male target and masker spectrum (lower panel of Fig. 4.6), the predicted SRTs for the two parameter sets are not much different from another. This is because the mr-sEPSM predictions fit the experimental data better in the first place (as seen for parameter set 1). For this choice of parameters, the SSN and AFS-SSN SRT50s are met very well and the RMSE is considerably lower than for the case of different gender of talker and masker gender (upper panel of Fig. 4.6) with the same parameter set. When using the AFS-SSN as a reference, the difference in SRT50s between the BB- and SAM-SSN is not as pronounced as if the SSN is used as reference. It is interesting to note that, although the AFS-SSN is chosen as reference for parameter set 2, the SRT50 for the SSN masker is still met very well. Remaining discrepancies occur mostly for predictions for the modulated SSN-based and speech-like maskers. For the speech-like maskers the overall predicted SRT50s fit better, but they show the exact opposite behavior to the empirical data, leading to the slightly higher RMSE for this parameter choice. In summary, the choice of parameters in the ideal observer stage is not crucial for predictions when target and masker have a similar spectrum, but has a great effect if they do not share a similar spectrum.

STOI

The STOI model performs well for most data from Kjems et al. (2009), but underestimates SI for the unprocessed (UN) condition in additive noise scenarios (car and bottle noise, see Taal et al., 2010) in that study. In the current study it fails to correctly predict SRT50s for all masker types. A possible reason could be that there is no time-frequency processing prior to the model analysis in the current study, as is the case for most conditions in Taal et al. (2010). But this explanation is unlikely, since the underestimation appears also for the UN condition in Taal et al. (2010), where preprocessing is omitted. A possible explanation that is mentioned in Taal et al. (2010), is the average noise spectrum, which is different from that of the clean speech tokens in that study. This is also the case for stimuli in the current study. The mismatch of spectra could explain the general overestimation of SRTs that is found for the predictions in the current study for the case of different gender of target and masker spectra. However, it cannot explain the deficits

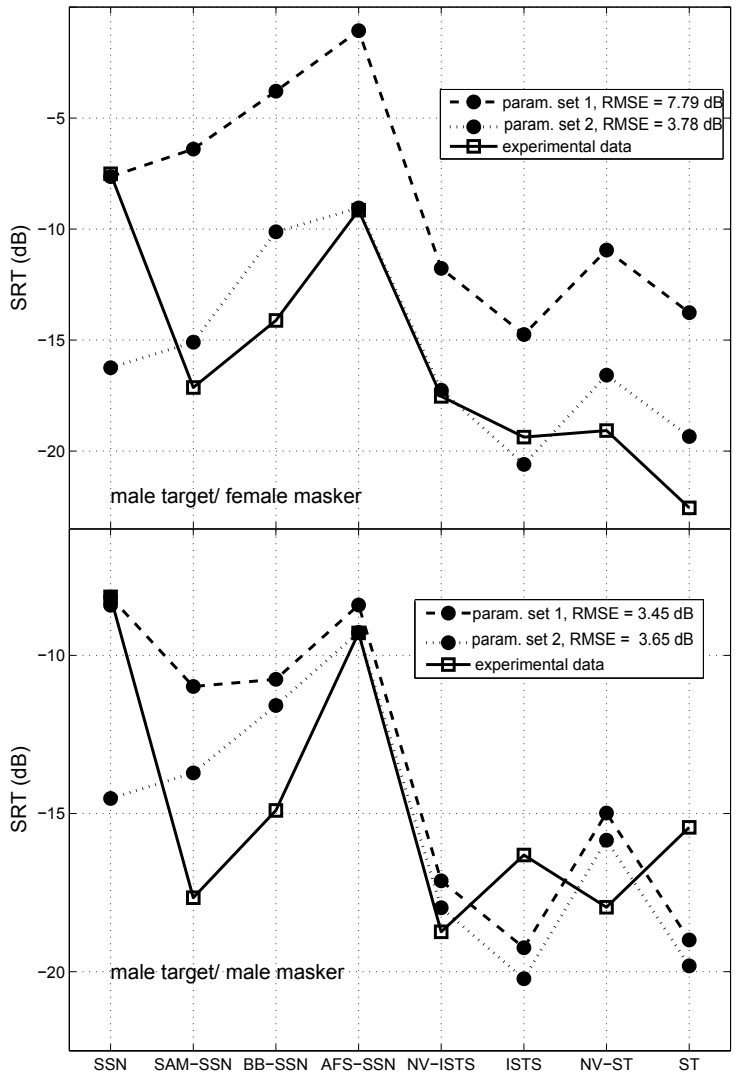


Figure 4.6: Predictions of the mr-sEPSM when using a different reference frame than the SRT50 of the SSN masker. The predictions are made for the combination of male target speech and female or male masker spectrum, respectively. The experimental data is depicted with open, the model predictions with closed symbols. Parameter set 1 is the one chosen to match the SRT50 in the SSN masker, thus data from Fig. 4.4 is re-plotted (dashed line). Parameter set 2 was chosen to match the SRT50 in the AFS-SSN masker for each combination (dotted line). For the upper panel (female masker) the RMSE decreases greatly when a reference frame other than the SRT50 of the SSN is used. For the lower panel (male masker) this is not the case. Here the three RMSEs do not differ much from another. This is presumably the case because parameter set 1 provides a small RMSE to begin with.

for the combinations where target and masker are of the same gender and therefore the, spectra similar or even matched (in the case of male target and SSN-based masker and male ISTS). Taal et al. (2010) state that in general, the model deficits could be overcome by introducing band-importance functions in the analysis. This could lead to better predictions, especially when target and masker are of the same gender.
Although STOI shows much less predictive power, compared to SII and ESII, it has a potential advantage, since it has additional parameters (a and b, see Tab. 4.2) that can be adjusted to match the prediction to the experimental data. Thus, STOI can, in principle, also be matched to references other than the SRT50 of the SSN masker. A further analysis of different parameter sets to fit the psychometric function is, however, omitted here for STOI, because STOI predictions do not represent the overall course of empirical SRT50s (e.g., a masking release) in the first place. To which extent the overall course of the predictions would be altered when the parameters are changed to match a reference condition other than the SSN, is subject to investigation in a further study. A first step towards this direction is shown in Fig. D.1 in the appendix. This figure shows STOI predictions that are gained when the input signal to the model consists of concatenated target sentences, instead of speech-shaped noise, or when a reference condition other than the SSN threshold is used. Fig. D.1 shows that indeed model predictions are hardly influenced by those changes. In general, STOI does not seem to account for effects such as masking release or listening in the dips in the current study.

4.6.5 Implications for the role of energetic, amplitude modulation, and informational masking

Altogether, results from the current study show that SRTs (SDTs) are lower for the speech-like maskers than for the SSN-based maskers. This is most probably caused by the influence of EM and AMM, less by IM, as is supported by the model predictions of the ESII, ESIIsen and mr-sEPSM. A possible hierarchy of the masking effects in the current data set would then look as follows: energetic masking explains most of the masking, followed by amplitude modulation masking, and informational masking provides the least amount of masking for speech intelligibility and detection in the noisy backgrounds of the current study.
However, even though informational masking does not appear to have a strong effect, empirical SRTs (SDTs) cannot be explained by the presence of EM and AMM alone. Model predictions by the ESII (ESIIsen) show an offset of $10 - 15$ dB to the SRTs in speech-like maskers and this can be attributed to IM, following Rhebergen et al. (2006). IM is thought to be most prominent for speech-on-speech masking (Brungart, 2001) and is influenced by features like fundamental frequency, number of interfering talkers, and the similarity between target and masker. In the current study, IM is addressed with the different speech-like maskers. This is somewhat different from studies such as Arbogast et al. (2002) and Brungart (2001), where IM is often realized with either interfering talkers that are identical to the target talker or with identical masking speech material. In these studies, the words of the target

and masker sentences are also temporally aligned so that word length and pauses are similar for target and masker. Nevertheless, the current study can draw conclusions on some influential factors of informational masking, since the deviations of the empirical data from the model predictions can be caused by informational masking effects that are not incorporated in any of the speech prediction models in the current study.

It can be hypothesized that the removal of the fundamental frequency (F0) increases the masking effect of the noise-vocoded speech-like maskers, because a separation of target and interferer due to F0 differences in target and masker is made more difficult. This is reflected in model predictions by the ESII (ESIIsen) and mr-sEPSM, where the SRT50 for the noise-vocoded maskers are slightly higher than for the original interfering talkers. But this is not seen in the empirical SRT50s from the listening experiments. Here, SRT50s for noise-vocoded and original maskers are either not statistically different or show the opposite behavior (lower panel in Fig. 4.3). This does not suggest an influence of fundamental frequency information on IM, which is reasonable when considering the stages in the auditory pathway at which IM is thought to arise and at which the fundamental frequency is analyzed. IM is thought to arise outside the auditory periphery, but F0 (differences) are analyzed at the cochlear stage of the pathway (Durlach et al., 2003a). Therefore, the fundamental frequency is most probably no dominant influential factor on informational masking.

Alternatively, the lack of significant differences between SRTs and SDTs for intact and noise-vocoded speech-like maskers in Fig. 4.3 could be caused by the chosen target sentence material, which leaves only little room for uncertainties and thus IM in general. Since the OLSA sentence material is very structured and hence predictable, listeners might know quite well what the target sentence will be like and can therefore concentrate better on the target material, ignoring the masker signal. This would actually lead to a de-masking effect and could explain the similar thresholds for all speech-like maskers. Then again, there is a large similarity between target and masker for the male ST masker, suggesting that IM occurs (Durlach et al., 2003a; Lutfi et al., 2013). A greater variance between the individual maskers could appear in more realistic settings, e.g., when the beginning of the sentence is unclear in timing, the target material itself more irregular (no matrix sentence tests) or when the maskers themselves are more realistic (real environment recordings).

In contrast to the described masking effects and used speech prediction models, there is another approach to describe masking effects based on salient time-frequency segments of the auditory signal. The concept of time-frequency segments has recently come up in the field of computational auditory scene analysis (CASA). There, so-called “glimpses” (Cooke, 2006) are used for a representation of the dominating source (in terms of SNR) in the mixture of signal and background noise. A glimpse could thus be defined as a spectro-temporal region where speech is least affected by the masker. Due to the redundant information of speech across the spectro-temporal plane, a sparse distribution of glimpses is often enough for speech perception

(Cooke, 2006). Brown and Wang (2005) proposed that SRTs can be derived from these glimpses and that the usage of glimpses can often sufficiently explain the perception of an auditory signal by the listeners. A glimpsing approach could be seen as a generalized analysis, combining elements of the "classic" EM and AMM by considering short-time SNRs in the time-frequency plane. Conceptually, even IM could be incorporated by means of processing efficiency of the provided (time-frequency) information. This would yield different intelligibility scores for comparable time-frequency distributions of target and masker, depending on the context of the masking situation.
The current data set provides a systematic approach to quantify masking effects in monaural speech processing and might provide a helpful benchmark for (joint) psychoacoustic, speech perception, and CASA model development. The maskers are publically available under http://www.uni-oldenburg.de/mediphysik-akustik/mediphysik/downloads.

4.7 Conclusions

Speech reception and speech detection were measured in various monaural masking conditions. Speech reception thresholds were also compared to predictions of five speech intelligibility models. The obtained results lead to the following conclusions:

1. Generally, there is a constant offset of 4 dB for the SSN-based maskers in this study when comparing SRT50s and SRT80s. For the speech-like maskers, this offset increases to 6 dB. This is robust for all gender combinations of target and masker.

2. A statistically significant co-modulation masking release appears in SRTs for all gender combinations of target and masker. There is a significant effect of the introduction of coherent, but irregular modulations across the frequency spectrum of the masker (AFS- versus BB-SSN condition). Regularity of the applied coherent modulations in the masker (SAM-SSN) further increases speech intelligibility and yields statistically lower SRTs for most measurements, compared to the SRTs from the masker with irregular modulations (BB-SSN).

3. Informational masking effects do not prominently arise in the current study for SI measurements, done with a matrix sentence test, such as the Oldenburger Satztest. There is no significant difference for thresholds obtained with one or more interfering talkers and there is no effect of presence or absence of fundamental frequency information in the speech-like maskers.

4. When a stationary masker (SSN) SRT is used for model calibration, SRT predictions show the best results for the ESII and ESIIsen. Empirical data for the case of male target material and male masker are best predicted by the mr-sEPSM. Prediction accuracy for the mr-sEPSM increases considerably when the predictions are matched to other reference masker conditions, such as AFS-SSN. Altogether, the ESII (and

ESIIsen) supports the assumption that the influence of EM can be seen in the speech detection data. The mr-sEPSM correctly predicts the influence of amplitude modulation masking, despite problems with the calibration.

5. Comparison of SRTs with SDTs and model data allows qualitative and quantitative statements regarding the three masking effects: Qualitatively, energetic masking seems to have the largest influence on SI and speech detection, followed by amplitude modulation masking and informational masking (in the setup of the current study). With respect to results of the ESII (ESIIsen) model and comparison of SI and detection data, the amount of amplitude modulation masking can be determined to be at least $3 - 4$ dB for the modulated SSN-based maskers. The masking for the speech-like maskers in the current study, can then be separated into contributions by AMM ($3 - 4$ dB) and IM ($10 - 15$ dB).

4.8 Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich "Das aktive Gehör" (DFG SFB/TRR31). We thank the Medical Physics group for fruitful discussions, Søren Jørgensen for providing his implementation for the mr-sEPSM model and discussion about it, and Thomas Biberger for helpful discussions on parameters of the modulation filterbank of the model.

Chapter 5

Summary, concluding remarks and possible future studies

In this dissertation, across-frequency processing in the identification of vowels in spoken CVCs (chapter 2) and the recognition of sentences from the Oldenburger Satztest in various masking backgrounds and spatial configurations was investigated (chapters 3 and 4). Observed speech reception thresholds (SRTs) were examined with respect to the influence of binaural cues (ILDs, IPDs; see chapter 3) and the three masking aspects energetic, amplitude modulation, and informational masking (see chapter 4). Moreover, observed SRTs were compared to predictions of various current speech prediction models. Data are provided for maskers that, due to their spectro-temporal complexity, challenge the existing models. In presenting a large amount of speech recognition data, this thesis provides empirical data to the discussion of the different masking aspects. It also sets a benchmark for further studies on speech recognition and the improvement of speech prediction models.

5.1 Findings on speech recognition from monaural and binaural measurements

The studies in chapters 3 and 4 measured speech recognition with the same masker material. In both studies, it is found that SRTs are similar for a stationary masker (SSN) and a masker that has incoherent modulations across the entire frequency spectrum (AFS-SSN). This suggests a transition from modulated to stationary maskers when the modulations become incoherent. Moreover, both studies show a decrease in SRTs when across-frequency modulations are present in the masker. In the binaural study, the coherence of these modulations is a factor that leads to significantly lower SRTs. In the monaural study, there are two factors that lead to significantly lower SRTs – the coherence and the temporal regularity of the across-frequency modulations.
Regarding the three masking aspects that are subject of interest in this

dissertation, the results from chapters 3 and 4 show that energetic masking (EM) is most influential on speech recognition. The other two aspects appear as less influential, but are not negligible. Comparing observed SRTs (and SDTs[1]) with ESII model predictions, it is found that short-term EM explains much of the observed speech recognition in modulated maskers. But there are still deviations between observed and predicted SRTs. A possible explanation for this deviation is the aspect of amplitude modulation masking that occurs for the modulated maskers. Regarding informational masking (IM), which is thought to appear mostly in the speech-like maskers, the two studies in chapters 3 and 4 show different results. In the monaural study, there is no prominent effect of IM visible. This is most probably caused by the predictability of the target material (OLSA) and a possible separation of the target speech from the interfering talkers due to different fundamental frequencies of the speakers. Contrarily, using the same target and masker material, the binaural study shows that human listeners have great difficulties in recognizing the target speech in situations with the speech-like maskers. Observed SRTs for the case of a co-located, single interfering talker (ST), are almost positive and significantly higher than SRTs for a stationary masker (SSN). This suggests that it is especially difficult for human listeners to separate target speech from a masker, consisting of the same sentences uttered by a speaker of the same gender. According to Durlach et al. (2003b), Brungart (2001), Brungart et al. (2001), and Rosen et al. (2013), the reason for these high SRTs can be IM (due to similarity of target and masker), although, as discussed in chapter 4, it is not easy to exactly pinpoint this masking aspect. Regarding a possible binaural "better-ear glimpsing" (as proposed by Brungart and Iyer, 2012), the study in chapter 3 shows that glimpses do not suffice for true binaural cues. SRTs from a masker that provides better-ear glimpses from both ears (IMBM) are generally higher than SRTs observed in a masker with ILDs and IPDs. This suggests that binaural listening in humans cannot be explained alone by a "fast switching better-ear mechanism", where those glimpses from both ears are utilized that provide the larger SNR, but that IPD information is necessary.

