

MACQUARIE UNIVERSITY

PH.D. THESIS

Understanding speech in complex
acoustic environments: the role of
informational masking and auditory
distance perception

Author:

Adam Westermann

Supervisors:

Dr. Jörg M. Buchholz

Dr. Harvey Dillon

Department of Linguistics, Faculty of Human Sciences

and

National Acoustic Laboratories, Australian Hearing

February 2015

Declaration of Authorship

I, Adam WESTERMANN, declare that this thesis titled, 'Understanding speech in complex acoustic environments: the role of informational masking and auditory distance perception' and the work presented in it are my own. I confirm that:

- This thesis has been submitted solely to Macquarie University for consideration for the doctoral degree.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Ethics approvals have been obtained from the Australian Hearing Human Research Ethics Committee. The signed approvals with number AHHREC2012-13 and AHHREC2013-1, can be found in Appendix A.

Signed:

Date:

Abstract

One of the greatest challenges for the auditory system is communicating in environments where speech is degraded by multiple spatially distributed maskers and room reverberation. This “cocktail-party” situation and the related auditory mechanisms have been a topic for numerous studies. This thesis primarily investigated speech intelligibility in such environments— specifically considering the role of differences in distance between talkers and the contribution of informational masking (IM).

The first two studies investigated the role of differences in distance between competing talkers on spatial release from masking (SRM) in normal hearing (NH) and subsequently, hearing impaired (HI) listeners. Intelligibility improved for both NH and HI listeners when moving the masker further away from the target. Contrastingly, when the target was moved further away and the maskers were kept near the listener, the results varied significantly across subjects. While intelligibility improved for some NH listeners, the HI listeners performed substantially worse. It was hypothesized that in this condition IM was caused by masker distraction rather than confusion. In the third study, the role of IM was investigated in a simulated cafeteria environment. Substantial IM effects were only observed when the target and masking talker were colocated and the same person. In conditions that resemble real life, no significant IM effects were found. This suggests that IM is of low relevance in real-life listening and is exaggerated by target-masker similarities and the colocated spatial configuration often used in previous listening tests.

The final study investigated the effect of nearby masking talkers in a simulated cafeteria environment with NH and HI listeners. The study demonstrated that for realistic conditions, nearby distracters introduce a significant amount of IM in both NH and HI subjects. However, the observed IM was likely not due to target-masker confusions, but rather caused by the nearby masker distracting the listener. Overall, this work suggests that (i) NH and HI listeners use distance related cues in the cocktail-party environment, (ii) in such environments IM related to target-masker confusions is of little relevance, and (iii) nearby maskers introduce IM - likely due to distraction of attention. These findings contribute to our understanding of auditory processing and could potentially have implications on signal processing methods for hearing devices.

Preface

This thesis work was conducted from September 2011 to September 2014. What an incredible three years it has been relocating from chilly Denmark to the land Down Under and living by the beach while having my daily life at the National Acoustic Laboratories (NAL).

First and foremost, I have to express my deepest gratitude to Jörg. From the first day more than five years ago, when I unknowingly stepped into his office at the Centre for Applied Hearing Research (CAHR) and went home with a long list of psychoacoustic effects to study, to the final leg of the Ph.D., where he has sacrificed nights and weekends, Jörg has been the best supervisor one could ask for.

Secondly, I would acknowledge my co-supervisor, Harvey. Many thanks for allowing me to spend these years at NAL and always showing enthusiasm for my work.

I am very grateful for the funding received from Widex A/S and Macquarie University. Without their support none of this would have been possible. Besides funding, we also need ears to study; therefore I am indebted to more than 75 participants who have volunteered their time for this work.

On the other side of the globe, I have to give a big thanks to Torsten Dau for hosting me on several occasions during the years. While in Denmark I would like to acknowledge the people at Widex I have had contact with: Kristoffer, Thomas, Morten and Søren.

I have had the pleasure of sharing some great (more or less work-related) moments with some great people at NAL. While the list is long, I will especially mention Toby, Chris, Fabrice and Bram. It is ironic that you move all the way to Australia and then learn about mateship from a dynamic group of Europeans.

Last but not least, I have to give my heartfelt thanks to my partner Abbie. Her cheery disposition, invaluable help and kind words of encouragement have been instrumental in completing this thesis.

List of publications

Journal articles, letters and book chapters:

Westermann, A. and Buchholz, J. M. (2012), “Binaural dereverberation based on interaural coherence histograms”, *J. Acoust. Soc. Am.* **133**, 2767–2777.

Tsilfidis, A., Westermann, A., Buchholz, J. M., Georganti, E. and Mourjopoulos, J. (2013), “Binaural Dereverberation”, in *The Technology of Binaural Listening* edited by Blauert, J, 359–396 (Springer Berlin Heidelberg).

Westermann, A. and Buchholz, J. M. (2014), “The effect of spatial separation in distance on the intelligibility of speech in rooms”, *J. Acoust. Soc. Am.*, **137**(2), 757–767.

Westermann, A. and Buchholz, J. M., “The effect of a hearing impairment on source-distance dependent speech intelligibility in rooms”, *J. Acoust. Soc. Am. Express Letter*, *in prep.*.

Westermann, A. and Buchholz, J. M. (2014), “The influence of informational masking in reverberant, multi-talker environments”, *J. Acoust. Soc. Am.*, *submitted*.

Westermann, A. and Buchholz, J. M., “The effect nearby maskers in reverberant, multi-talker environments”, *J. Acoust. Soc. Am.*, *in prep.*.

Conference articles:

Westermann, A. and Buchholz, J. M. (2013), “Release from masking through spatial separation in distance in hearing impaired listeners”, *Proceedings of Meetings on Acoustics (ICA/ASA)*, Montreal, Canada.

Westermann, A. and Buchholz, J. M. (2013), “The influence of informational masking in complex real-world environments”, *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY.

Published abstracts:

Westermann, A. and Buchholz, J. M. (2012), “Release from masking through spatial separation in distance”, *International Hearing Aid Research Conference (IHCON)*, Lake Tahoe, CA.

Westermann, A. and Buchholz, J. M. (2014), “The Influence of Nearby Maskers on Informational Masking in Complex Real-world Environments”, *Association for Research in Otolaryngology (ARO) MidWinter meeting*, San Diego, CA.

Westermann, A. and Buchholz, J. M. (**2014**), “Informational masking and listening in complex real-world environments”, World Congress of Audiology (WCA), Brisbane, Australia.

Westermann, A. and Buchholz, J. M. (**2014**), “Informational Masking in Complex Real-world Environments”, International Hearing Aid Research Conference (IHCON), Lake Tahoe, CA.

Contents

Declaration of Authorship	ii
Abstract	iii
Preface	iv
List of publications	v
Contents	vii
Abbreviations	xi
1 General introduction	1
1.1 Aims of this thesis	4
1.2 Overview of chapters	4
2 The effect of spatial separation in distance on the intelligibility of speech in rooms	7
2.1 Introduction	8
2.2 Methods	11
2.2.1 Subjects	11
2.2.2 Stimuli	11
2.2.3 Spatialization of sounds	13
2.2.4 Procedures	14
2.3 Results	16
2.3.1 Varying masker distance	16
2.3.2 Varying target distance	17
2.3.3 Diotic versus dichotic presentation	19
2.4 Discussion	20
2.4.1 Objective signal-based measures	20
2.4.2 Effect of equalization	24
2.4.3 Role of informational masking	26
2.5 Summary and conclusion	30
2.6 Acknowledgments	31
3 The effect of a hearing impairment on source-distance dependent speech intelligibility in rooms	33
3.1 Introduction	34
3.2 Methods	35
3.2.1 Stimuli	35

3.2.2	Procedures	37
3.2.3	Subjects	38
3.3	Results	38
3.4	Discussion and conclusions	40
3.5	Acknowledgments	42
4	The influence of informational masking in reverberant, multi-talker environments	43
4.1	Introduction	44
4.2	Methods	48
4.2.1	Subjects	48
4.2.2	Stimuli	48
4.2.3	Spatialization of sounds	51
4.2.4	Procedures	52
4.3	Results	53
4.3.1	Two-dialogue cafeteria	54
4.3.2	Seven-dialogue cafeteria	56
4.4	Discussion	57
4.4.1	Perspectives	62
4.5	Summary and conclusion	63
4.6	Acknowledgments	64
5	The effect of nearby maskers in reverberant, multi-talker environments	65
5.1	Introduction	66
5.2	Methods	68
5.2.1	Subjects	68
5.2.2	Stimuli	68
5.2.3	Equipment	70
5.2.4	Spatialization of sounds	71
5.2.5	Cognitive measures	73
5.2.6	Procedures	73
5.3	Results	74
5.3.1	Speech intelligibility measures	74
5.3.2	Cognitive measures	78
5.4	Discussion	78
5.4.1	Effect of informational masking	78
5.4.2	The region of informational masking	81
5.4.3	Cognition and informational masking	83
5.4.4	Perspectives	84
5.5	Summary and conclusion	85
5.6	Acknowledgments	86
6	General summary and discussion	87
6.1	Perspectives and limitations of this work	89

A Binaural dereverberation based on interaural coherence histograms	93
A.1 Introduction	94
A.2 The coherence-based dereverberation algorithm	96
A.2.1 Signal processing	96
A.2.2 Signal decomposition and coherence estimation	97
A.2.3 Coherence-to-gain mapping	98
A.2.4 Reference systems	101
A.3 Evaluation methods	102
A.3.1 Objective evaluation methods	103
A.3.2 Subjective evaluation methods	105
A.4 Results	106
A.4.1 Effects of reverberation on speech in different acoustic environments	106
A.4.2 Effects of dereverberation processing on speech	108
A.5 Discussion	112
A.6 Summary and conclusion	115
A.7 Appendix	115
A.7.1 Measuring binaural impulse responses	115
A.7.2 Acknowledgments	116
 B Ethics approvals	 117
 Bibliography	 121