5.2 Performance of speech prediction models in binaural and monaural listening conditions

Considering the study on speech recognition in binaural listening conditions, it is seen that predictions by a short-term binaural model[2] show a decrease in SRTs for modulated maskers, as is also seen in the listener's data. In contrast, this is not the case for the long-term analysis models[3]. The monaural study in chapter 4 also shows that a short-term analysis is necessary to account for decreased SRTs in fluctuating maskers and underlines the importance of this aspect in terms of modeling human speech perception. It can be stated that general trends in the observed data can only be captured when the speech

[1] speech detection thresholds
[2] BSIM, $\text{BSIM}_{\text{begl}}$, ADD
[3] $\text{BSIM}_{\text{long}}$ and SNR_{long}

prediction models incorporate a short-term analysis of the input signals.
In chapter 4, the mr-sEPSM (Jørgensen et al., 2013) shows a surprisingly poor prediction accuracy for SRTs in modulated maskers. Moreover, this model is very sensitive towards a spectral (mis-) match of target and masker material, which is neither the case for the other prediction models, nor for the observed SRTs. This inaccuracy is caused by an overestimation of high-frequency modulation filters, which appears to cause problems when the mr-sEPSM is applied to an arbitrary choice of target and masker material. The model's prediction accuracy can to some degree be improved by choosing another reference SRT than the SRT50 in a stationary masker, but in summary, it is found that the mr-sEPSM cannot be used "out of the box" to provide reliable speech recognition results. The statistically inspired STOI model generally fails in predicting SRTs in modulated and speech-like maskers. Findings from section 4.6.4 suggest that this model is not suited for signals that have no signal enhancement (i.e., time-frequency weighting or single-channel noise reduction) prior to the model analysis. Another surprising result is the good performance of the ESII, which provides the best agreement with observed data, despite its rather simple analysis approach. The ESII accounts well for "dip-listening" (Bronkhorst, 2000) and can explain much of the observed SRTs patterns. Generally, BSIM (Beutelmann et al., 2010) predictions show a good agreement with human speech recognition in the various symmetric masker conditions in chapter 3. If the binaural information in the masker is limited to level differences, a model version that accounts for a "better-ear" listening ($\text{BSIM}_{\text{begl}}$), disregarding the equalization-cancellation mechanism, is sufficient for predicting SRTs that are close to the observed data. But such ILD-based model processing fails when the masker also contains IPD information.
Considering the performance of models that generally account well for SRTs of the modulated maskers, it is found that a background of interfering talkers constitutes a challenging situation. SRTs for the speech-like maskers are underestimated by BSIM (and all its derivatives[4]) in chapter 3 and by ESII (and ESIIsen) in chapter 4 by about $10 - 15$ dB. It is tempting to ascribe this entirely to the lack of informational masking in the models, especially since Rhebergen and Versfeld (2005) state that "when speech-like maskers are used, it is expected that the obtained thresholds are [lower] than predicted by the extended SII model due to additional i.e., informational masking". But this consideration is too simple, because there can also be the lack of across-frequency processing or general analysis problems (as discussed in sections 3.4 and 4.6.4) that account for a mismatch of observed and predicted SRTs.
In summary, the results from chapters 3 and 4 show that the tested models (SII, ESII, mr-sEPSM, STOI, BSIM) cannot be easily applied to arbitrary listening conditions and do not provide reliable prediction data in all masker conditions. Considering the fact that speech recognition models are developed to replace measurements with human listeners, the complex maskers provided in this dissertation show that this is not yet possible. Instead, there are

[4] $\text{BSIM}_{\text{begl}}$, BSIM_{mon}, ADD

numerous listening conditions where the model predictions fail. In particular the maskers that show complex spectro-temporal features (such as the BB-SSN, AFS-SSN, ISTS or ST) provide conditions in which the current models reach their limits. But exactly those challenging conditions need to be tested in order to improve the performance of speech prediction models. Thus, the observed data provide an excellent benchmark to test future model versions. Another disadvantage of the current models is that all have to be adjusted to match a certain reference SRT, which leads to the question of how generally applicable these models are. There are (except for the SII and ESII) several parameters that need to be adjusted to reach a good agreement between the model outcomes and a reference SRT. But as long as this is the case, the predictions cannot be entirely objective. This undermines the idea of applying speech prediction models to universal listening conditions and therefore, an independence of reference should be incorporated in future model versions. A first step towards this direction is presented by Schädler et al. (2015b) and Kollmeier et al. (2015), where a reference-free automatic speech recognition approach is able to reasonably predict the performance of human listeners across various noise conditions and languages.

5.3 Extensions towards measurements with hearing-impaired listeners

In future studies, it would be of interest to apply the various spectro-temporal maskers in speech recognition measurements with hearing-impaired listeners. On the one hand, it may be hypothesized that hearing-impaired listeners would not show the same pronounced SRT patterns as normal-hearing listeners do. The SRT pattern would differ due to the loss of audibility, loss of dynamic range (recruitment), increase in internal noise, and factors involving binaural functions (Kollmeier, 1999; Marrone et al., 2008) that hearing-impaired listeners face. On the other hand, e.g., Micheyl et al. (2000) state that across-channel processing is unaffected in hearing-impaired listeners, suggesting similar SRTs for normal-hearing and hearing-impaired listeners in case of the modulated maskers from this dissertation. Therefore, the AFS- and BB-SSN (and SAM-SSN) maskers would serve as perfect masker material to verify or falsify this hypothesis.
By comparing observed and predicted SRTs from experiments with hearing-impaired listeners, it could be investigated to what extent the current models are able to account for speech recognition in this group of listeners. First steps towards the simulation of a hearing-impaired signal analysis could be, for example, the simulation of the increased hearing thresholds. Recent research by Scheidiger and Dau (2015) on the predecessor of the mr-sEPSM (sEPSM, Ewert and Dau, 2000) shows that much of the data from a speech recognition experiment (Christiansen and Dau, 2012) can already be explained by incorporating the individual hearing thresholds of the listeners. For SII and ESII (ESIIsen) predictions, the actual audiograms of the listeners can be loaded prior to the model analysis in the SIP-Toolbox (Kollmeier et al., 2011), so that the individual hearing loss is accounted for. Besides, a study

by Rhebergen et al. (2010) shows that extending the SII by implementing a compressive input-output function leads to speech reception predictions that show a better agreement with the observed data than the standard SII predictions. Adjusting this function to match the loss of compression in hearing-impaired listeners could allow a prediction of SRTs in experiments with these listeners. However, these suggestions are only first steps towards modeling SRTs of hearing-impaired listeners. They are mentioned here because they can easily be implemented in the existing model versions and hence serve as indicators of the model performance. These suggestions are by no means complete and do not replace a detailed investigation on the signal processing of listeners with impaired hearing.

On the whole, this dissertation examined human speech recognition in many different background noises. It was found that factors such as the coherence and regularity of masker modulations and the presence of interfering talkers significantly influence the observed speech reception thresholds. Regarding binaural speech perception, it was found that both ILDs and IPDs are necessary for a substantial speech perception in complex masking backgrounds. Much of the observed data could be explained by established knowledge, but there were also new questions raised. It is left to the course of science to establish a thorough understanding of informational masking and to refine speech prediction models in a way that human speech perception in complex listening conditions can be explained.

Appendices

Appendix A

Supporting material for Chapter 2

A.1 The rationalized arcsine transformation

The statistical analysis in chapter 2 was done on rationalized arcsine transformed units (rau), instead of a limited range of values such as percent correct. A limited range can be a problem for statistical analyses when percentages appear that are close to the upper or lower ends of the scale and violate the assumption of a normal distribution of values. The transformation to rau has the advantage that resulting values are numerically close to the original percentage scores, but retain the desired statistical properties of the arcsine transformation (normal distribution of the data). Thus, the outcomes of the transformation can be interpreted like percentage, although they are no percentage scores. Fig. A.1 shows an example of the transformation from percent correct to rau units. It is clearly visible that percentage and rau values are very similar for the range of $15-85\%$ and only deviate for numbers that are closer to the extremes. The equations used in chapter 2 are taken from Studebaker (1985) and read

$$T = \arcsin\sqrt{\frac{X}{N+1}} + \arcsin\sqrt{\frac{X+1}{N+1}} \tag{A.1}$$

$$R = a \cdot T - 23, \text{with } a = 46.47324337, \tag{A.2}$$

whereas T denotes the transformed observed scores from the measurement and R the transformation to rau. Equation A.1 was first proposed by Thornton and Raffin (1978) (and later by Mosteller and Youtz, 2006) and is supposed to be used for sample sizes smaller than 150. Thus, X denotes the number of samples observed and N the total number of samples.

A.2 Confusion matrices

Figures A.2 - A.6 show the confusion matrices that were calculated for experiment 1 and experiment 2 on vowel identification in chapter 2. The squares in

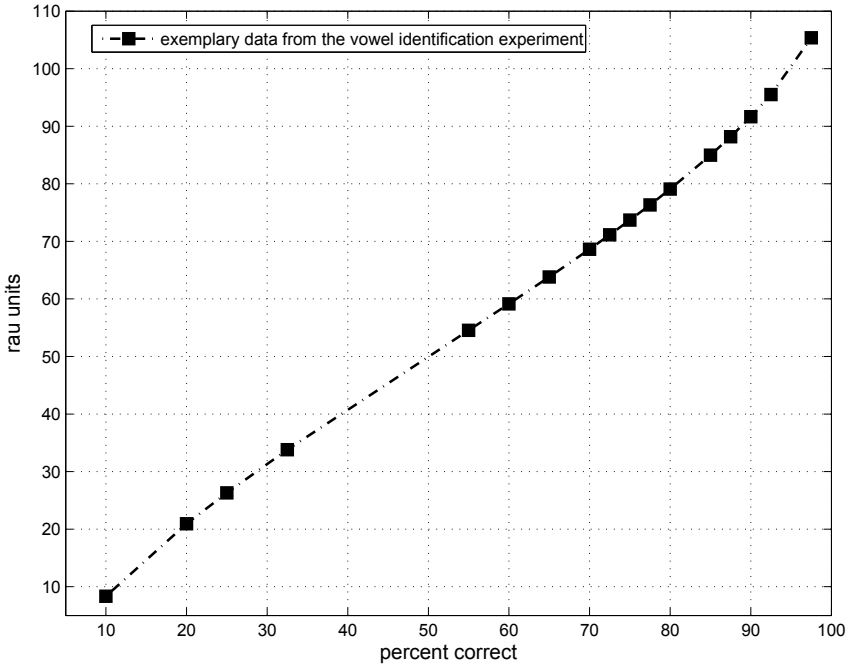


Figure A.1: Presentation of percentage versus rationalized arcsine transformed units for exemplary data of the study on vowel identification in chapter 2. The rau transformation maps the percentage values onto an open scale (rau units can be larger than 100) and provides rau values that are close to the original percentage valus for the range of $15 - 85\%$. The two values become increasingly different as as the two end of the percentage scale are approached.

each matrix denote the individual confusions and the numbers the normalized confusion rates. Black denotes a perfect identification, while white denotes no correct identification. Accordingly, shades of grey denote identification rates in between.
Comparing Figs. A.2 and A.3 it is apparent that vowel identification is generally larger for the higher SNR (-14 dB) than for the lower SNR. The confusions appear in sub-matrices for the vowels [e:,i:] and [o:,u:], but most vowels are identified correctly (large identification rates on the main diagonals of the matrices). This pattern of sub-matrices is more pronounced when low-pass filtered speech (LFS) instead of LPC-vocoded speech (LPC) is present in the low-frequency range of the stimulus. Besides, the pattern in most pronounced when band-pass filtered speech serves as a HF cue (HFS) and can be seen especially well in the case of the lower SNR. For this SNR it is also best visible that the vowel [a:] stands out from the sub-matrices, since there is no "confusion partner" present. This is caused by the position of the vowel in the formant triangle (Pätzold and Simpson, 1997) that is far apart from the other vowels tested in the study in chapter 2.
The main diagonal is even more pronounced in the case of vowel identification in experiment 2 (see Figs. A.4,A.5, and A.6). This is caused by the LFS in the low-frequency range that generally allows a better identification, but there are also confusions in the sub-matrices as found in experiment 1. When only high-frequency cues are present (lower rows of Figs. A.4 and A.5) there is no clear identification pattern. Instead, confusions arises equally for all vowels and prove that HF cues alone do not carry any valuable information on the vowels. In contrast, when HFS is presented as a cue (Fig. A.6), vowel identification shows the established pattern. The performance with the HFS cue alone is comparable to that of LFS alone, but the overall rates are lower (see Fig. 2.4).

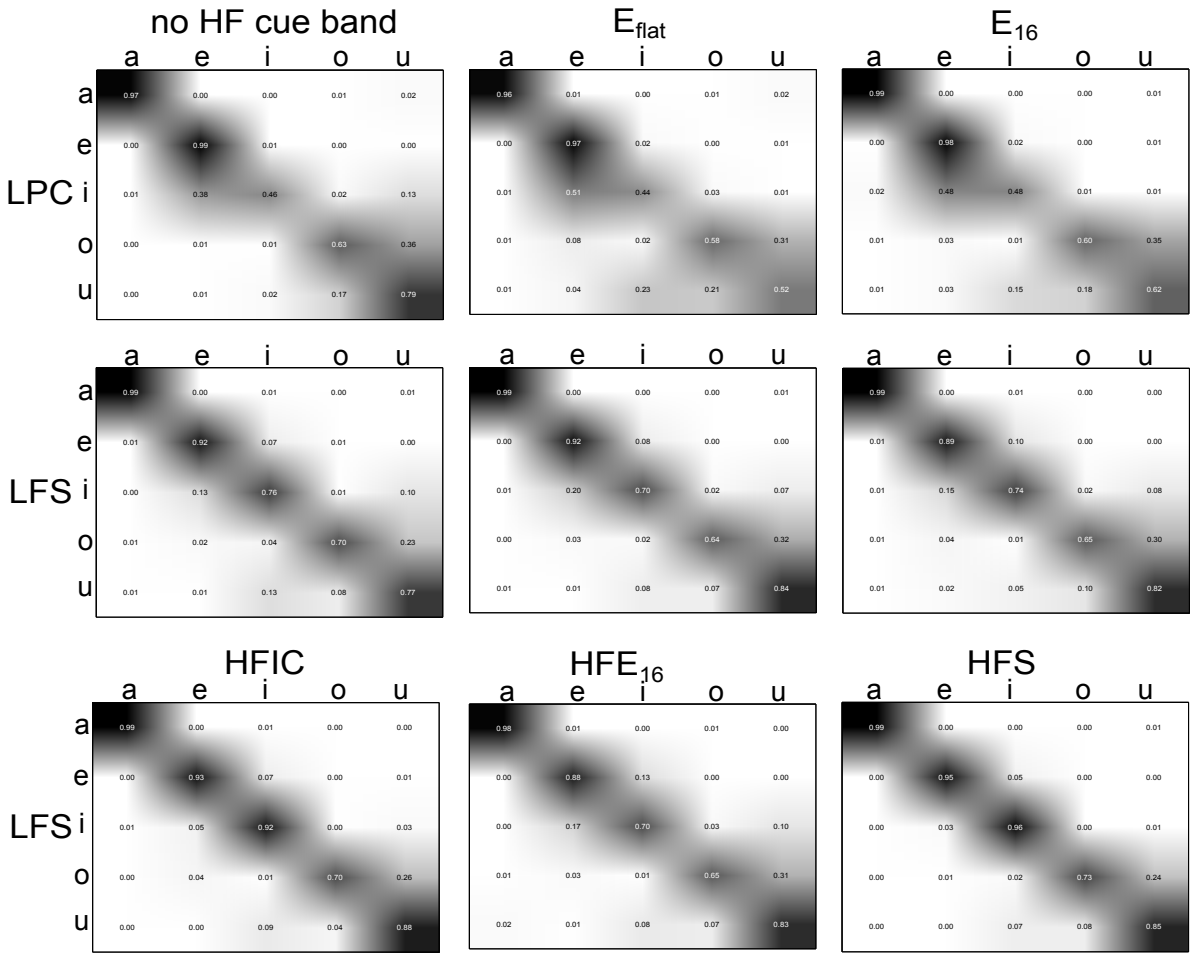


Figure A.2: Confusion matrix for experiment 1 in the study on vowel identification for SNR = −14 dB. The rows indicate if LPC or LFS was used in the low-frequency range of the stimulus and columns the type of high-frequency cue band that was presented. The color shading represents the identification rates. Black denotes perfect identification and white indicates no correct identification. The label on the left side of the matrix denotes those vowels that were presented to the listeners, the label on the upper side of the matrix denotes the vowels that were identified by the listeners. The normalized numbers in the matrices correspond to the percentage of this certain confusion.