Abbreviations

4FAHL	F our- F requency A verage H earing L oss
AMR	A daptive M ulti- R ate speech coder
ANOVA	A nalysis O f V ariance
BKB	B amford- K owal- B ench
BRIR	B inaural R oom I mpulse R esponse
CRM	C oordinate R esponse M easure
DRR	D irect-to- R everberant R atio
EM	E nergetic M asking
FIR	F inite I mpulse R esponse
GUI	G raphical U ser I nterface
HATS	H ead A nd T orso S imulator
HI	H earing I mpaired
HL	H earing L oss
IBM	I deal B inary M ask
IC	I nteraural C oherence
IG	I nsertion G ain
IR	I mpulse R esponse
ILD	I nteraural L evel D ifference
IM	I nformational M asking
ITD	I nteraural T ime D ifference
IPD	I nteraural P hase D ifference
JND	J ust- N oticeable D ifference
LiSN-S	L istening in S patialized N oise- S entence test
LoRA	L oudspeaker-based R oom A uralization
LTASS	L ong- T erm A verage S peech S pectrum
MUSHRA	M ultiple S timuli with H idden R eference and A ncor test
NAL	N ational A coustic L aboratories
NH	N ormal H earing

NMR	Noise-Mask R atio
RIR	Room Impulse R esponse
RMS	Root Mean S qaure
RST	Reading Span T est
segSRR	Segmental Signal-to- R everberation R atio
SNR	Signal-to-Noise R atio
SRM	Spatial R elease from M asking
SPL	Sound P ressure L evel
SRT	Speech R eception T hreshold
STFT	Short-Time F ourier T ransform
TMR	Target-to-Masker R atio

Chapter 1

General introduction

The ability to hear speech and thereby participate in conversations is a cornerstone in our daily lives. However, people with a hearing impairment (HI) are often faced with difficulties when communicating with people around them. The detrimental effect of a hearing impairment especially becomes evident in situations where normal hearing (NH) listeners already struggle. A noisy and reverberant environment with multiple spatially distributed sound-sources is a common example where the NH auditory system copes but HI listeners have difficulties. Cherry (1953) aptly defined such scenarios as the “cocktail party problem” and raised the question, “*how do we recognize what one person is saying when others are speaking at the same time?*”. Today, the cocktail party problem unifies research that involves multiple stages of auditory processing from low-level peripheral processing to binaural processing and localization all the way to auditory scene analysis, perceptual grouping and selective listening (Bronkhorst, 2000). While particular aspects are well understood, there is not yet a complete picture that accurately accounts for how the NH auditory system performs in the cocktail party and that can precisely explain why the HI auditory system fails.

One of the most commonly applied outcome measures when considering the ability to communicate is speech intelligibility. Speech reception thresholds (SRTs) are often applied to adaptively measure the signal-to-noise ratio (SNR) which results in 50 % intelligibility on an underlying psychometric function. SRTs are currently used in clinics in addition to the pure-tone audiogram to estimate a person’s speech understanding and their benefit from receiving a hearing device (Dillon, 2001). In clinics, the target is presented in a background of speech-shaped noise. SRTs and other speech intelligibility measures are widely employed by psychoacoustic researchers. They have shown important effects such as reduced benefit from temporal masker fluctuations with a hearing

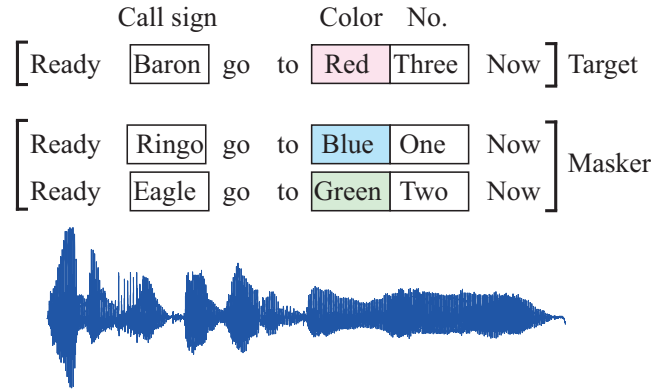


FIGURE 1.1: Example from the CRM speech corpus. The subject is instructed to report the color/number coordinate in the presence of similar maskers. A corpus such as the CRM will exhibit a strong IM effect.

impairment (Festen and Plomp, 1990) or the increased masking capabilities of a speech masker compared to a noise masker (Brungart *et al.*, 2001).

Such observations have resulted in the concept of masking being divided between energetic masking (EM) and informational masking (IM). EM describes cochlear masking effects occurring in the auditory periphery before the auditory nerve and central auditory system. Thereby, many aspects of EM can be accounted for by applying an auditory model which incorporates the cochlear behavior (Dau *et al.*, 1996; Oxenham and Moore, 1994). On the other hand, IM relates to masking of a central nature which makes it harder to concisely define and pinpoint. In regard to speech perception, IM often occurs because of confusions in discrimination between a target and masking talker (Kidd *et al.*, 2007). In psychoacoustic experiments, these confusions are often further exaggerated by the use of speech corpora that have inherent confusions from either target-masker similarity or structure. This for example, could be by using the same talker for the masker and target (Cameron and Dillon, 2007) or by time-aligning utterances and including masker uncertainty (as illustrated in Fig. 1.1; Bolia *et al.*, 2000). Studies have shown differences between target and masker, such as gender, language spoken and time-reversal, all substantially reduce IM (Brungart *et al.*, 2001). Concepts from auditory scene analysis such as grouping, streaming and stream selection are all used to explain aspects of IM (Bregman, 1994; Kidd *et al.*, 2007). Hence, any cues that aid stream segregation are expected to reduce IM, which includes talker differences, spectral separation, temporal de-synchronization and spatial separation.

The ability to use differences in location between sound sources to better understand

speech is an aspect of the cocktail-party effect that has received considerable attention. This spatial unmasking, or spatial release from masking (SRM), is typically measured as the difference in speech intelligibility when a target and masker are at the same place in space (colocated) and when they are angularly separated on the horizontal plane. In addition to the reduction in IM from the perceptual segregation of the sound sources, SRM is mainly accredited to SNR fluctuations across ears caused by head-shadow as well as binaural interaction (Brungart, 2012; Glyde *et al.*, 2013). However, where SRM resulting from angular separation has been the topic of numerous studies, very little research has looked at the effect of separation in distance on SRM. As noted by Darwin (2008) on SRM and distance-related cues, “*There has been almost no work on the effectiveness of these cues...*”.

In general, auditory distance perception has received less attention compared to angular localization. Mainly, it is dominated by vision and when we rely only on our auditory system to estimate the distance of a sound source it is rather imprecise. The main cue for auditory distance perception is the intensity of sound arriving at the listener. By predicting the initial level of a sound source from vocal effort or *a priori* knowledge about the source, the auditory system estimates its distance. When a sound source is placed in a room, the surrounding surfaces will give rise to reflections and reverberation. As reverberation is almost independent of distance, distance-dependent direct-to-reverberant energy ratio (DRR) can further aid the listener in establishing the distance of a sound source. Akeroyd *et al.* (2007) investigated DRR just-noticeable-differences (JNDs) in NH and HI listeners. While NH-listener JNDs corresponded to a doubling of distance, HI listeners were not able to reliably use DRR cues. Beyond this initial study of JNDs, the ability of the impaired auditory system to use distance-related cues is not well understood.

In a complex scene such as the “cocktail party” (Cherry, 1953), a target talker will be masked by reverberation and other sound sources distributed in directions and distances. Conversations will be dynamic, involve talkers with various vocal characteristics and cover a range of semantical meanings. The ability to communicate in such environments will be linked to a listener’s ability to use differences between talkers, spatial separation and glimpses of high SNR combined with visual and semantical cues. More recently, researchers have successfully employed intricate room acoustical models combined with three-dimensional loudspeaker-arrays to incorporate many of these features

into laboratory-based experiments (Best *et al.*, 2013a; Favrot and Buchholz, 2010; Seeber *et al.*, 2010). By using such tools, we can reevaluate and revise some of the assumptions about how the auditory system fares in Cherry’s cocktail party environment.

1.1 Aims of this thesis

The overall aims of this thesis are:

- To investigate speech intelligibility in NH and HI listeners when changing the distance between target and masking talkers placed inside a room.
- To study the influence of IM in a simulated multi-talker reverberant environment and thereby gain a better understanding of the involvement of IM in everyday life.
- To uncover and further analyze critical scenarios in which substantial IM effects occur, especially those potentially related to auditory distance perception.

1.2 Overview of chapters

This thesis presents five interconnected studies. **Chapter 2** describes an experiment investigating the effect of spatial separation in distance on speech intelligibility in NH listeners. Here, the perceived distances of target and masking sound sources are varied to create three conditions: colocated, the target further away and the masker further away. In addition, the effect of IM is investigated by applying either speech or speech-modulated noise maskers. To try to explain the measured intelligibility improvement between the colocated and spatially separated conditions, several signal-based measures are implemented – such as short-term and long-term intelligibility weighted SNR, cross-ear glimpsing and modulation domain SNR. Finally, the role of IM is discussed, particularly in regards to conditions with the masker nearby and the target far away.

Chapter 3 expands the study of Chapter 2 to investigate the effect of spatial separation in distance on speech intelligibility in HI listeners. Again, target and masking sound sources presented at various distances. To compensate for the hearing loss, linear amplification suited for each particular hearing loss is applied to (partly) restore audibility. The findings and discrepancies between the NH and HI listeners are discussed; especially the effect of nearby maskers in the presence of a target that is further away.

Chapter 4 aims to quantify the amount of IM that can be observed in a multi-talker reverberant environment resembling the cocktail party. A 3D sound-reproduction technique is used to simulate and auralize an environment that includes room acoustics and a more natural background. In this environment, a speech experiment investigates the effect of spatial separation, amount of masking talkers and talker similarity. Applying a masker comprised of unintelligible vocoded speech or speech with the same talker as the target, the upper and lower boundaries of IM were quantified. Similarly, the effect of spatial separation was measured by comparing intelligibility in conditions where sources were colocated or spread throughout the room.

The work presented in Chapter 2, 3 and 4 is tied together in **Chapter 5**. It presents a study where NH and HI listeners are again placed in an auralized cafeteria environment, but this time it included maskers closer to the listener than the target. SRTs were measured either with or without these nearby maskers, and with nearby maskers with different degrees of angular separation. As before, the contribution of IM in the masker was varied by applying a speech or vocoded masker. The chapter discusses the effects of nearby maskers and how these relate to IM and real-life listening. As the effect and susceptibility to IM is often linked to cognition, additional cognitive tests possibly related to IM were conducted, and the cognitive outcomes were compared to intelligibility performance.