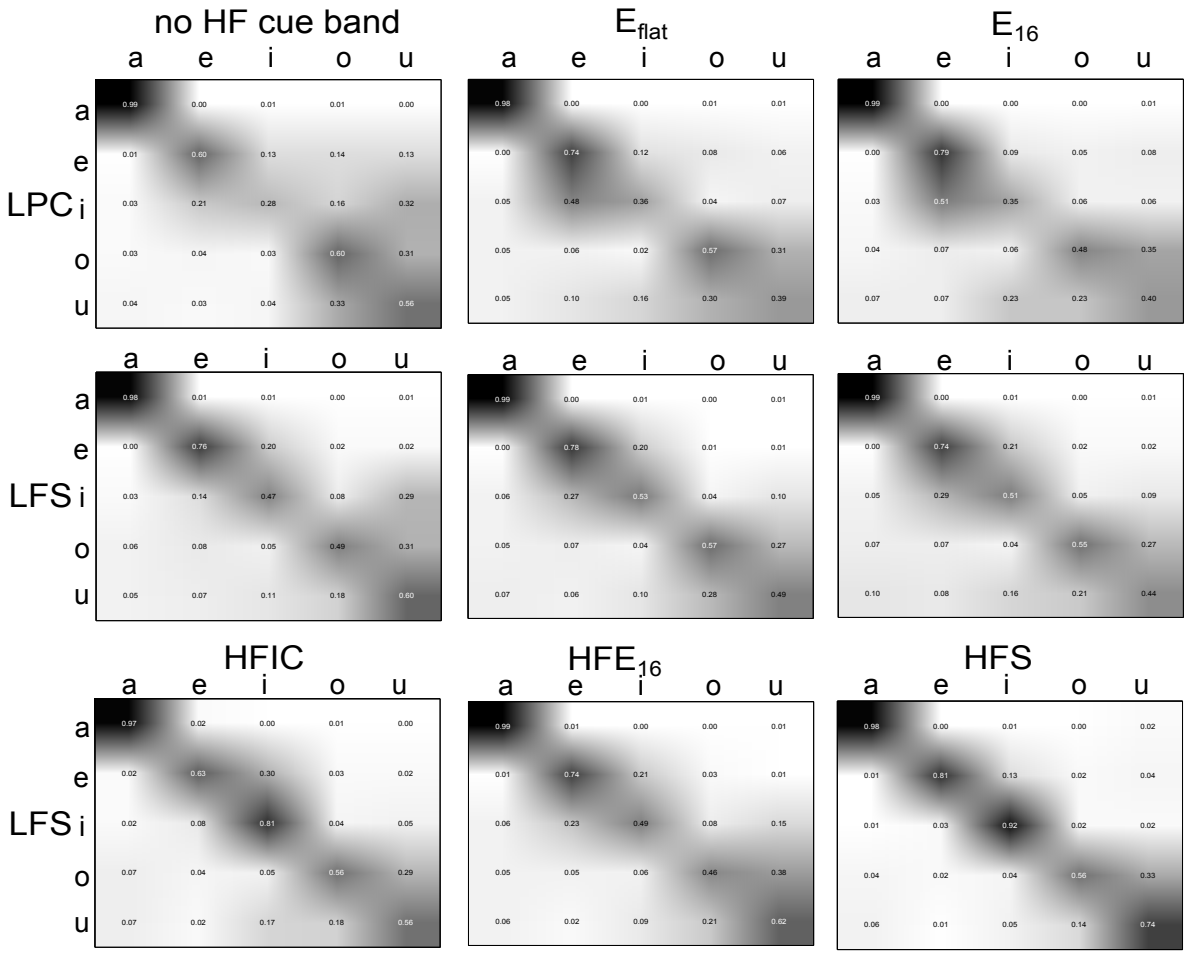


Figure A.3: Confusion matrices for experiment 1 in the study on vowel identification for SNR = −18 dB. The rows indicate if LPC or LFS was used in the low-frequency range and columns the type of high-frequency cue band that was presented. The color shading represents the identification rates. Black denotes perfect identification and white indicates no correct identification. The label on the left side of the matrix denotes those vowels that were presented to the listeners, the label on the upper side of the matrix denotes the vowels that were identified by the listeners. The normalized numbers in the matrices correspond to the percentage of this certain confusion.

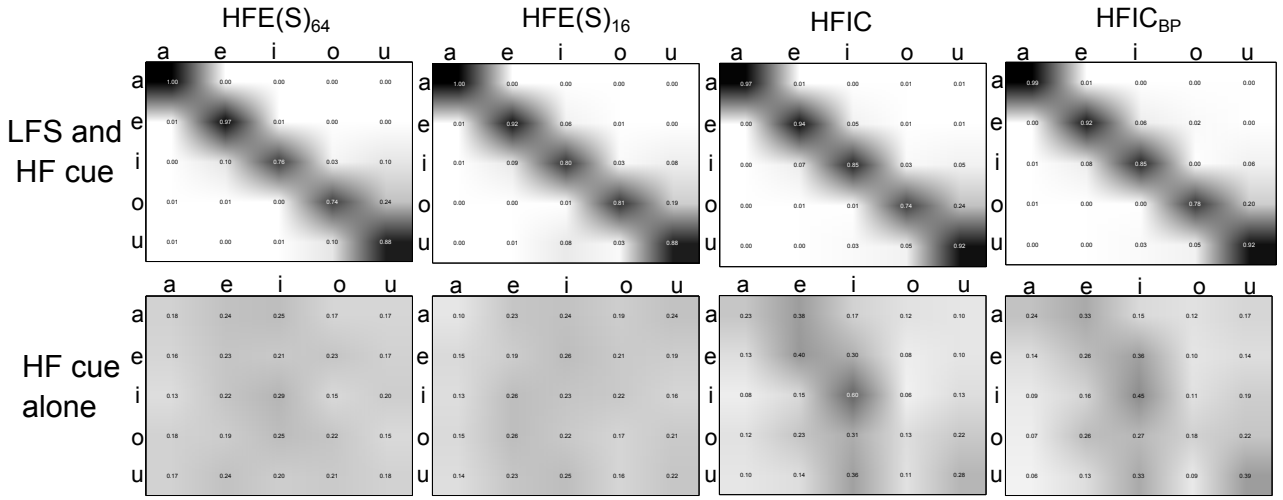


Figure A.4: Confusion matrices for SNR = −14 dB for experiment 2. The rows indicate whether LFS was present in the stimulus (upper row) or not (lower row) and columns indicate the individual HF cues. The color shading represents the identification rates (black indicates perfect identification, white no identification). The label on the left side of the matrices denotes the vowels that were presented to the listeners, the label on the upper side of the matrix denotes the vowels that were identified by the listeners. The normalized numbers correspond to the percentage of this certain confusion. While confusions show the characteristic pattern when LFS is present, HF cues alone do not lead to substantial identification, except for the case of HFS.

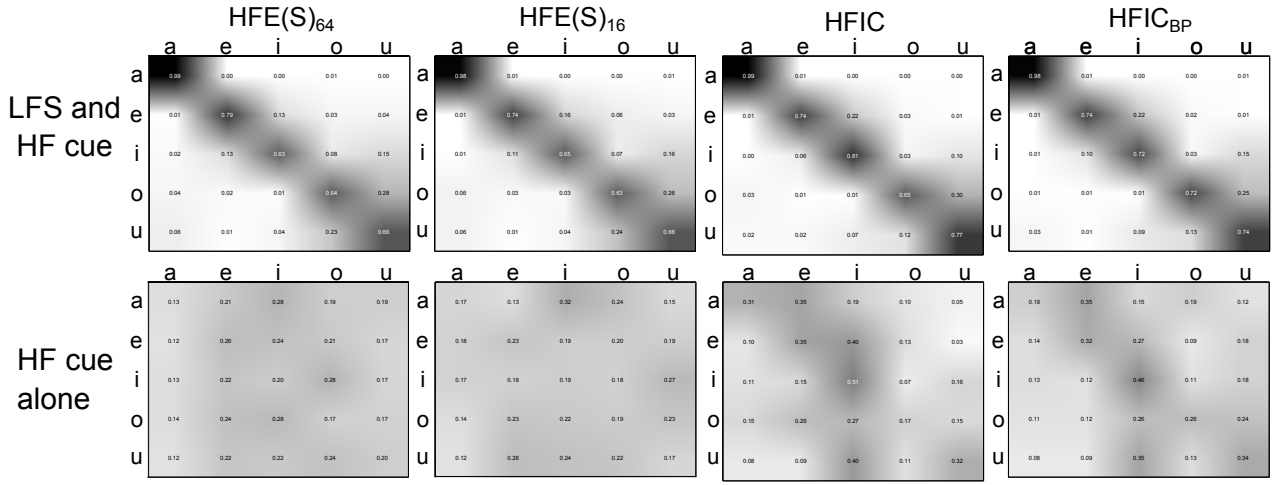


Figure A.5: Confusion matrices for SNR = −18 dB in experiment 2. The rows indicate whether LFS was present in the stimulus (upper row) or not (lower row) and columns indicate the individual HF cues. The color shading represents the identification rates (black indicates perfect identification, white no identification). The label on the left side of the matrices denotes the vowels that were presented to the listeners, the label on the upper side of the matrix denotes the vowels that were identified by the listeners. The normalized numbers in the matrices correspond to the percentage of this certain confusion. The patterns of confusions are similar to those from the higher SNR, but the sub-matrices are smeared out.

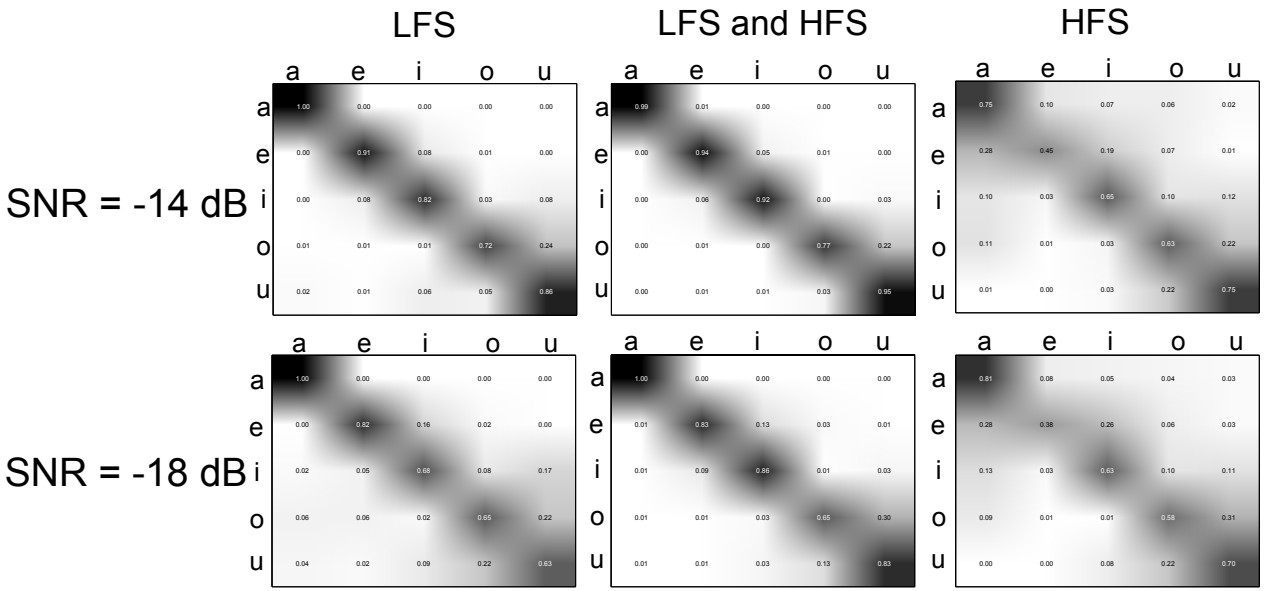


Figure A.6: Confusion matrices for both SNRs (SNR = −14 dB and SNR = −18 dB) the for the stimuli of experiment 2 when the intact low- and high-frequency speech parts are presented. The columns indicate whether LFS and HFS are presented alone or in combination. The color shading is the same as for Figs. A.4 and A.5

Appendix B

Supporting material for Chapter 3

B.1 Statistical differences in SRTs and MR across SSN-based and speech-like maskers

Table B.1 shows some of the statistically significant differences in SRTs that arise for the different masker types used in the study in chapter 3. The statistically different SRTs within the SSN-based and speech-like masker types were reported in section 3.3.1, but differences across these two masker groups are shown here. It is to be noted that differences arise only between the ST and all SSN-based maskers, but not between the $\text{ISTS}_{\text{male}}$ and the SSN-based maskers. This is most probably caused by the large amount of informational masking that is conveyed in the ST and leads to very high SRTs for this certain masker. As seen in Fig. 3.2, SRTs are about -3 dB for the ST masker, while they are in the range of -8 dB to -10 dB for the other (SSN-based and $\text{ISTS}_{\text{male}}$) masker types.
The MR that arises when masker sequences are independent in both ears, instead of partly correlated across the ears, was discussed in section 3.3.3. There, a statistical analysis was performed that investigated the differences in SRTs within the SSN-based and speech-like masker groups. Table B.2 shows the statistically different MR across these two groups. The MR is statistically different between the $\text{ISTS}_{\text{male}}$ as well as the ST and the SSN-based maskers (except the SAM-SSN). This suggests that an independence of the masker sequences lowers SRTs significantly for every kind of masking background.

masker configuration	HRTF condition	SSN-based masker	speech-like masker	
			$ISTS_{male}$	ST
co-location	$HRTF_{full}$	SSN		**
		SAM-SSN		***
		BB-SSN		***
		AFS-SSN		***
	ILD_{only}	SSN		***
		SAM-SSN		***
		BB-SSN		***
		AFS-SSN		***
	IMBM	SSN		**
		SAM-SSN		***
		BB-SSN		***
		AFS-SSN		**

Table B.1: Statistically significant differences in co-located SRTs across SSN-based and speech-like maskers for the HRTF conditions $HRTF_{full}$, ILD_{only}, and IMBM (Fig. 3.2, panels a-b), d)). Significant differences are marked with stars ($p < 0.05$ is *, $p < 0.01$ is **, and $p < 0.001$ is ***) and are determined with simple pairwise comparisons of the six masker types in each panel. Statistically significant differences within the SSN-based and speech-like maskers are presented in Fig. 3.2 and in section 3.3.1.

masker	HRTF condition	SSN-based masker	speech-like masker	
			$ISTS_{male}$	ST
independent masker sequences	$HRTF_{full}$	SSN	***	***
		SAM-SSN	***	**
		BB-SSN	**	**
		AFS-SSN	***	***
	IMBM	SSN	***	***
		SAM-SSN		**
		BB-SSN	**	***
		AFS-SSN	***	***

Table B.2: Results of the simple comparisons of MR for the different masker types for each HRTF condition in Fig. 3.5. Statistically different MR within SSN-based and speech-like maskers was presented in Fig. 3.5, while this table shows the different MR across SSN-based and speech-like maskers. Statistically significant differences are marked with stars ($p < 0.05$ is *, $p < 0.01$ is **, and $p < 0.001$ is ***).

Appendix C

Predicting recognition with a binaural speech intelligibility model (BSIM)

Binaural speech recognition was examined in chapter 3 with a variety of maskers showing different amount of spectro-temporal speech features and different binaural cues. In section 3.4 of that chapter, the binaural speech intelligibility model BSIM (Beutelmann et al., 2010) was used to predict observed SRTs in these different maskers. BSIM was used in a variety of versions, each representing different mechanisms of the human auditory system, but results were only shown for few representative listening conditions. In the following section, model predictions for the SRTs are shown for all HRTF conditions ($\text{HRTF}_{\text{full}}$, ILD_{only}, IPD_{mag0}, $\text{IPD}_{\text{mag60}}$, and IMBM) and all masker configurations (co-located and spatially separated). Moreover, predicted spatial release from masking (SRM), masking release due to the independence of masker sequences in both ears (MR), and the root-mean square errors (RMSEs) of the predictions are shown.

C.1 Model versions

The individual model versions in chapter 3 differ in their time constants, usage of ILD and IPD information and the stimulus manipulation prior to the model analysis. Predictions in chapter 3 are shown for the SNR_{long} analysis, four BSIM versions and the ADD approach. In this section, results of an additional model version, BSIM_{mon}, are shown. Table C.1 gives an overview of the individual analysis stages of each model version.
SNR_{long} and $\text{BSIM}_{\text{long}}$ are based on the long-term analysis of the masker signal, i.e. the SNR that is present when the entire masker sequence of 3 s (the length of the masker signal is chosen slightly longer than the duration of the OLSA sentences) is used in the signal analysis. While SNR_{long} is only based on the SNR at the listener's eardrum, $\text{BSIM}_{\text{long}}$ additionally contains the analysis stages described in Beutelmann et al. (2010). These include the equalization-cancellation (EC) stage, where the SNR is maximized according

model versions	time analysis		model stages	
	long-term	short-term	EC	BE
SNR_{long}	x			
$BSIM_{long}$	x		x	
BSIM		x	x	
$BSIM_{begl}$		x		x
$BSIM_{mon}$		x		
ADD[a]		x	x	

Table C.1: Overview of the model versions used for predicting binaural speech intelligibility in chapter 3. The models differ in the time constant of the analysis (full signal length or 23 ms time windows) and the usage of binaural cues. Binaural information is analyzed with an equalization-cancellation (EC-) and better-ear (BE-) stage. Results from $BSIM_{mon}$ have not been shown in chapter 3.

[a] Binaural summation prior to model analysis

to Durlach (1963), and a stage where the SNR in the analysis channels after the EC-processing is compared to the SNR in the analysis channels after the Gammatone filterbank processing. Those channels that provide the larger SNR are then further processed in the SII calculations of BSIM. In contrast to SNR_{long} and $BSIM_{long}$, all other model versions (BSIM, $BSIM_{begl}$, $BSIM_{mon}$, and ADD) contain a short-term analysis in 23 ms time frames (with 50% overlap). Thus, they are more suited for speech intelligibility predictions in a fluctuating masker. The BSIM setup used here is identical to the version presented in Beutelmann et al. (2010), whereas $BSIM_{begl}$ resembles an analysis based on binaural better-ear glimpsing only, where the EC-stage is disabled. Thus, $BSIM_{begl}$ contains only a better-ear (BE-) analysis stage. $BSIM_{mon}$ incorporates neither the EC- nor the BE- stage and thus it is not possible to select the better-ear in each time frame from either one of the two ears in that version. Instead, the output is restricted to one ear only to find the best possibly SNR. This procedure is equivalent to a short-term standard SII calculation and can be regarded as a simplified version of the ESII from Rhebergen et al. (2006). $BSIM_{mon}$ can serve as a baseline model in which the binaural processing is disabled but all other BSIM processing stages enabled. It is hypothesized that this configuration provides worse intelligibility predictions than configurations that enable a binaural processing.

The ADD approach resembles a simplistic binaural processing and uses a stimulus optimization prior to the short-term BSIM analysis. The two ear signals are added and then passed on to the model processing. In this case, interaurally correlated stimulus parts (mainly from the target speech) are ideally enhanced and their SNR is improved by 3 dB over uncorrelated stimulus parts (from the maskers). It is expected that ADD provides speech intelligibility predictions that are closer to the observed SRTs when compared to predictions by BSIM.

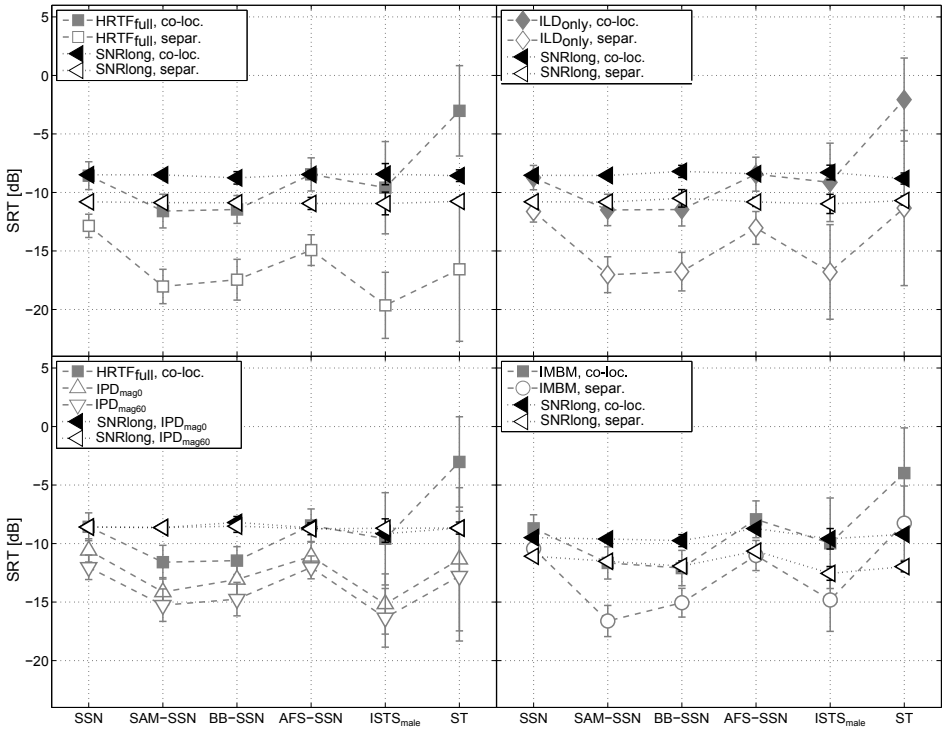


Figure C.1: $\mathrm{SNR}_{\mathrm{long}}$ predictions for binaural speech intelligibility in the different HRTF conditions and masker configurations along with the standard deviations. Observed SRTs from Fig. 3.2 are shown with gray and model predictions with black symbols. Closed symbols denote SRTs for the co-located masker, open symbols SRTs for the spatially separated masker.

C.2 Predictions of SRTs

C.2.1 Long-term prediction models

The presentation of the model data is the same for all figures: SRTs from the listening experiments in chapter 3 are depicted in gray, model predictions in black. Closed symbols denote the SRTs observed in a co-located masker, open symbols those observed in a spatially separated masker. Model predictions are shown in four sub-plots, whereas each denotes a different HRTF condition (Figs. C.1 - C.6). Figure C.7 shows the model predictions for the case of independent masker sequences in both ears. The color coding is the same as for the previous figures.

Figures C.1 and C.2 show SRT predictions of the two long-term analysis models for all HRTF conditions. Predicted SRTs are nearly constant across the individual masker types in all three HRTF conditions and the IMBM. This is expected, since both model predictions are based on the long-term energy of the input signals and this is very similar across the masker types. Considering Fig. C.1 ($\mathrm{SNR}_{\mathrm{long}}$), predicted SRTs for the spatially separated masker are generally $2-3$ dB lower than those of the co-located masker for all HRTF conditions (as is seen in the observed SRTs), except for the IPD

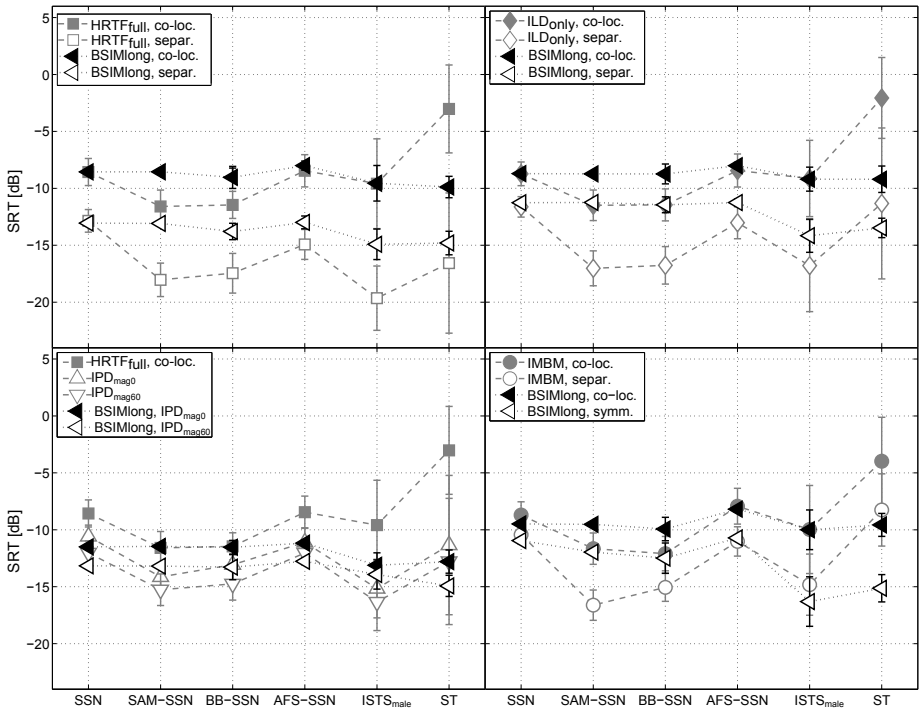


Figure C.2: $\text{BSIM}_{\text{long}}$ predictions and standard deviations for binaural speech intelligibility in the different masking conditions. The different panels denote the HRTF conditions and open and closed symbols the masker configurations (co-located and spatially separated). Model predictions are shown with black, observed SRTs with gray symbols.

condition. This discrepancy to the observed data can be explained by the lack of a filterbank analysis and band-importance function in SNR_{long}. In the generation of IPD_{mag0} and $\text{IPD}_{\text{mag60}}$, the original masker amplitude in each auditory filter is changed, meaning that certain filters have a higher masker amplitude and others a lower amplitude than before. If speech would be analyzed in auditory filters in SNR_{long}, the contribution of each filter would be different for IPD_{mag0} and $\text{IPD}_{\text{mag60}}$ and there would be a difference in predictions for these two HRTF manipulations.
$\text{BSIM}_{\text{long}}$ incorporates a Gammatone filterbank analysis and the SPIN (speech-in-noise) band-importance function (rightmost column in Table B1 in ANSI, 1997) and weights the contribution of each frequency channel differently. $\text{BSIM}_{\text{long}}$ predictions show a 2 dB difference for the two IPD conditions, which is closer to the observed SRTs. But the overall pattern slightly overestimated SRTs, suggesting that the analysis of IPD and ILD cues is not optimal in $\text{BSIM}_{\text{long}}$. Another difference towards SNR_{long} predictions are the SRTs for the $\text{HRTF}_{\text{full}}$ condition: The SRTs for the spatially separated maskers that are predicted by SNR_{long} are about 2 dB lower than those of the co-located maskers, while SRTs predicted by $\text{BSIM}_{\text{long}}$ are about 5 dB lower. Predictions by $\text{BSIM}_{\text{long}}$ are thus closer to the observed SRTs.

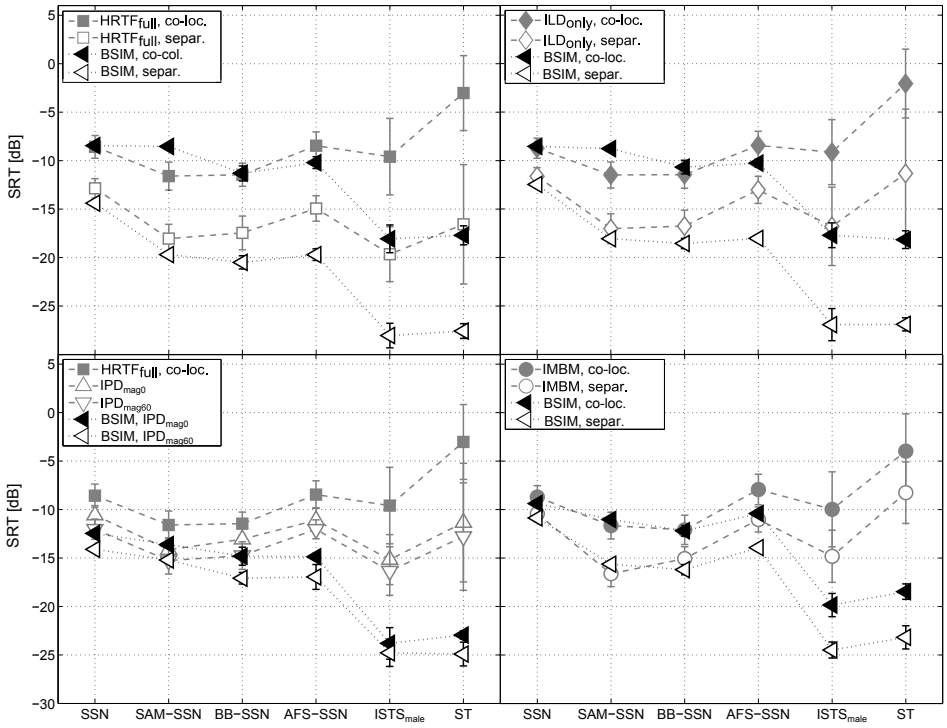


Figure C.3: BSIM predictions and standard deviations for binaural speech intelligibility in the different masking conditions. The different panels denote the HRTF conditions and open and closed symbols the masker configurations (co-located and spatially separated). Model predictions are shown with black, observed SRTs with gray symbols. The decrease in SRTs for the speech-like maskers is well captured with this model configuration.

Moreover, $\text{BSIM}_{\text{long}}$ shows a slight decrease in SRTs for speech-like maskers, when the long-term masker level is reduced due to the temporal gaps, which is also not captured in SNR_{long}. Taken together, a long-term analysis fails to correctly predict SRTs in fluctuating and speech-like maskers.

C.2.2 BSIM

Contrarily, BSIM describes the decrease in SRTs for modulated maskers (see Fig. C.3) well. SRTs for the spatially separated masker are slightly underestimated, but the general pattern for this masker position is captured. There are two difficulties, however, for this model version. Firstly, there is a general underestimation of the masking exhibited by the speech-like maskers. Predicted SRTs are generally $10 - 15$ dB too low. Secondly, the model fails to correctly predict the decrease in SRTs from the SSN to the SAM-SSN in the case of co-location of the masker. This is most probably caused by the crosstalk in the two ear signals due to the fixed phase difference (90°) of the SAM-SSN. Troughs from one masker sequence are filled by hills of the second sequence, so there are no time frames in the analysis where the

target is perceived in absolute quiet. Consequently, the resulting SII values and predicted SRTs cannot be arbitrarily low. The dip listening that can principally be accounted for by BSIM is thus overruled by the crosstalk in the two ear signals. Crosstalk is reduced when the masker is spatially separated from the target (open symbols) and for that case the predicted SRTs match the observed SRTs much better. To conclude on this issue, further research could be done with maskers that are gradually moved from co-location to a spatially separated position and where the amount of crosstalk is monitored. Besides, it could be investigated in how far the mean averaging of SII values from each time frame affects the outcomes for SSN and SAM-SSN maskers. It can be hypothesized that replacing the mean average by a procedure where frames with a large SII (masker gap) are weighted more than those with a small SII (masker hill) would lead to an actual difference in SRTs for SSN and SAM-SSN also in case of the co-located masker. BSIM predictions for the IMBM are generally close to the observed SRTs, supporting the findings from chapter 3 that, once glimpsing is simulated, the information about the glimpses can be optimally used by the listeners and the model.
Despite an overall better agreement between observed data and predictions by BSIM, there is an overestimation of SRTs for co-located and underestimation of SRTs for spatially separated maskers. A possible explanation for the underestimation could be the "binaural sluggishness" (Culling and Colburn, 2000) of the human auditory system, which is not accounted for by BSIM and its variations ($\text{BSIM}_{\text{begl}}$, BSIM_{mon}, and ADD). The analysis window of 23 ms in the model versions could be too short to account for the perception of changes in the interaural cues that human listeners experience. An extension to the existing BSIM version (Beutelmann et al., 2010) could be a change in the time constant of the EC-stage to a slightly larger value (e.g., 100 ms) as was already proposed by Rhebergen and Versfeld (2005). This could possibly decrease the discrepancy of observed and predicted thresholds as binaural sluggishness is then accounted for.

C.2.3 $\text{BSIM}_{\text{begl}}$

Figure C.4 shows predictions from the $\text{BSIM}_{\text{begl}}$ model version. The analysis in the BE-ear stage is based on ILD information only, therefore the predictions are expected to be worse for the $\text{HRTF}_{\text{full}}$ condition compared to BSIM predictions. For the ILD_{only} and IMBM condition, however, predictions by $\text{BSIM}_{\text{begl}}$ should be very similar to those of BSIM.
The upper left panel of Fig. C.4 shows the predictions for the $\text{HRTF}_{\text{full}}$ condition and these are indeed slightly higher than those of the same panel in Fig. C.3. $\text{BSIM}_{\text{begl}}$ predictions show higher SRTs when IPD information is neglected in the model analysis, but present in the masker. Despite this effect, the overall fit for the SRTs in a spatially symmetric masker is better for $\text{BSIM}_{\text{begl}}$ than for BSIM, as seen, for example, in the upper right panel for the case of ILD_{only}. For the IPD masker configuration (lower left panel), predictions by $\text{BSIM}_{\text{begl}}$ are about 5 dB too high due to the disabling of the EC-stage, but the overall pattern is identical to the observed SRTs. Model predictions for the IMBM are ruled by the ILD information that is available

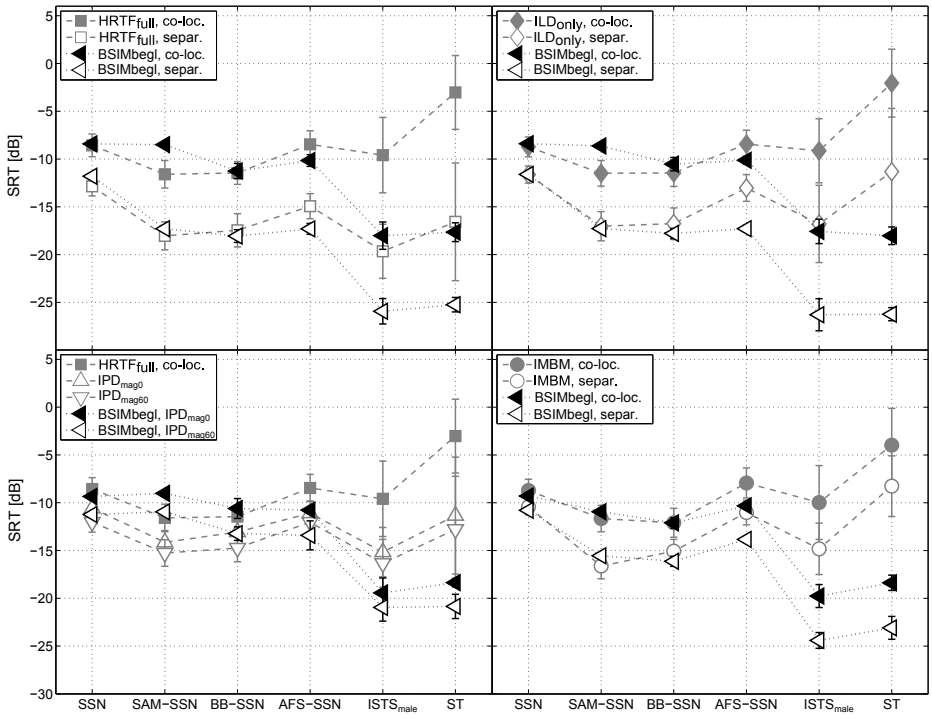


Figure C.4: $\mathrm{BSIM_{begl}}$ predictions and standard deviations for binaural speech intelligibility in the different masking conditions. The different panels denote the HRTF conditions and open and closed symbols the masker configurations (co-located and spatially separated). Model predictions are shown with black, observed SRTs with gray symbols. Predicted SRTs for the IMBM masker are identical to those of BSIM and predicted SRTs for the IPD masker condition are shifted upwards. The offset between observed and predicted SRTs in the spatially separated masker configuration is reduced for $\mathrm{HRTF_{full}}$ and $\mathrm{ILD_{only}}$, despite the disabling of the EC-stage.

in the glimpses and indeed, prediction for BSIM and $\text{BSIM}_{\text{begl}}$ are almost identical for the IMBM condition.

C.2.4 BSIM_{mon}

Fig. C.5 shows SRT predictions that arise when the usage of binaural cues is disabled and consequently, the analysis in BSIM similar to a simplified ESII. Thus, when comparing BSIM_{mon} to BSIM predictions, the accuracy of the predictions is expected to be worse for BSIM_{mon}. Indeed, predicted SRTs are about 5 dB higher than those of BSIM for the spatially separated maskers in the $\text{HRTF}_{\text{full}}$ and ILD_{only} conditions. For a co-located masker, the predictions of BSIM_{mon} are similar to BSIM. Considering predicted SRTs for the IMBM, BSIM_{mon} yields predictions that are almost identical to those of $\text{BSIM}_{\text{begl}}$ and BSIM. This is, because the glimpses in the IMBM already provide the best possible SNR in each analysis frame. Thus, BSIM_{mon} works on an ideally pre-processed stimulus. In general, the SRTs for the spatially separated maskers are overestimated with this model version, as is especially apparent for the two IPD manipulations in the lower left panel of Fig. C.5 and in the upper two panels .