Finally, **Chapter 6** summarizes the main findings of this thesis and discusses their role in the overall understanding of the cocktail party problem. It also presents implications and recommendations for hearing device processing and suggests further research in the field of IM and spatial hearing.

Initially this thesis considered the effect of distance-related reverberation on binaural acoustics and a binaural dereverberation algorithm was proposed. While this work inspired the subsequent directions of the project, it did not fit into the remainder of the thesis as a whole. The resulting publication is therefore presented in **Appendix A**.

Chapter 2

The effect of spatial separation in distance on the intelligibility of speech in rooms¹

The influence of spatial separation in source distance on speech reception thresholds (SRTs) is investigated. In one scenario, the target was presented at 0.5 m distance, and the masker varied from 0.5 m distance up to 10 m. In a second scenario, the masker was presented at 0.5 m distance, and the target distance varied. The stimuli were synthesized using convolution with binaural room impulse responses measured on a dummy head in a reverberant auditorium, and they were equalized to compensate for distance-dependent spectral and intensity changes. All sources were simulated directly in front of the listener. SRTs decreased monotonically when the target was at 0.5 m and the speech-masker was moved further away, resulting in a SRT improvement of up to 10 dB. When the speech masker was at 0.5 m and the target was moved away, a large variation across subjects was observed. Neither short-term signal-to-noise ratio (SNR) improvements nor cross-ear glimpsing could account for the observed improvement in intelligibility. However, the effect might be explained by an improvement in the SNR in the modulation domain and a decrease in informational masking. This study demonstrates that distance-related cues can play a significant role when listening in complex environments.

¹Manuscript submitted for publication to The Journal of the Acoustical Society of America.

2.1 Introduction

One of the greatest challenges for the auditory system is listening in environments where speech is degraded by spatially distributed maskers and room reverberation. This “cocktail-party” situation and the related auditory mechanisms have been topic for numerous studies (see Cherry, 1953 or Bronkhorst, 2000 for an overview). Spatial release from masking (SRM) characterizes the ability to use the angular difference in location between a target and a masker to better understand the target. With a frontal target talker, moving speech maskers from a colocated position (at 0°) to the side ($\pm 90^\circ$) in reference to the listener has shown improvements in speech intelligibility of up to 20 dB (Cameron and Dillon, 2007; Freyman *et al.*, 1999; Kidd *et al.*, 1998). However, in real-life listening environments, maskers are not only separated in angular direction but also in distance. Very few studies have looked into the effect of distance between a target and a masker on speech intelligibility.

SRM was originally measured in noise or speech babble (e.g. Bronkhorst, 2000; Bronkhorst and Plomp, 1990; Kock, 1950). Later studies found that the effect is more pronounced in presence of speech maskers (Best *et al.*, 2013b; Freyman *et al.*, 1999; Kidd *et al.*, 1998). This difference has been explained by the concepts of *energetic* and *informational masking*. Energetic masking occurs because of overlap between target and masker in the auditory periphery. This is the only type of masking offered by noise or speech babble. Complimentary informational masking has been used to describe masking effects that occur subsequent to the auditory periphery and as providing additional masking on top of the energetic component (for full review see Kidd *et al.*, 2007). Informational masking is often associated with auditory grouping and auditory stream segregation, cognitive abilities, working memory and attention. For speech recognition Kidd *et al.* (2007) mentions two different sources of informational masking: one due to failures in segregation of target and masker because of similarity (here ascribed as “confusions”) and a second due to the masker misdirecting or stealing the attention of the listener (here ascribed as “distractions”).

Different binaural mechanisms are underlying SRM. One of the main factors is fluctuating signal-to-noise ratios (SNRs) across ears caused by head-shadow as maskers are moved to the side. Brungart (2012) used ideal-binary masks to combine binaural signals over time and frequency to a monaural better-ear representation. Using this method

he showed that for maskers located at $\pm 60^\circ$, head-shadow alone could account for 5 dB out of a 6 dB SRM. In addition to head-shadow, so-called binaural interaction has been linked to SRM. Binaural interaction describes the ability to use interaural phase differences (IPDs) occurring when maskers are moved to the side. However, in Brungart's results the binaural interaction only amounted to 1 dB of the SRM. In addition to better-ear listening and binaural interaction, perceived spatial separation of target and masker reduces informational masking. Glyde *et al.* (2013) conducted similar experiments as Brungart (2012) using the LISN-S corpus, which allowed them to vary the amount of informational masking by either using masking talkers that were different (but same sex) to the target talker or identical to the target talker. They found a SRM of 12 dB in the same-talker condition and 9 dB in the different-talker condition. In both conditions, 6 dB of the SRM could be explained by better-ear processing, and thus by a reduction in energetic masking, whereas the remaining 6 dB or 3 dB, respectively, were attributed to a release in informational masking. Best *et al.* (2013b) and Brungart *et al.* (2001) looked into the difference in intelligibility between a speech masker similar to the target and a speech-modulated noise masker with the same energetic masking properties as the speech masker while providing no informational masking. The difference between the two, which is considered a measure of the involved informational masking, was up to 10 dB for the colocated condition, but smaller for the separated condition.