C.2.5 ADD

SRT predictions by the ADD approach are shown in Fig. C.6. ADD predictions show the same pattern as predictions by BSIM (and $\text{BSIM}_{\text{begl}}$), but the overall shift of the pattern is different for the individual HRTF conditions for the case of the spatially separated maskers. For the co-located maskers, the predictions are almost identical to those of BSIM and $\text{BSIM}_{\text{begl}}$. Comparing ADD and BSIM predictions for the case of the $\text{HRTF}_{\text{full}}$ condition, the possible enhancement of coherent target speech parts generally moves the model predictions to higher SRTs. A similar behavior is also seen for the ILD_{only} condition and this leads to an overestimation for some of the modulated SSN-based masker SRTs (i.e., SAM-SSN and BB-SSN). Comparing the predictions for the two IPD manipulations, the results are very similar for ADD and BSIM. This suggest that ADD can, to a large degree, account for the use of IPD information as done in the EC-stage in BSIM, but provides a much simpler approach to binaural processing of IPD information. However, the use of ILD information (as seen in SRTs for $\text{HRTF}_{\text{full}}$ and ILD_{only}) cannot be accounted for by the ADD approach. Overall, a binaural summation approach only changes the overall shift of the prediction pattern, but does not characteristically improve predictions for certain maskers.
Summarizing, much of the observed SRTs in binaural listening conditions (Figs. C.1 - C.6) can be explained with a short-term energy analysis as proposed in Beutelmann et al. (2010). SRT predictions for the case of a co-located masker are very similar across the different model versions. This is because the interaural cues (ILD, IPD) are small when a signal is in a frontal position and thus, the influences on binaural speech reception is limited. Regarding SRT predictions for the speech-like maskers, the SRTs are constantly underestimated. This suggests (as already mentioned in section

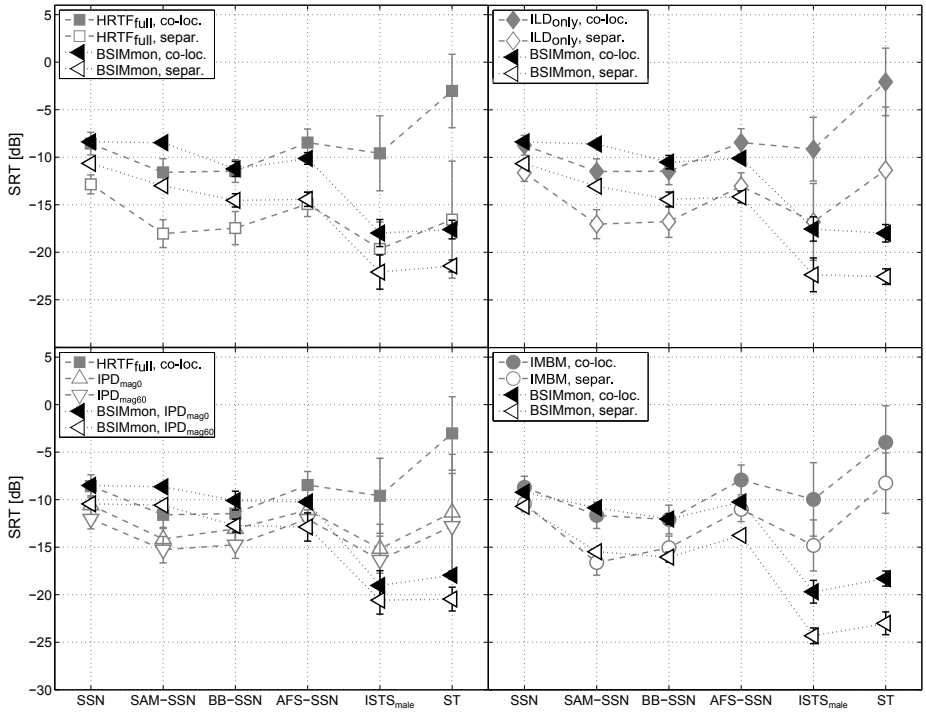


Figure C.5: BSIM_{mon} predictions and standard deviations for binaural speech intelligibility in the different masking conditions. The different panels denote the HRTF conditions and open and closed symbols the masker configurations (co-located and spatially separated). Model predictions are shown with black, observed SRTs with gray symbols. The EC- and BE-stage are omitted in this configuration, thus the model is expected to show a worse prediction accuracy than configurations where the binaural cues are fully utilized.

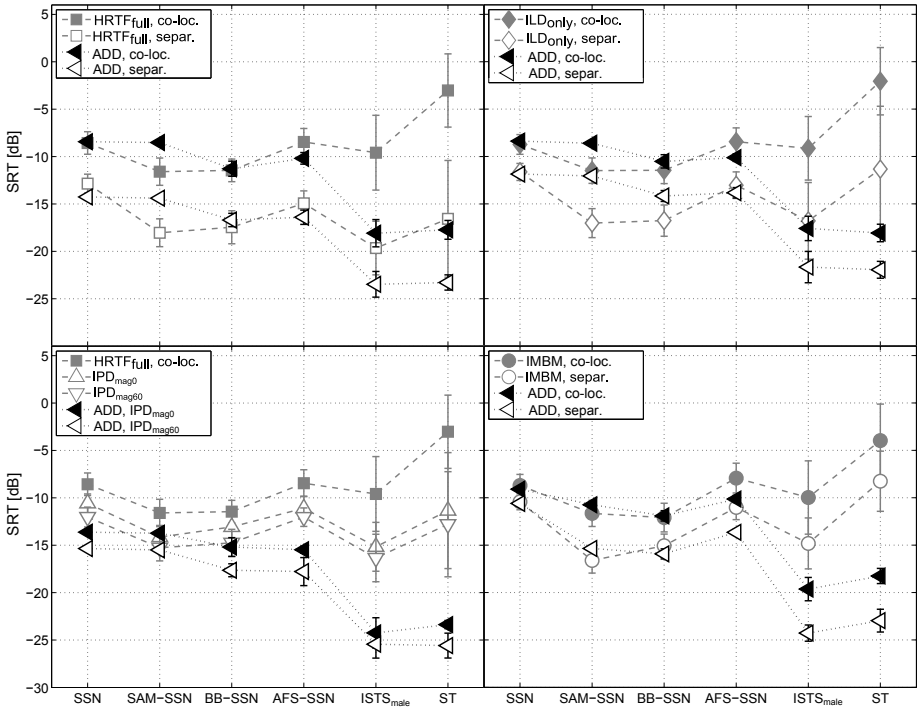


Figure C.6: ADD predictions and standard deviations for binaural speech intelligibility in the different masking conditions. The different panels denote the HRTF conditions and open and closed symbols the masker configurations (co-located and spatially separated). Model predictions are shown with black, observed SRTs with gray symbols.

3.5.5) that the aspect of informational masking is not captured in the model versions.

C.2.6 Independent masker sequences

Model predictions for the case of independent masker sequences in both ears are shown in Fig. C.7. Predictions are very different for the various models, suggesting that certain versions are not able to account for this listening situation. Generally, it is found that a short-term analysis is necessary to capture the overall pattern of SRTs that arises for the different maskers, as was already seen in the predictions for the partly correlated maskers. Both long-term analysis models yield similar SRTs across the individual masker types. Comparing SNR_{long} and $\text{BSIM}_{\text{long}}$ predictions in detail, it is again found that a band-importance function leads to predicted SRTs that are closer to the observed data. For BSIM, there is now a difference between the SSN and SAM-SSN thresholds for the case of a co-located masker, which is most probably caused by the missing crosstalk. The missing crosstalk also influenced the SRTs for the independent masker sequences in general. These are overall lower than those of the spatially separated maskers (i.e., Fig. C.1). As for SRTs with correlated masker sequences, predictions for BSIM and $\text{BSIM}_{\text{begl}}$ are very similar. This is, because there are no IPDs for the independent masker sequences and consequently the EC-stage cannot improve the SNR and does not influence the predictions. Regarding predictions by BSIM_{mon} it is apparent that predictions do not match observed data well for the case of the independent $\text{HRTF}_{\text{full}}$ masker. This is most likely caused by the lack of the BE-stage in this model version. Interestingly, the pattern of predictions is similar to that of $\text{BSIM}_{\text{long}}$ and suggests that BSIM_{mon} is "in between" the long- and short-term version of BSIM.
Altogether, regarding predictions for the individual masker types, it is found that predictions are too low for the two speech-like maskers in each model version. As stated earlier (see section 3.5.5), this could be caused by informational masking that is not captured in any of the presented approaches.

C.2.7 RMSE for the model SRTs

Differences in the accuracy of the different model predictions can be seen in Figs. C.1 – C.7, but can also be numbered by calculating the root-mean-square errors (RMSEs) for each model version. Table C.2 shows the RMSEs that were calculated across the different masker types for each version. Each row represents a certain model and each column denotes a certain masker condition. The lowest and largest values for each masker condition are highlighted.
These numbers have, however, to be regarded carefully, because deviations between observed and predicted SRTs can be very different across the individual masker types. RMSEs for the co-located $\text{HRTF}_{\text{full}}$ masker are, for example, quite low for SNR_{long} and $\text{BSIM}_{\text{long}}$, but the reason is that the deviations from the observed data are similar for each masker type (see Figs. C.1 and C.2) and a mismatch for certain masker types is canceled out. In

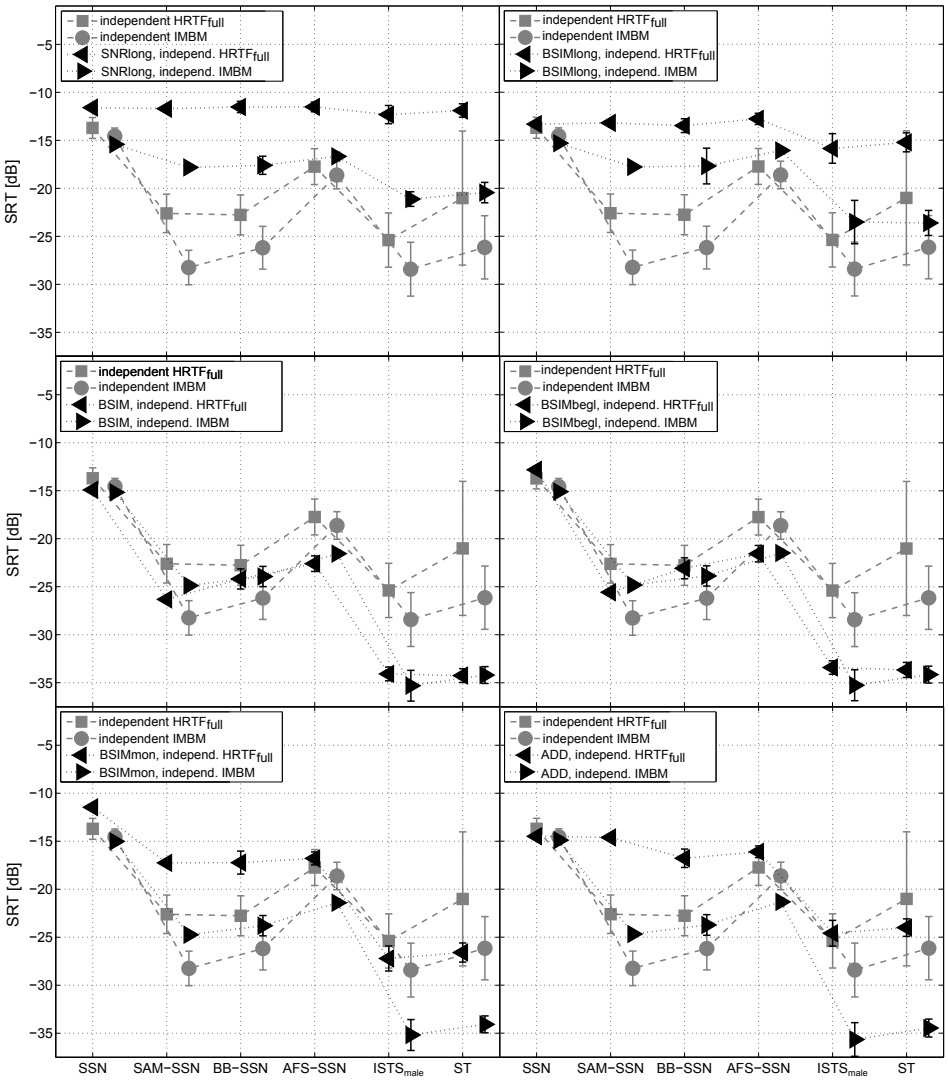


Figure C.7: Model predictions of SRTs for independent masker sequences in both ears. The predicted thresholds are shown along with the corresponding standard deviations. Observed SRTs are shown with gray symbols that correspond to those from Fig. 3.4 in chapter 3. Model predictions for the independent $\text{HRTF}_{\text{full}}$ condition are shown with left-pointing triangles, predictions for the independent IMBM with right-pointing triangles.

contrast, BSIM, which provides a good agreement for SRTs in the SSN-based maskers, has the largest RMSEs for the $\text{HRTF}_{\text{full}}$ condition. The reason for that is the large underestimation of SRTs for the speech-like maskers (see Fig. C.3) and this greatly influences the calculated RMSE. Another example is the RMSE for the spatially separated $\text{HRTF}_{\text{full}}$ masker: For this certain masker setup, BSIM_{mon} shows the lowest RMSE, although observed SRTs for the SSN-based maskers are clearly not met very well (see Fig. C.5). But the SRTs predicted by BSIM_{mon} for the speech-like maskers are much closer to the observed data, compared to all other model configurations. This leads to the overall lowest RMSE.
Regarding the RMSEs for the independent masker sequences, it is apparent that the errors are generally larger compared to those for the correlated maskers (both co-located and spatially separated). This reflects the fact that SRT predictions are worse in this certain masker condition in the first place. The RMSEs presented in this section are calculated across all masker types, because the study in chapter 3 investigates, among other things, if the used model versions can model listener's data at all. For a detailed view on the reliability of the calculated RMSEs and the accuracy of the model predictions for the individual masker types and conditions, RMSEs should be calculated for SSN-based and speech-like maskers separately, as deviations are generally larger for the latter masker types.

model version	co-located masker			spatially separated masker					independent maskers	
	$HRTF_{full}$	ILD_{only}	IMBM	$HRTF_{full}$	ILD_{only}	IPD_{mag0}	IPD_{mag60}	IMBM	HRTF	IMBM
SNR_{long}	**2.86**	**3.31**	**2.53**	6.11	4.43	4.23	5.49	**3.04**	**9.51**	**6.74**
$BSIM_{long}$	**2.86**	3.32	2.62	6.11	**3.57**	**1.67**	**1.77**	3.61	7.34	6.04
BSIM	**7.07**	**7.57**	**7.24**	**6.17**	**7.91**	6.20	6.48	**7.38**	6.98	4.79
$BSIM_{begl}$	7.04	7.51	7.18	4.51	7.44	4.11	4.28	7.33	6.44	4.77
$BSIM_{mon}$	7.02	7.49	7.14	**3.38**	5.48	4.13	4.21	7.28	**4.08**	**4.74**
ADD	7.08	7.52	7.11	3.60	5.30	**6.58**	**7.06**	7.25	4.33	4.96

Table C.2: Root-mean-square errors (RMSE) in dB for model predictions of the speech reception thresholds presented in Figs. C.1 - C.7. Model predictions for the different HRTF conditions are presented in columns for each model version. Smallest and largest RMSEs, denoted in bold, are found for both long-term analyses and BSIM (in case of correlated masker sequences). For the independnet masker sequences in both ears, the smallest RMSEs are found for $BSIM_{mon}$ and largest for the SNR_{long} analysis. The RMSEs only correspond to the overall match of model predictions to the data, but do not resemble exact matches to certain masker types.

C.3 Predictions of SRM and MR

Although SRTs are not always optimally predicted by the various models, the resulting spatial release from masking (SRM) and masking release (MR), due to the independence of the masker sequences, are shown for each model version in Figs. C.8 and C.9. This is reasonable, because even if systematic errors or misconceptions arise in the predictions of the SRTs, these are removed when SRT differences are considered. It is thus still interesting to discuss SRM and MR that is predicted by the various model versions. Open symbols in Fig. C.8 denote the observed SRM from the listening experiments (Fig. 3.3) and closed symbols the model SRM, calculated as the difference between predicted SRTs from the co-located and spatially separated masker configuration (as done for the SRTs in Fig. 3.3). The symbols correspond to the individual HRTF conditions in Fig. 3.3. As stated in section 3.3.2, SRM for the two IPD conditions is obtained by subtracting the predicted SRTs from the IPD_{mag0} ($\text{IPD}_{\text{mag60}}$) masker from the SRT from the co-located $\text{HRTF}_{\text{full}}$ masker. The MR is calculated by subtracting the predicted SRTs from the independent masker sequences from those of the co-located masker versions.

The two long-term models (upper panels of Fig. C.8) show a SRM that is very similar for each masker type. The reason are the predicted SRTs, which are also similar due to a similar overall long-term energy of the individual masker types. As for the SRTs, $\text{BSIM}_{\text{long}}$ shows a difference for the two IPD conditions, while SNR_{long} does not. Generally, $\text{BSIM}_{\text{long}}$ shows larger SRM than SNR_{long} and thus a predicted SRM that is closer to the observed data. But as for the SRTs, there is a large deviation of predictions from the observed data for the speech-like maskers. This leads to a predicted SRM for speech-like maskers that is much too small (about 6 dB compared to observed 10 dB).

Generally, predicted SRM is larger for the short-term model BSIM (middle left panel) than for the long-term models, but the resulting SRM in the $\text{HRTF}_{\text{full}}$ and ILD_{only} conditions is largely overestimated. This is caused by the underestimation of about 2 dB for the SRTs in the case of a spatially separated masker (see Fig. 3.6). While the absolute value of the predicted SRM for $\text{HRTF}_{\text{full}}$ and ILD_{only} is too large, the relative difference between the two is similar to the observed data. Observed SRM is always larger for the $\text{HRTF}_{\text{full}}$ than for the ILD_{only} condition, supporting the hypothesis that both ILD and IPD cues are used in binaural processing and that this is correctly captured in BSIM. The best match for BSIM predictions arises for the IMBM. Here, the observed SRM is met for most masker types due to the good agreement between predicted and observed SRTs (see Fig. 3.6).

Considering predicted SRM by $\text{BSIM}_{\text{begl}}$ in the middle right panel, the same conclusions can be drawn as in considering the predicted SRTs. Since $\text{BSIM}_{\text{begl}}$ only utilizes ILD information, the predictions are identical to those of BSIM for the ILD_{only} condition and the IMBM. Consequently, predicted SRM is smaller than the observed SRM for $\text{HRTF}_{\text{full}}$ and the two IPD conditions, as IPD information is disregarded in the $\text{BSIM}_{\text{begl}}$ analysis. More

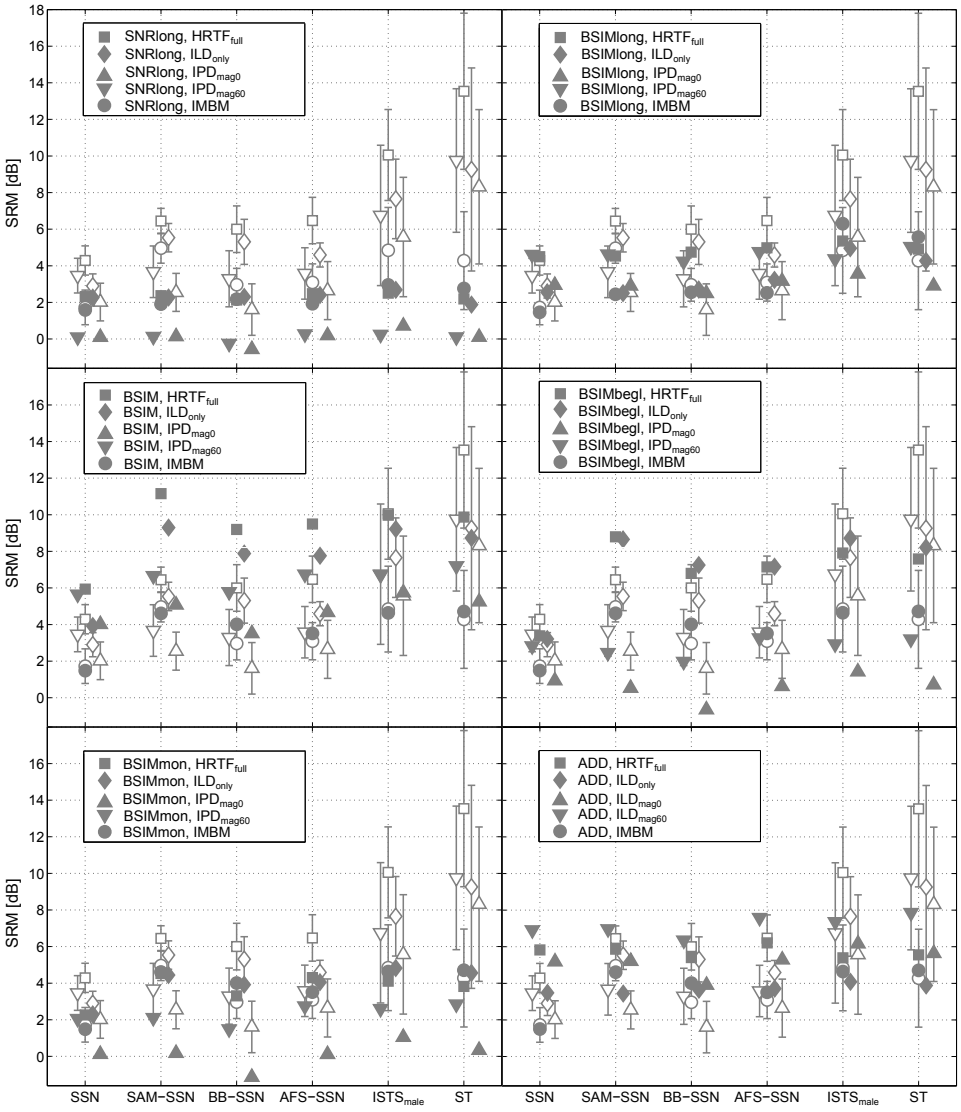


Figure C.8: Model predictions for the spatial release from masking (SRM) due to the spatial separation between target and masker. The different panels show the predictions for the individual model versions. Open symbols correspond to the observed SRM in Fig. 3.3, while closed symbols denote the model predictions. The different symbols denote the different HRTF conditions. Model SRM is shown without error bars.

precisely, predicted SRM for the IPD_{mag0} is almost absent and only very small for the $\text{IPD}_{\text{mag60}}$. This can be explained by the difference in spectral coloration of IPD_{mag0} and $\text{IPD}_{\text{mag60}}$. The IPD_{mag0} has the same amplitude spectrum as the co-located $\text{HRTF}_{\text{full}}$ condition. Predicted SRTs are similar for both and the subtraction leaves about zero dB as a SRM. This is different for the $\text{IPD}_{\text{mag60}}$, which has a different amplitude spectrum than the $\text{HRTF}_{\text{full}}$ masker. Thus, $\text{BSIM}_{\text{begl}}$ does predict a (small) SRM for this certain masker. It is to be noted, however, that the predicted SRM for the SSN masker is generally met better with $\text{BSIM}_{\text{begl}}$ than with BSIM for the various HRTF conditions.
Predictions by BSIM_{mon} are shown in the lower left panel of Fig. C.8. Here, predictions are clearly worse than for BSIM and $\text{BSIM}_{\text{begl}}$, but better than for $\text{BSIM}_{\text{long}}$, as was expected. This version bridges the gap between the long-term and short-term BSIM version, but regarding the predicted SRM it is found that binaural SRM cannot be explained by a monaural analysis approach alone. It is to be noted, however, that the monaural effect of spectral coloration (between IPD_{mag0} and $\text{IPD}_{\text{mag60}}$) is captured correctly in BSIM_{mon}.
The lower right panel of Fig. C.8 shows model predictions by ADD. Here, the spread of predicted SRM is even smaller than for $\text{BSIM}_{\text{begl}}$, most SRM ranges between $4-8$ dB, regardless of masker type or HRTF condition. ADD predictions meet roughly the observed data for all SSN-based maskers, but are clearly too small for speech-like maskers. As for the other panels in Fig. C.8, the best match between model and observed data arises for the SRM that appears in the IMBM.
In general, SRM predictions do not perfectly match observed data for any of the masker types or conditions, as can be expected considering the predicted SRTs. However, some aspects are recurring:
SRM is in general underestimated for speech-like maskers, especially for the ST, which means that human listeners have a larger advantage of the spatial separation of target and masker than is captured in BSIM or any of the other model versions. This suggests an underestimation of the aspect of informational masking, as was already reasoned in section 3.5.5. For the SSN-based maskers, there is often an overestimation of SRM, which is especially the case for BSIM (modulated SSN-based maskers) and $\text{BSIM}_{\text{begl}}$ ($\text{HRTF}_{\text{full}}$, ILD_{only}). Generally, BSIM and $\text{BSIM}_{\text{begl}}$ predict a larger SRM when only ILD cues instead of only IPD cues are present in the masker, which is in line with the findings from the listener's data and could suggest a priority of ILD over IPD cues. But this is contrasted by considering the SRM for the $\text{HRTF}_{\text{full}}$ condition. Comparing all HRTF conditions, $\text{HRTF}_{\text{full}}$ provides the largest SRM and this suggests an interplay of both interaural cues (ILDs and IPDs).

The predicted MR is shown in Fig. C.9 with open symbols, while MR from the listening experiment (see Fig. 3.5) is shown with closed symbols. As for the observed data, the predicted MR is calculated by subtracting the SRTs from the independent $\text{HRTF}_{\text{full}}$ and independent IMBM from those of the respective co-located masker configurations.

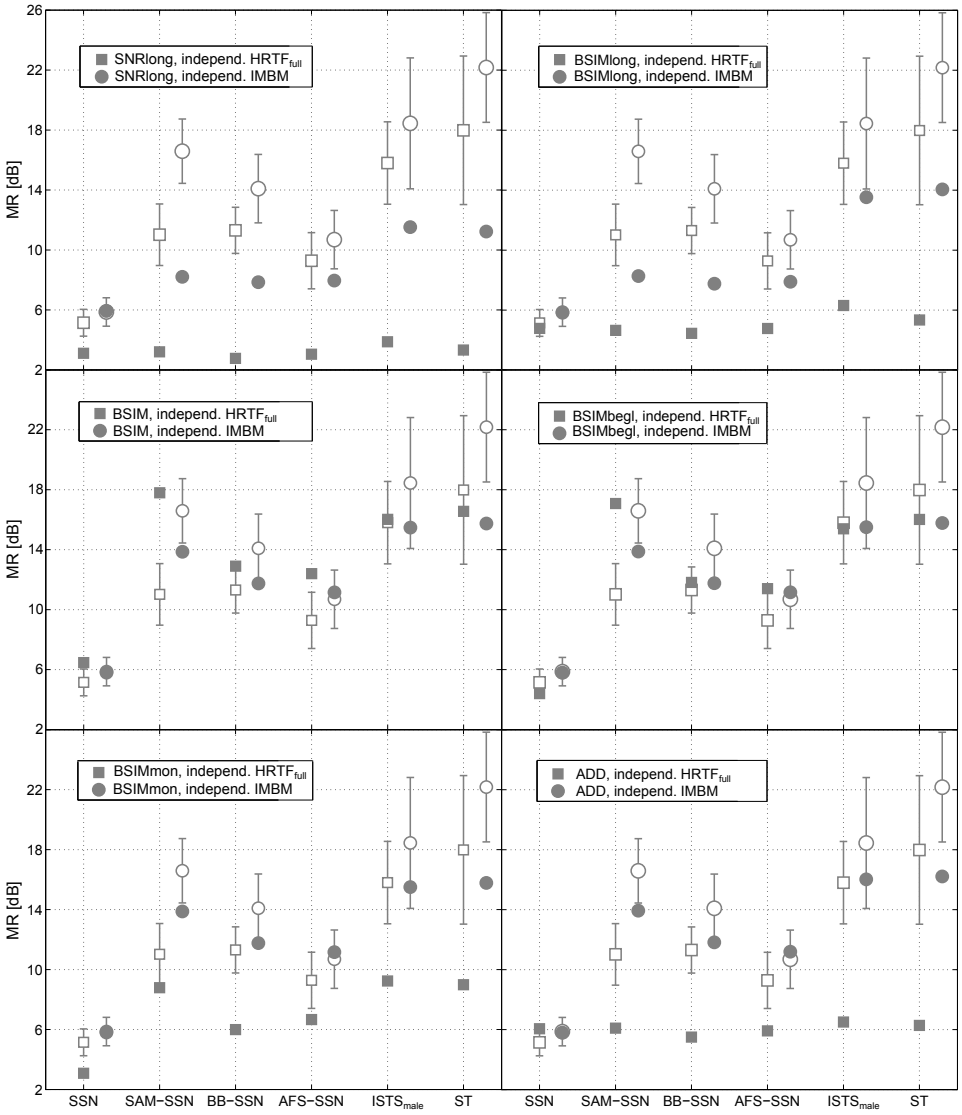


Figure C.9: Predicted masking release (MR) for the predictions of the individual binaural speech prediction model versions. The MR is shown in the same style as the predictions for the SRM in Fig. C.8. Open symbols denote the observed MR from the listening experiment, while closed symbols denote the predicted MR. The different symbols correspond to the different independent maskers (independent $\text{HRTF}_{\text{full}}$ and independent IMBM). Model data are shown without error bars.

Again, the two long-term analyses show a constant underestimation of MR for all masker types in the two independent masker conditions. Predicted MR is similar for SNR_{long} and $\text{BSIM}_{\text{long}}$ for the independent $\text{HRTF}_{\text{full}}$ condition, but the match between $\text{BSIM}_{\text{long}}$ and the observed data is slightly better. The MR pattern for the independent IMBM is met by both long-term analyses, but SRTs are underestimated by $5 - 10$ dB. There is in general a larger MR for the independent IMBM than for the independent $\text{HRTF}_{\text{full}}$ masker. The predicted MR for the two speech-like maskers is higher than for the modulated SSN-based masker, but still generally underestimated.
The MR predicted by BSIM is shown in the middle left panel of Fig. C.9. Predictions show an overall larger MR, but the pattern is exactly opposite to the observed MR. This suggests that, once glimpsing is anticipated, human listeners can utilize the glimpses better than BSIM. In contrast, when glimpses have to be extracted from the masker, the glimpsing process implemented in BSIM is more efficient. As for the SRTs, predicted MR is naturally the same for BSIM and $\text{BSIM}_{\text{begl}}$ in case of the IMBM, since the level information that is used in both model versions is nearly the same. For the independent $\text{HRTF}_{\text{full}}$ condition, the level information is not the same, but predicted MR by BSIM and $\text{BSIM}_{\text{begl}}$ is still very similar for these two model versions.
MR predictions by BSIM_{mon} are shown in the lower left panel of Fig. C.9. It is apparent that, again, prediction are "between" those of $\text{BSIM}_{\text{long}}$ and BSIM for the case of the independent $\text{HRTF}_{\text{full}}$ condition. In contrast, BSIM_{mon} predictions for the independent IMBM are more similar to those of BSIM. Thus, observed MR in the independent IMBM seems to be accounted for even by a short-term monaural model approach.
The lower right panel shows the predicted MR for the ADD approach. As for the SRM, ADD predictions are similar to those of other short-term BSIM versions for the case of the independent IMBM. Predictions for the case of the independent $\text{HRTF}_{\text{full}}$ are surprisingly poor, the predicted MR is about 6 dB, regardless of masker type or condition. This is surprising, because the enhancement of correlated target speech parts should be especially large in a masker condition, where the masker sequences in each are independent.

C.3.1 RMSE for the model SRM and MR

The RMSEs that result from comparing predicted SRM and MR with those observed in the listening experiment are shown in Tab. C.3. As for the predicted SRTs, the smallest and largest RMSEs for each HRTF condition are depicted in bold in the columns of the table. As mentioned earlier, in calculating the SRM and MR, the systematic errors in the predictions are eliminated, thus RMSEs can actually be used to estimate the performance of each model version.
As expected, but in contrast to the predicted SRTs, the largest RMSEs are found for SNR_{long} (followed by BSIM_{mon}). This model shows large RMSEs for all different HRTF conditions. Large RMSEs are also found for $\text{BSIM}_{\text{long}}$ and ADD, showing that the ADD approach is not a superior predictor for SRM or MR, except for the case of the IPD maskers (IPD_{mag0} and $\text{IPD}_{\text{mag60}}$). The smallest overall RMSEs are found for BSIM, $\text{BSIM}_{\text{begl}}$, and ADD for the

case of the IMBM. This is, because the outcome of the analysis of the IMBM is almost identical in all these models and therefore predicted SRTs are also nearly identical. Generally, RMSEs are larger for the case of independent masker sequences in both ears. This was also seen in the RMSEs for the SRT predictions and suggests that the models cannot well cope with this situation. As this certain masker condition does not contain IPDs, all model versions that rely on the EC-stage are expected to fail in providing accurate SRT (MR) predictions. Models relying on level information generally yield lower RMSEs in that situation, but taken together human listeners can better utilize the level information that arise in maskers with independent masker sequences in both ears.

model version	HRTF condition					independent maskers	
	$\mathrm{HRTF}_{\mathrm{full}}$	$\mathrm{ILD}_{\mathrm{only}}$	$\mathrm{IPD}_{\mathrm{mag0}}$	$\mathrm{IPD}_{\mathrm{mag60}}$	IMBM	$\mathrm{HRTF}_{\mathrm{full}}$	IMBM
$\mathrm{SNR}_{\mathrm{long}}$	**6.29**	**4.17**	**4.31**	**5.52**	**1.69**	**9.43**	**6.88**
$\mathrm{BSIM}_{\mathrm{long}}$	4.16	2.89	2.43	**2.32**	1.33	7.72	5.88
BSIM	3.10	2.40	**2.14**	2.47	**0.53**	3.21	3.25
$\mathrm{BSIM}_{\mathrm{begl}}$	**2.81**	**1.93**	3.86	3.19	**0.53**	**2.77**	3.23
$\mathrm{BSIM}_{\mathrm{mon}}$	4.99	2.38	4.23	3.49	**0.53**	5.30	3.23
ADD	3.84	2.89	2.48	2.95	**0.53**	6.99	**3.00**

Table C.3: Root-mean-square errors (RMSEs) in dB for the predictions of spatial release from masking and masking release due to independent masker sequences at the two ears. Model predictions for the different HRTF conditions are shown in columns for each model version. Smallest overall RMSEs, denoted in bold, are found for the predictions of SRM in the case of the IMBM. The predictions for this masker are identical for BSIM, $\mathrm{BSIM}_{\mathrm{begl}}$, $\mathrm{BSIM}_{\mathrm{mon}}$, and ADD. Largest RMSEs, also denote in bold, are generally found for the long-term analyses of the stimuli and in the case of independent masker sequences in both ears.