Distance perception is an aspect of localization which has been given considerably less attention than horizontal localization, particularly in connection with speech intelligibility measures. Where human horizontal localization is sensitive down to just noticeable differences of 1 degree (Blauert, 1997), distance is a less salient measure and is often dominated by vision. For auditory distance perception several cues are available (for a review see Zahorik, 2005). The most predominant cue is signal intensity. As distance increases, signal levels for omni-directional sound sources decrease proportionally to the inverse of their distance (Kuttruff, 2000). This cue is especially relevant for speech signals, as listeners are able to estimate the source level from the applied vocal effort. When sounds are presented in reverberant environments, the auditory system can additionally use the signal's direct-to-reverberant ratio (DRR) to determine distance. The strength of reverberation is almost independent of position, and as the direct sound energy will decrease with distance the DRR decreases accordingly. Zahorik (2002a) measured DRR just-noticeable differences (JNDs) in NH listeners and found that this cue only provided

a coarse estimate of distance as the lowest JNDs required a doubling of distance. In addition to intensity and DRR, other distance related cues mainly include interaural level differences (ILDs) (Brungart and Rabinowitz, 1999) at very close distance and spectral cues from air absorption at very far distances (Zahorik, 2005).

Shinn-Cunningham *et al.* (2001) and later Brungart and Simpson (2002) investigated SRM related to differences in distance combined with angular separation in an anechoic environment. However, both studies focused on ILD cues which occur at very near distances (< 1 m) when the sources are to the side of the listener (45° and 90°). They also considered only anechoic environments. Both studies found a substantial effect of distance on intelligibility, especially when the masker was similar to the target. Bronkhorst and Plomp (1990) included both reflections and reverberation and also considered sources either in a direct or mainly reverberant (near and far) sound field, however they only applied modulated and unmodulated speech-shaped noise maskers. Their results showed speech reception threshold (SRT) improvements of about 1 dB when moving the masker from near to far while keeping the target near. To the best knowledge of the authors, no studies have systematically investigated SRM resulting from differences in distance further than 1 m in a reverberant environment using speech maskers.

The current study investigates SRM occurring from differences in distance mainly considering room effects as intensity and spectral cues are equalized. Binaural room impulse responses (BRIRs) measured in an auditorium are used to spatialize the speech signals. The colocated condition is compared to separated conditions where either the target or masker is moved further away. Two speech corpora with different characteristics are applied. The Coordinate Response Measure (CRM) is used as main measure and the Listening in Spatialized Noise-Sentence test (LISN-S) is applied to verify the results. Furthermore, an objective analysis is performed to better understand the physical cues underlying the findings. Here the concepts of segmental (or short-term) SNR improvements, cross-ear glimpsing and modulation domain signal to noise ratio changes are investigated. Furthermore, a discussion focusing on the potential involvement of informational masking effects is provided.

2.2 Methods

2.2.1 Subjects

Sixteen subjects (11 female and 5 male) aged between 20-49 years (mean 33.8) participated in this study. All subjects had normal hearing (< 20 HL), determined by a pure-tone audiogram from 500 Hz to 8 kHz, and were native Australian English speakers. Subjects were either employed at the National Acoustic Laboratories or students at Macquarie University and gave written consent before participating in the study. Subjects not connected to the National Acoustic Laboratories were given a gratuity for their participation.

2.2.2 Stimuli

Two speech corpora were used in this experiment, the CRM (Bolia *et al.*, 2000) and sentences from the LISN-S (Cameron and Dillon, 2007). Both corpora are often used when measuring SRM and apply speech maskers that provide significant informational masking.

The CRM corpus consists of sentences spoken by four male and four female talkers. Here only the four male talkers were used. Each sentence has the structure: “Ready [call sign] go to [color] [number] now”, with eight call-signs (“Arrow”, “Baron”, “Charlie”, “Eagle”, “Hopper”, “Laker”, “Ringo”, “Tiger”), four colors (red, green, blue and white) and eight numbers (1 through 8), resulting in 256 sentences for each talker. The target was always given the call-sign “Baron”, but the color/number coordinate and talker was randomly chosen.

Two types of maskers were applied with the CRM corpus; a speech masker and a speech-modulated noise masker. For each target sentence, the speech masker consisted of two random CRM sentences with talker, call-sign and number/color coordinate different from the target. To measure the contribution of energetic masking, a speech-modulated noise masker according to Best *et al.* (2013b) was applied. The speech-modulated noise masker was realized by applying the low-pass filtered (50 Hz) Hilbert envelope of two randomly chosen CRM sentences to noise with the long-term spectrum of the entire male CRM corpus. This masker carries most of the temporal fluctuations of the speech

masker but avoids talker confusions (i.e. it only provides energetic masking). The task is thus reduced to identifying the correct color/number coordinate in a noise-like mixture.

The main advantages of the CRM corpus are the minimal learning effects, allowing an indefinite amount of repetitions, and the large masking release resulting from spatial separation. The main disadvantage of the CRM corpus is that the sentences are roughly time-aligned and thus, changes in the temporal behavior of the target and masker might change the masking characteristics of the corpus. Applying BRIRs to the target and masker signal to introduce room-related distance cues (see Sec. 2.2.3) causes differences in arrival time as well as temporal smearing (or temporal spread of energy). This may result in significant parts of the target signal to stand out from the masker signal and thereby providing artificial cues that are only relevant to the CRM corpus. Arrival time differences were accounted for in this study by removing the initial delay of all applied BRIRs. However, the varying temporal spread of reverberant energy might still have an effect on the masking properties of the CRM corpus that cannot be avoided.