Appendix D

Supporting material for Chapter 4

D.1 The usage of STOI

In chapter 4 human speech recognition was predicted by using, among others, the short-time objective intelligibility model STOI (Taal et al., 2010). Fig. 4.3 shows STOI predictions that were matched to the SRT50, observed in the SSN masker. The input to the model, as it was applied in chapter 4, were the clean and degraded speech signals. The clean speech signal that was used in chapter 4 was the OLnoise, which is a stationary noise, spectrally matched to the OLSA sentences. This is different from the procedure in Taal et al. (2010) where the target signal was composed of 30 concatenated sentences of the respective target material.
The three panels of Fig. D.1 show STOI predictions for the three combinations of target and masker spectrum and the different input signals. Open squares denote SRTs that are identical to those shown in Fig. 4.5 (gained with OLnoise as target) and open circles SRTs that were gained when ten OLSA sentences were concatenated. Outcomes are matched to correctly predict the SRT in the SSN masker for both procedures. It is apparent that the type of input signal does not affect model outcomes in any way. Therefore, predicting SRTs with the OLnoise as it was done in chapter 4 is reasonable.
It was noted in section 4.6.4 that RMSEs decrease dramatically for the mr-sEPSM when model outcomes are matched to a reference SRT other than the SRT50 from the SSN masker. This process is omitted for STOI, because choosing another reference does not improve the model predictions. The stars in Fig. D.1 denote predictions, where the reference was the SRT50, reported in Wagener et al. (1999). It is apparent that this change of reference SRT does not improve the predictions, but only generally shifts the predicted SRTs upwards. From that, it can be hypothesized that choosing yet another reference frame, e.g. the SRT50 from the AFS-SSN masker, would generally lower predicted SRTs, but would not influence the overall prediction pattern. Thus, predictions of STOI with other reference SRTs were omitted in chapter 4.

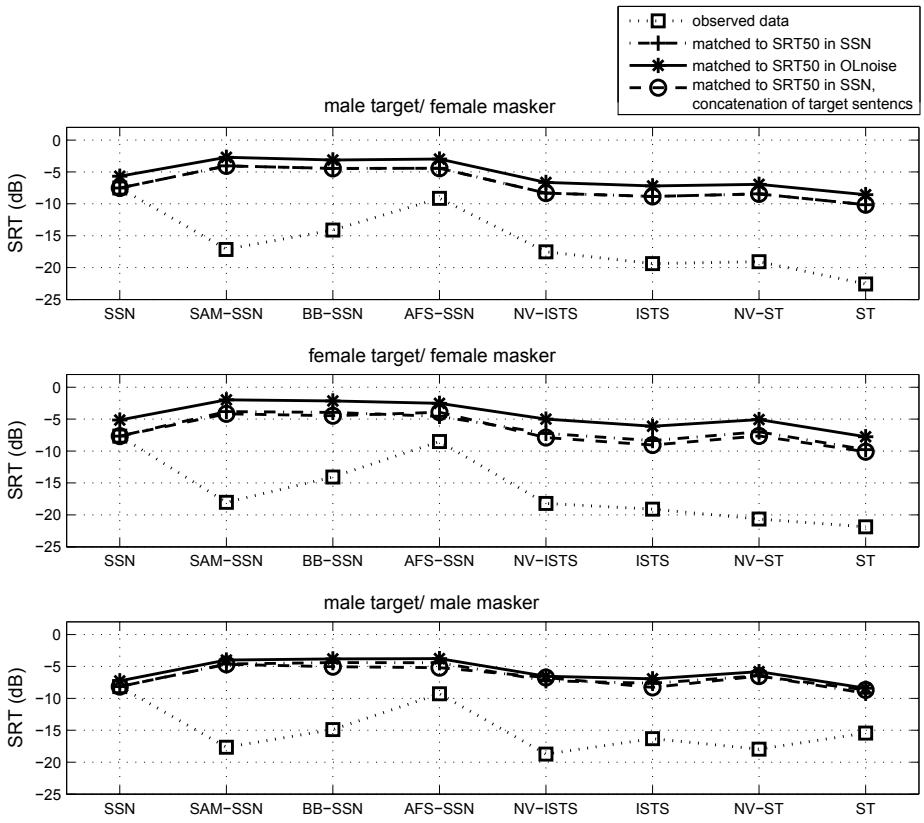


Figure D.1: Outcomes of the short-time objective intelligibility measure (STOI) for three combinations of target and masker spectrum. Open squares denote the observed SRTs, plus signs predictions when the input signal is the OLnoise, and open circles the predictions when the input signal are ten concatenated sentences. Stars denote predictions that arise when the outcome is matched as to correctly predict the SRT50 in the OLnoise as reported in Wagener et al. (1999).

Appendix E

Using a matrix sentence test

E.1 Assessing speech recognition with a matrix sentence test

Measurements in chapter 3 and 4 were done with speech material developed at the University of Oldenburg, the Oldenburger Satztest (OLSA, Wagener et al., 1999).
There are two categories of sentence tests that are typically used for assessing speech recognition and the distinction is made with respect to the sentence material. The first group consists of sentence tests that are made up of predictable, everyday sentences like the German Göttinger Satztest (Kollmeier and Wesselkamp, 1997), the American HINT test (Nilsson et al., 1994) or the Dutch Plomp and Mimpen (Plomp and Mimpen, 1979) sentences. The other group consists of matrix sentence tests, among others the Swedish Hagerman test (Hagerman, 1982) and the Dantale II (Wagener et al., 2003), to which the OLSA belongs to. The OLSA matrix test is made up of ten basic sentences, presented in the rows of Tab. E.1. A multitude of sentences can be generated from this base list by random combination of the words of the different categories. These sentences seem rather unusual at first glance, because they constitute semantically unpredictable (nonsense) sentences, but this avoids learning effects and thus provides reproducible SRTs also over long and repeated measurements (Wagener, 2004). Speech recognition in chapters 3 and 4 was determined with word scoring, meaning that the SNR during the measurement was adjusted in correspondence to the amount of words that were repeated correctly from the preceding sentences (Brand and Kollmeier, 2002).
During the measurements for both studies certain words were given as answers that sounded similar to the words from the matrix E.1, but were incorrect. Tab. E.2 shows an "alternative" matrix for OLSA, derived from false answers of the listeners.

Name	Verb	Numeral	Adjective	Object
Peter	bekommt	drei	große	Blumen
Kerstin	sieht	neun	kleine	Tassen
Tanja	kauft	sieben	alte	Autos
Ulrich	gibt	acht	nasse	Bilder
Britta	schenkt	vier	schwere	Dosen
Wolfgang	verleiht	fünf	grüne	Sessel
Stefan	hat	zwei	teure	Messer
Thomas	gewann	achtzehn	schöne	Schuhe
Doris	nahm	zwölf	rote	Steine
Nina	malt	elf	weiße	Ringe

Table E.1: The base list of test sentences from the Oldenburger Satztest (OLSA, Wagener et al., 1999). A variety of test sentences can be derived from this matrix by randomizing the words of the respective columns. An example sentence would be "Peter kauft achtzehn nasse Tassen." All sentences have a fixed grammatical structure (name-verb-numeral-adjective-object), but are not semantically predictable.

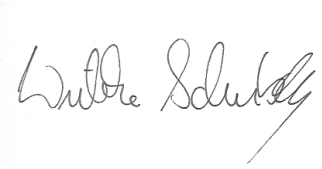
Name	Verb	Numeral	Adjective	Object
Richard	liebt	zehn	weite	Blusen
Boris	verkauft	ein	goldene	Rosen
Mia	vergibt	acht	große	Hosen
Kevin	verleiht			Taschen
	fängt			

Table E.2: An "alternative" OLSA matrix with listener's responses. These words were false, but were repeated more than once by the listeners.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe. Die Dissertation hat weder in Teilen, noch in ihrer Gesamtheit einer anderen wissenschaftlichen Hochschule zur Begutachtung in einem Promotionsverfahren vorgelegen.

Oldenburg, den 30.10.2015



Danksagung

An dieser Stelle ist nun Platz um Danke zu sagen, Danke für Großes und Kleines, für Allgemeines und Spezielles.
Zunächst möchte ich mich bei Prof. Dr. Dr. Birger Kollmeier für die Möglichkeit bedanken, meine Dissertation in der Arbeitsgruppe Medizinische Physik anfertigen zu können und dies auf einer Stelle, die angemessen bezahlt wurde. Außerdem möchte ich mich für die wertvolle Diskussion zum „roten Faden“ dieser Arbeit bedanken. Prof. Dr. Volker Hohmann danke ich ganz herzlich für die schnelle Zusage, die Zweitkorrektur dieser Arbeit zu übernehmen.
Ein spezielles *Danke* möchte ich meinen beiden Betreuern Dr. Thomas Brand und Dr. Stephan Ewert sagen. Die Ergebnisse unserer zahlreichen Gespräche finden sich auf dem einen oder anderen Weg in dieser Arbeit wieder. Unsere Zusammenarbeit war zwar nicht immer einfach, aber ich denke, letztendlich haben wir alle voneinander gelernt und von unseren unterschiedlichen – um nicht zu sagen kontroversen – Sichtweisen profitiert.
Des Weiteren möchte ich an dieser Stelle ausdrücklich Dr. Daniel Lübbert danken, den ich vor mittlerweile fünf Jahren als Mentor kennengelernt habe und dessen Satz: „Wenn sie erstmal gearbeitet haben, dann machen Sie eh keine Promotion mehr“ mir immer noch im Ohr ist, und der mich letztendlich dazu gebracht hat, überhaupt zu promovieren.
Ein ganz großes DANKE geht an die Hard- und Softwareabteilung der Medi-und-umzu, namentlich Frank Grunau, Anita Gorges, Felix Grossmann, Ingrid Wusowski, Katja Warnken, Ilona Dwehus und Annegret Bullermann-Wessels. Vielen Dank für die schnelle und unkomplizierte Hilfe bei Rechner-, Abrechnungs- und Formularfragen und die hervorragende Infrastruktur, die ihr bereitstellt!
Ein Dankeschön möchte ich auch dem Rest der Medi sagen und mich für die vielen schlauen und lustigen Gespräche bedanken, die das Arbeiten hier sehr angenehm gestaltet haben.
Ganz besonders möchte ich mich bei den Menschen bedanken, die mir in meiner Zeit in Oldenburg sehr an Herz gewachsen sind:
Zunächst sind das die Insassen von W2-0-071, Carolin Iben, Regina Baumgärtel, Marc René Schädler, Martin Klein-Hennig und Thomas Biberger, mit denen ich in wechselnder Besetzung die Höhen und Tiefen der Promotion durchlebt habe – herzlichen Dank für diese spannende Zeit!
Des Weiteren sind das die „Mädels“ vom OlWiN Mentoring-Programm, Esther Schoenmaker, Angela Josupeit, Geneviève Laumen, Christiane Stroth,

Cordula Walder, Dorothee Hodapp, Nicole Schwartz, Katharina Gandras, Sandra Tolnai, Inga Schepers, Oxana Ivanova, Heidi Wichmann und Verena Freytag, mit denen ich ein tolles Jahr verbracht habe und deren Bekanntschaft ich wirklich bereichernd finde.
Schließlich ist das noch die Aku-Mittagsrunde, bestehend aus Torben Wendt, Stefan Klockgether, Stephan Töpken, Esther Schoenmaker, Christina Imbery und Ewald Strasser, die sich meiner so unproblematisch und herzlich „angenommen“ hat – Danke euch für diese kurze, aber intensive Zeit!

Und zum Schluss möchte ich meinem Mann, Björn Opitz, danken, der in dieser ganzen wilden Zeit am Steuer geblieben ist, auch wenn wir zwischendurch ziemlich hohen Wellengang hatten. Es ist schön, dass wir wieder in ruhigeren Gewässern schippern, und ich freue mich auf unsere weitere Fahrt!

Curriculum Vitae

Wiebke Schubotz
Melanchthonstraße 33
22525 Hamburg

wiebke.schubotz@posteo.de

Born on 7^{th} of September 1984

in Rathenow

Nationality: German

Married to Dr. Björn Opitz

School & University Education

11/2011 – 01/2016 Research assistant in Project B1 in the SFB/TRR 31 "The active auditory system"

04/2011 – 10/2011 Stipend of the SFB/TRR 31 "The active auditory system"

04/2011 – 12/2015 Promotion at the Carl von Ossietzky University in Oldenburg

01/2010 – 01/2011 Diploma thesis at the University of Hamburg, Title: "The density distribution of ultra-high-energy protons in the local galaxy supercluster", Grade: *very good*

08/2007 – 05/2008 Studies of Physics and Astronomy at Helsingin Yliopisto, Helsinki, Finland

10/2004 – 01/2011 Studies of Physics at the University of Hamburg

08/2001 – 06/2002 Stay abroad at the Lakeview Highschool, Oregon, USA

09/1997 – 06/2004 School leaving examination (Abitur) at the Alexander-von-Humboldt Gymnasium, Premnitz, Grade: *1,2*

Active Involvement in

03/2014 the organisation of the "SFB-Doktorandenworkshop" at Langeoog

10/2010 the organisation of the Workshop "Cosmic Radiation Fields" at DESY Hamburg

03/2009 the organisation of the DPG-Frühjahrstagung (SAMOP) in Hamburg

Bibliography

Allen, J. B. (1994). How do humans process and recognize speech? *Speech and Audio Processing, IEEE Transactions on*, 2(4):567–577.

ANSI, A. (1997). S3. 5-1997, methods for the calculation of the speech intelligibility index. *New York: American National Standards Institute*.

ANSI, S. (1969). 5: Methods for the calculation of the articulation index. *American National Standards Institute, New York*.

Arbogast, T. L., Mason, C. R., and Kidd Jr, G. (2002). The effect of spatial separation on informational and energetic masking of speech. *The Journal of the Acoustical Society of America*, 112(5):2086–2098.

Arbogast, T. L., Mason, C. R., and Kidd Jr, G. (2005). The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 117(4):2169–2180.

Assmann, P. F. and Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, 88(2):680–697.

Barker, J. and Cooke, M. (2007). Modelling speaker intelligibility in noise. *Speech Communication*, 49(5):402–417.

Best, V., Mason, C. R., Kidd Jr, G., Iyer, N., and Brungart, D. S. (2015). Better-ear glimpsing in hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 137(2):EL213–EL219.

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

Bird, J. and Darwin, C. (1998). Effects of a difference in fundamental frequency in separating two sentences. *Psychophysical and physio logical advances in hearing*, pages 263–269.

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, 107(2):1065–1066.

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

Bregman, A. S., Abramson, J., Doehring, P., and Darwin, C. J. (1985). Spectral integration based on common amplitude modulation. *Perception & Psychophysics*, 37(5):483–493.

Bregman, A. S., Levitan, R., and Liao, C. (1990). Fusion of auditory components: Effects of the frequency of amplitude modulation. *Perception & psychophysics*, 47(1):68–73.

Bregman, A. S. and Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 32(1):19.

Broadbent, D. and Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *The Journal of the Acoustical Society of America*, 29(6):708–710.

Brokx, J. and Nooteboom, S. (1981). *Intonation and the perceptual separation of simultaneous voices*. Institute for Perception Research.

Bronkhorst, A. and Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 83(4):1508–1516.

Bronkhorst, A. and Plomp, R. (1992). Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *The Journal of the Acoustical Society of America*, 92(6):3132–3139.

Bronkhorst, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica*, 86(1):117–128.

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, pages 1–23.

Brown, G. J. and Wang, D. (2005). Separation of speech by computational auditory scene analysis. In *Speech enhancement*, pages 371–402. Springer.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, 109(3):1101–1109.

Brungart, D. S. and Iyer, N. (2012). Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *The Journal of the Acoustical Society of America*, 132(4):2545–2556.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, 110(5):2527–2538.

Carlyon, R. P. (1996). Encoding the fundamental frequency of a complex tone in the presence of a spectrally overlapping masker. *The Journal of the Acoustical Society of America*, 99(1):517–524.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979.

Christiansen, C. and Dau, T. (2012). Relationship between masking release in fluctuating maskers and speech reception thresholds in stationary noise. *The Journal of the Acoustical Society of America*, 132(3):1655–1666.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, 119(3):1562–1573.

Coughlin, M., Kewley-Port, D., and Humes, L. E. (1998). The relation between identification and discrimination of vowels in young and elderly listeners. *The Journal of the Acoustical Society of America*, 104(6):3597–3607.

Culling, J. F. and Colburn, H. S. (2000). Binaural sluggishness in the perception of tone sequences and speech in noise. *The Journal of the Acoustical Society of America*, 107(1):517–527.

Culling, J. F. and Darwin, C. (1993). Perceptual separation of simultaneous vowels: Within and across-formant grouping by f0. *The Journal of the Acoustical Society of America*, 93(6):3454–3467.

Culling, J. F. and Summerfield, Q. (1995). Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *The Journal of the Acoustical Society of America*, 98(2):785–797.

Darwin, C. J. (1997). Auditory grouping. *Trends in cognitive sciences*, 1(9):327–333.

Dau, T., Piechowiak, T., and Ewert, S. D. (2013). Modeling within-and across-channel processes in comodulation masking release. *The Journal of the Acoustical Society of America*, 133(1):350–364.

Deeks, J. M. and Carlyon, R. P. (2004). Simulations of cochlear implant hearing using filtered harmonic complexes: Implications for concurrent sound segregation. *The Journal of the Acoustical Society of America*, 115(4):1736–1746.

Delgutte, B. (1990). Physiological mechanisms of psychophysical masking: Observations from auditory-nerve fibers. *The Journal of the Acoustical Society of America*, 87(2):791–809.

Dreschler, W. A., Verschuure, H., Ludvigsen, C., and Westermann, S. (2001). Icra noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment: Ruidos icra: Señates de ruido artificial con espectro similar al habla y propiedades temporales para pruebas de instrumentos auditivos. *International Journal of Audiology*, 40(3):148–157.