To ensure that the potential effect of differences in source distance on speech intelligibility is not only a methodological artifact, a speech corpus with very different properties, the LISN-S, was additionally applied. The LISN-S provides a sentence recall task in a continuous two-talker background (Cameron and Dillon, 2007). The target consists of four sets of 30 short sentences (e.g. “Mom is driving carefully”). The onset of each sentence is signaled by a preceding 200 ms long tone at 1000 Hz. The masker consists of two simultaneously presented children stories, with a duration of approximately 150 seconds, continuously looped throughout the test. Both the target and masker talkers are native Australian English speaking females. The informational masking of the masker was adjusted by using either maskers spoken by the target talker or two different female. Since the masker consists of continuous speech and the target is presented at a random instance in the masker mixture, any “pop-out” effect of the target signal due to a temporal misalignment in energy between target and masker should be minimized by using the LISN-S instead of the CRM corpus. A major draw-back of the LISN-S is that, due to the limited number of sentences, it only allows testing of four conditions.

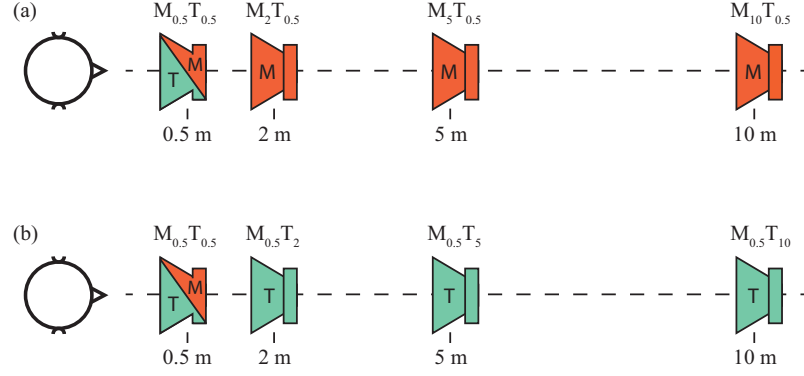


FIGURE 2.1: Schematic of the applied conditions and labeling. Top: Target fixed at 0.5 m and varying masker distance. Bottom: Masker fixed at 0.5 m and varying target distance.

2.2.3 Spatialization of sounds

To recreate the auditory sensation of listening in a room with sources at different distances, the anechoic speech stimuli were convolved with measured BRIRs. The BRIRs were recorded, with a sample rate of 44.1 kHz, in an auditorium using a Brüel & Kjær Type 4100 Head and Torso Simulator (HATS) and a DynAudio BM6P two-way loudspeaker. The auditorium had a volume of approximately 1150 m³ and the reverberation time (T_{30}) in octave bands shown in Tab. 2.1. The BRIRs were measured using 30 second long logarithmic sweeps (see Muller and Massarani, 2001). The position of the HATS was kept constant and BRIRs were measured with the source at 0.5 m, 2 m, 5 m and 10 m distance directly in front of the HATS and the DRR calculated with a 2.5 ms direct-sound window (as in Zahorik, 2002a) was 15.1 dB, 5.3 dB, 0.0 dB and -7.7 dB, respectively.

TABLE 2.1: Auditorium reverberation time in octave bands

F_c (Hz)	250	500	1000	2000	4000	8000
$T_{30}(s)$	1.10	1.72	1.92	1.88	1.43	0.95

The different spatial configurations tested with the CRM corpus are illustrated in Fig. 2.1 and summarized in Tab. 2.2. Note that the indices of the spatial conditions show the applied distances as well as the masker type, i.e. M_{s10} is a speech masker at 10 m distance and $M_{n0.5}$ is a speech-modulated noise masker at 0.5 m distance. Due to the limited number of target sentences, the LISN-S was only tested for $M_{s05}T_{s05}$ and $M_{s10}T_{s05}$, but both the same talker as the target and the different talker condition was tested to examine the effect of different amounts of informational masking.

Increasing the distance of a sound source delays the arrival of the direct sound, changes the overall spectrum and reduces the overall sound intensity. Arrival time differences were compensated by removing the initial delay of the BRIRs and intensity changes were removed by normalizing the root mean square (RMS) level of the convolved speech signals. To remove long-term spectral differences, the masker spectrum was always equalized to match the long-term spectrum of the masker colocated with the target using a 512 tap finite impulse response (FIR) equalization filter designed and applied with the MATLAB commands `fir2` and `filter`. For the CRM, the target was either at 0.5 m or 10 m distance and hence the maskers were equalized to either of these two long-term spectra illustrated in the left panel of Fig. 2.2. For the LISN-S both the same voice and different voice masker were equalized so that the long-term spectrum matched that of the same voice masker in the colocated condition shown in the right panel of Fig. 2.2.

To investigate the effect of binaural processing, conditions with diotic presentation were tested. The diotic stimuli were realized by supplying either the left or right ear signals to both ears. The test subjects were divided so that half of the subjects received the version with the left ear and the other half the right ear signals.