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *Journal of the Acoustical Society of America*, 97(1):585–592.

Drullman, R., Festen, J. M., and Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, 95(5):2670–2680.

Dubbelboer, F. and Houtgast, T. (2008). The concept of signal-to-noise ratio in the modulation domain and speech intelligibility. *The Journal of the Acoustical Society of America*, 124(6):3937–3946.

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

Durlach, N. I., Mason, C. R., Kidd Jr, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (2003a). Note on informational masking (l). *The Journal of the Acoustical Society of America*, 113(6):2984–2987.

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd Jr, G. (2003b). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *The Journal of the Acoustical Society of America*, 114(1):368–379.

Edmonds, B. A. and Culling, J. F. (2006). The spatial unmasking of speech: Evidence for better-ear listening. *The journal of the Acoustical Society of America*, 120(3):1539–1545.

Elie, J. E., Mariette, M. M., Soula, H. A., Griffith, S. C., Mathevon, N., and Vignal, C. (2010). Vocal communication at the nest between mates in wild zebra finches: a private vocal duet? *Animal Behaviour*, 80(4):597–605.

Ewert, S. (2013). Afc—a modular framework for running psychoacoustic experiments and computational perception models. In *Proc Conf Acoust AIA-DAGA*, pages 1326–1329.

Ewert, S. D. and Dau, T. (2000). Characterizing frequency selectivity for envelope fluctuations. *The Journal of the Acoustical Society of America*, 108(3):1181–1196.

Festen, J. M. (1987). Explorations on the difference in srt between a stationary noise masker and an interfering speaker. *The Journal of the Acoustical Society of America*, 82(S1):S4–S4.

Festen, J. M. (1993). Contributions of comodulation masking release and temporal resolution to the speech-reception threshold masked by an interfering voice. *The Journal of the Acoustical Society of America*, 94(3):1295–1300.

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

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

French, N. and Steinberg, J. (1947). Factors governing the intelligibility of speech sounds. *The journal of the Acoustical society of America*, 19(1):90–119.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (2008). Spatial release from masking with noise-vocoded speech. *The Journal of the Acoustical Society of America*, 124(3):1627–1637.

Ghitza, O. (2001). On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *The Journal of the Acoustical Society of America*, 110(3):1628–1640.

Glasberg, B. R. and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1):103–138.

Glyde, H., Buchholz, J., Dillon, H., Best, V., Hickson, L., and Cameron, S. (2013). The effect of better-ear glimpsing on spatial release from masking. *The Journal of the Acoustical Society of America*, 134(4):2937–2945.

Gordon, P. C. (1997). Coherence masking protection in speech sounds: The role of formant synchrony. *Perception & psychophysics*, 59(2):232–242.

Gordon, P. C. (2000). Masking protection in the perception of auditory objects. *Speech Communication*, 30(4):197–206.

Griffiths, T. D. and Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5(11):887–892.

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scandinavian audiology*, 11(2):79–87.

Hall, J. W., Haggard, M. P., and Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. *The Journal of the Acoustical Society of America*, 76(1):50–56.

Hall III, J. W. and Grose, J. H. (1988). Comodulation masking release: Evidence for multiple cues. *The Journal of the Acoustical Society of America*, 84(5):1669–1675.

Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (1999). Speech intelligibility and localization in a multi-source environment. *The Journal of the Acoustical Society of America*, 105(6):3436–3448.

Hawley, M. L., Litovsky, R. Y., and Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 115(2):833–843.

Hochmuth, S., Jürgens, T., Brand, T., and Kollmeier, B. (2014). Multilingualer cocktailparty-einfluss von sprecher- und sprachspezifischen faktoren auf die sprachverständlichkeit im störschall. In *17. Jahrestagung der Deutschen Gesellschaft für Audiologie*. DGA.

Hoen, M., Meunier, F., Grataloup, C.-L., Pellegrino, F., Grimault, N., Perrin, F., Perrot, X., and Collet, L. (2007). Phonetic and lexical interferences in informational masking during speech-in-speech comprehension. *Speech communication*, 49(12):905–916.

Hohmann, V. (2002). Frequency analysis and synthesis using a gammatone filterbank. *Acta Acustica united with Acustica*, 88(3):433–442.

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (2010). Development and analysis of an international speech test signal (ists). *International journal of audiology*, 49(12):891–903.

House, A. S., Williams, C. E., Hecker, M. H., and Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *The Journal of the Acoustical Society of America*, 37(1):158–166.

Houtgast, T. (1989). Frequency selectivity in amplitude-modulation detection. *The Journal of the Acoustical Society of America*, 85(4):1676–1680.

Houtsma, A. J. and Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. *The Journal of the Acoustical Society of America*, 87(1):304–310.

Johnson, D. H. (1980). The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *The Journal of the Acoustical Society of America*, 68(4):1115–1122.

Jones, G. L. and Litovsky, R. Y. (2011). A cocktail party model of spatial release from masking by both noise and speech interferersa). *The Journal of the Acoustical Society of America*, 130(3):1463–1474.

Jørgensen, S. and Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *The Journal of the Acoustical Society of America*, 130(3):1475–1487.

Jørgensen, S. and Dau, T. (2014). *Modeling speech intelligibility based on the signal-to-noise envelope power ratio*. PhD thesis, Technical University of DenmarkDanmarks Tekniske Universitet, Department of Electrical EngineeringInstitut for Elektroteknologi, Hearing SystemsHearing Systems.

Jørgensen, S., Ewert, S. D., and Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *The Journal of the Acoustical Society of America*, 134(1):436–446.

Josupeit, A., Hohmann, V., and van de Par, S. (2012). Release from masking of low-frequency complex tones by high-frequency complex tone cue bands. *The Journal of the Acoustical Society of America*, 132(6):EL450–EL455.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (2008). Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 3933–3936. IEEE.

Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., and Kollmeier, B. (2009). Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP Journal on Advances in Signal Processing*, 2009:6.

Kidd Jr, G., Best, V., and Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America*, 124(6):3793–3802.

Kidd Jr, G., Mason, C. R., Rohtla, T. L., and Deliwala, P. S. (1998). Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. *The Journal of the Acoustical Society of America*, 104(1):422–431.

Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. (2009). Role of mask pattern in intelligibility of ideal binary-masked noisy speech. *The Journal of the Acoustical Society of America*, 126(3):1415–1426.

Kollmeier, B. (1990). *Meßmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache*. Habilitationsschrift, Georg-August-Universität, Göttingen.

Kollmeier, B. (1999). On the four factors involved in sensorineural hearing loss. *Psychophysics, Physiology and Models of Hearing, edited by T. Dau, V. Hohmann, & B. Kollmeier. Singapore: World Scientific Publishing.*

Kollmeier, B., Rennies, J., and Brand, T. (2011). Tools to predict binaural speech intelligibility in complex listening environments for normal and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 129(4):2669–2669.

Kollmeier, B., Schädler, M. R., Warzybok, A., Meyer, B., and Brand, T. (2015). Individual speech recognition in noise, the audiogram, and more: Using automatic speech recognition (asr) as a modelling tool and consistency check across audiological measure. In *International Symposium on Auditory and Audiological Research*.

Kollmeier, B. and Wesselkamp, M. (1997). Development and evaluation of a german sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America*, 102(4):2412–2421.

Kong, Y.-Y. and Carlyon, R. P. (2007). Improved speech recognition in noise in simulated binaurally combined acoustic and electric stimulationa). *The Journal of the Acoustical Society of America*, 121(6):3717–3727.

Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, 34(11):1689–1697.

Levitt, H. and Rabiner, L. (1967). Binaural release from masking for speech and gain in intelligibility. *The journal of the acoustical society of america*, 42(3):601–608.

Licklider, J. (1948). The influence of interaural phase relations upon the masking of speech by white noise. *The Journal of the Acoustical Society of America*, 20(2):150–159.

Lingner, A., Wiegrebe, L., Grothe, B., and Ewert, S. D. (2015). Binaural glimpses at the cocktail party. *The Journal the Association of Research in Otolaryngology*, currently under revision.

Litovsky, R. Y. (2012). Spatial release from masking. *Acoustics today*, 8(2):18–25.

Loizou, P. C. (2013). *Speech enhancement: theory and practice*. CRC press.

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

Lutfi, R. A. (1990). How much masking is informational masking? *The Journal of the Acoustical Society of America*, 88(6):2607–2610.

Lutfi, R. A., Gilbertson, L., Heo, I., Chang, A.-C., and Stamas, J. (2013). The information-divergence hypothesis of informational masking. *The Journal of the Acoustical Society of America*, 134(3):2160–2170.

Markel, J. E. and Gray, A. H. (1982). *Linear prediction of speech*. Springer-Verlag New York, Inc.

Marrone, N., Mason, C. R., and Kidd Jr, G. (2008). Tuning in the spatial dimension: Evidence from a masked speech identification task. *The Journal of the Acoustical Society of America*, 124(2):1146–1158.

Mattys, S. L. and Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of memory and Language*, 65(2):145–160.

Meddis, R. and O'Mard, L. (1997). A unitary model of pitch perception. *The Journal of the Acoustical Society of America*, 102(3):1811–1820.

Meyer, R. M. and Brand, T. (2013). Comparison of different short-term speech intelligibility index procedures in fluctuating noise for listeners with normal and impaired hearing. *Acta Acustica united with Acustica*, 99(3):442–456.

Micheyl, C., Arthaud, P., Reinhart, C., and Collet, L. (2000). Informational masking in normal-hearing and hearing-impaired listeners. *Acta oto-laryngologica*, 120(2):242–246.

Moore, B. C. (1990). Co-modulation masking release: spectro-temporal pattern analysis in hearing. *British journal of audiology*, 24(2):131–137.

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill.

Moore, B. C. and Ernst, S. M. (2012). Frequency difference limens at high frequencies: Evidence for a transition from a temporal to a place code. *The Journal of the Acoustical Society of America*, 132(3):1542–1547.

Moore, B. C. and Sęk, A. (2009). Sensitivity of the human auditory system to temporal fine structure at high frequencies. *The Journal of the Acoustical Society of America*, 125(5):3186–3193.

Moore, B. C. and Vickers, D. A. (1997). The role of spread excitation and suppression in simultaneous masking. *The Journal of the Acoustical Society of America*, 102(4):2284–2290.

Moore, G. A. and Moore, B. C. (2003). Perception of the low pitch of frequency-shifted complexes. *The Journal of the Acoustical Society of America*, 113(2):977–985.

Mosteller, F. and Youtz, C. (2006). Tables of the freeman-tukey transformations for the binomial and poisson distributions. In *Selected Papers of Frederick Mosteller*, pages 337–347. Springer.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2):1085–1099.

Oxenham, A. J., Micheyl, C., and Keebler, M. V. (2009). Can temporal fine structure represent the fundamental frequency of unresolved harmonics? *The Journal of the Acoustical Society of America*, 125(4):2189–2199.

Oxenham, A. J., Micheyl, C., Keebler, M. V., Loper, A., and Santurette, S. (2011). Pitch perception beyond the traditional existence region of pitch. *Proceedings of the National Academy of Sciences*, 108(18):7629–7634.

Palmer, A. and Russell, I. (1986). Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells. *Hearing research*, 24(1):1–15.

Patterson, R. D., Allerhand, M. H., and Giguere, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America*, 98(4):1890–1894.

Pätzold, M. and Simpson, A. P. (1997). Acoustic analysis of german vowels in the kiel corpus of read speech. *Arbeitsberichte des Instituts für Phonetik und digitale Spachverarbeitung Universität Kiel*, 32:215–247.

Piechowiak, T., Ewert, S. D., and Dau, T. (2007). Modeling comodulation masking release using an equalization-cancellation mechanism. *The Journal of the Acoustical Society of America*, 121(4):2111–2126.

Plack, C. J., Oxenham, A. J., and Fay, R. R. (2006). *Pitch: neural coding and perception*, volume 24. Springer Science & Business Media.

Plomp, R. and Mimpen, A. (1979). Improving the reliability of testing the speech reception threshold for sentences. *International Journal of Audiology*, 18(1):43–52.

Plomp, R. and Mimpen, A. (1981). Effect of the orientation of the speaker's head and the azimuth of a noise source on the speech-reception threshold for sentences. *Acta Acustica united with Acustica*, 48(5):325–328.

Pollack, I. (1975). Auditory informational masking. *The Journal of the Acoustical Society of America*, 57(S1):S5–S5.

Pompino-Marschall, B. (2009). *Einführung in die Phonetik*. Walter de Gruyter.

Qin, M. K. and Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *The Journal of the Acoustical Society of America*, 114(1):446–454.

Qin, M. K. and Oxenham, A. J. (2006). Effects of introducing unprocessed low-frequency information on the reception of envelope-vocoder processed speech. *Journal of the Acoustical Society of America*, 119(4):2417–2426.

Rhebergen, K. S., Lyzenga, J., Dreschler, W. A., and Festen, J. M. (2010). Modeling speech intelligibility in quiet and noise in listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America*, 127(3):1570–1583.

Rhebergen, K. S. and Versfeld, N. J. (2005). A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. *The Journal of the Acoustical Society of America*, 117(4):2181–2192.

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

Rosen, S., Souza, P., Ekelund, C., and Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America*, 133(4):2431–2443.

Schädler, M. R., Warzybok, A., Ewert, S. D., and Kollmeier, B. (2015a). A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception. *The Journal of the Acoustical Society of America*, submitted.

Schädler, M. R., Warzybok, A., Hochmuth, S., and Kollmeier, B. (2015b). Matrix sentence intelligibility prediction using an automatic speech recognition system. *International Journal of Audiology*, in press.

Scheidiger, C. and Dau, T. (2015). Modellierung der sprachverständlichkeit in schwerhörenden probanden. In *Fortschritte der Akustik, 41. Jahrestagung für Akustik*, pages 127–130.

Schoenmaker, E. and van de Par, S. (2013). Auditory streaming in cocktail parties: Better-ear versus binaural processing. In *AIA-DAGA 2013 Conference on Acoustics*, pages 1301–1303. AIA-DAGA.

Schubotz, W., Brand, T., Kollmeier, B., and Ewert, S. D. (2015). Monaural speech intelligibility and detection in maskers with varying amount of spectro-temporal speech features. *The Journal of the Acoustical Society of America*, currently under revision.

Scott, S. K., Rosen, S., Wickham, L., and Wise, R. J. (2004). A positron emission tomography study of the neural basis of informational and energetic masking effects in speech perception. *Journal of the Acoustical Society of America*, 115(2):813–821.

Seyfarth, R. M. and Cheney, D. L. (2003). Signalers and receivers in animal communication. *Annual review of psychology*, 54(1):145–173.

Steeneken, H. J. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, 67(1):318–326.

Stone, M. A., Füllgrabe, C., and Moore, B. C. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *The Journal of the Acoustical Society of America*, 132(1):317–326.

Studebaker, G. A. (1985). A rationalized arcsine transform. *Journal of Speech, Language, and Hearing Research*, 28(3):455–462.

Suzuki, T. N. (2014). Communication about predator type by a bird using discrete, graded and combinatorial variation in alarm calls. *Animal Behaviour*, 87:59–65.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4214–4217. IEEE.

Thornton, A. R. and Raffin, M. J. (1978). Speech-discrimination scores modeled as a binomial variable. *Journal of Speech, Language, and Hearing Research*, 21(3):507–518.

Turner, C. W., Gantz, B. J., Vidal, C., Behrens, A., and Henry, B. A. (2004). Speech recognition in noise for cochlear implant listeners: benefits of residual acoustic hearing. *The Journal of the Acoustical Society of America*, 115(4):1729–1735.

van de Par, S. and Kohlrausch, A. (1998). Analytical expressions for the envelope correlation of narrow-band stimuli used in cmr and bmld research. *The Journal of the Acoustical Society of America*, 103(6):3605–3620.

Verhey, J. L., Pressnitzer, D., and Winter, I. M. (2003). The psychophysics and physiology of comodulation masking release. *Experimental Brain Research*, 153(4):405–417.

Vom Hövel, H. (1984). *Zur Bedeutung der Übertragungseigenschaften des Außenohrs sowie des binauralen Hörsystems bei gestörter Sprachübertragung*. Dissertation, RWTH Aachen.

Wagener, K. (2004). *Factors influencing sentence intelligibility in noise*. BIS Verlag.

Wagener, K., Brand, T., and Kühnel, V. (1999). Entwicklung und evaluation eines satztestes für die deutsche sprache i-iii: Design, optimierung und evaluation des oldenburger satztestes. *Zeitschrift für Audiologie*, 38(1-3).

Wagener, K., Hochmuth, S., Ahrlich, M., Zokoll, M., and Kollmeier, B. (2014). Der weibliche oldenburger satztest. In *17. Jahrestagung der Deutschen Gesellschaft für Audiologie*. DGA.

Wagener, K., Josvassen, J. L., and Ardenkjær, R. (2003). Design, optimization and evaluation of a danish sentence test in noise: Diseño, optimización y evaluación de la prueba danesa de frases en ruido. *International journal of audiology*, 42(1):10–17.

Wan, R., Durlach, N. I., and Colburn, H. S. (2014). Application of a short-time version of the equalization-cancellation model to speech intelligibility experiments with speech maskers. *The Journal of the Acoustical Society of America*, 136(2):768–776.

Warren, R. M., Riener, K. R., Bashford, J. A., and Brubaker, B. S. (1995). Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits. *Perception & Psychophysics*, 57(2):175–182.

Wesker, T., Meyer, B. T., Wagener, K., Anemüller, J., Mertins, A., and Kollmeier, B. (2005). Oldenburg logatome speech corpus (ollo) for speech recognition experiments with humans and machines. In *Interspeech*, pages 1273–1276.

Zekveld, A. A., Kramer, S. E., and Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and hearing*, 32(4):498–510.