2.2.4 Procedures

Experiments were carried out in a double-walled booth, with the experimenter (for LISN-S) or subject (for CRM) interacting with a Windows-based silent computer (no moving components) running MATLAB. The signals were presented via equalized Sennheiser HD-215 circumaural headphones driven by a RME Hammerfall HDSPe AIO sound-card. For both speech corpora, the masker was kept at a level of 55 dB SPL, measured

TABLE 2.2: Conditions and labeling used in the experiment

Condition name	Masker type	Masker distance	Target distance
$M_{s05}T_{s05}$, $M_{s2}T_{s05}$, $M_{s5}T_{s05}$, $M_{s10}T_{s05}$	Speech	0.5 m, 2 m, 5 m, 10 m	0.5 m
$M_{s05}T_{s2}$, $M_{s05}T_{s5}$, $M_{s05}T_{s10}$	Speech	0.5 m	2 m, 5 m, 10 m
$M_{n05}T_{s05}$, $M_{n10}T_{s05}$	Speech-modulated noise masker	0.5 m, 10 m	0.5 m
$M_{n05}T_{s10}$	Speech-modulated noise masker	0.5 m	10 m

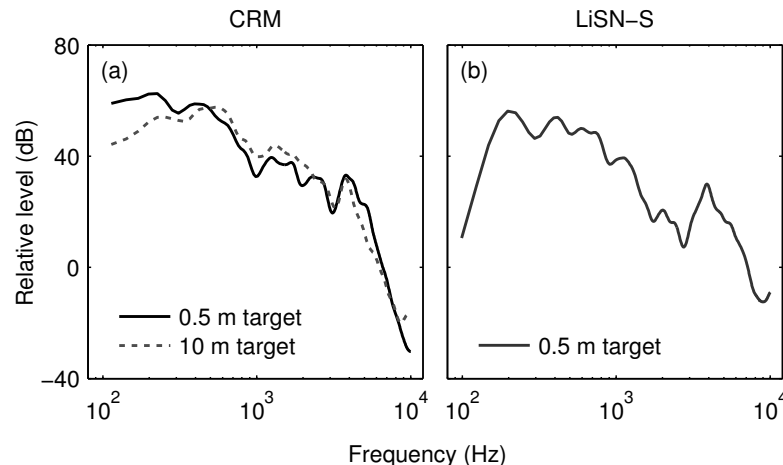


FIGURE 2.2: (Left) Equalized long-term spectra in critical bands of CRM masker and target sentences when the target is at 0.5 m (black solid line) and 10 m (gray dashed line). The masker was always equalized to the long-term spectrum in the target position. (Right) Equalized long-term spectra in critical bands of the LiSN-S maskers in all conditions. The LiSN-S maskers were spectrally matched with the same-talker masker in the colocated position.

in a Brüel & Kjær type 4153 artificial ear. In both tests, the level of the target sentences was initially set to 62 dB SPL and varied relative to the masker following a 1-up 1-down rule, thereby adaptively estimating the SRT (i.e. the 50% point on the psychometric function).

The tests were divided in two parts; in the first part the LiSN-S thresholds were measured. The four sets of sentences were matched with the four masker conditions (see table 2.2) according to a Latin square design. This design balanced effects of order and sentence list, which is particularly important because LiSN-S does not include training. As in Cameron and Dillon (2007), the test completed either when the subject reached 30 sentences in one condition or when the standard error fell below 1 dB after a minimum of 17 sentences. Furthermore, the subjects were instructed in accordance with Cameron and Dillon (2007) and this part took approximately 20 minutes.

In the second part, SRTs were measured using the CRM. Before testing, the subjects were instructed to listen for the color/number coordinate of the speaker with the “Baron” call-sign and press the corresponding button in a MATLAB GUI. Subjects were familiarized with the task by testing a random condition which was excluded from the final results. The presentation order of the conditions was randomized and all threshold measurements were repeated once. The test was completed after nine reversals. The familiarization took approximately 5 minutes and the main task 45 minutes, resulting in a total test duration of 90 minutes with instructions and breaks.

2.3 Results

2.3.1 Varying masker distance

Fig. 2.3 shows the results for the CRM corpus as a function of masker distance with the target fixed at 0.5 m distance. The top panel shows the mean SRTs and the corresponding 95% confidence intervals across subjects. The bottom panel shows the mean value of the spatial advantage and 95% confidence intervals. The spatial advantage is determined by subtracting the individual SRT in the spatially separated condition from the individual SRT in the colocated condition for each subject separately. The masker was either two-talker speech (circles) or fluctuating noise with a similar envelope to the speech masker (diamonds). In the case of the speech masker, increased masker distance decreased the SRT. When the masker was at 10 m ($M_{s10}T_{s0.5}$) the mean spatial advantage (as defined by the difference in SRT from the colocated condition) was approximately 10 dB. The SRT measured with the speech-modulated noise masker was independent of masker distance with SRTs similar to the SRT measured with the speech masker in the maximally separated condition. For the speech masker, a repeated measures two-way analysis of variance (ANOVA) showed significance for condition [$F(3, 45) = 345.2$, $p < 0.001$], but neither for repetition [$F(1, 15) = 0.2$, $p = 0.66$] nor interaction [$F(3, 45) = 1.8$, $p = 0.15$]. Post-hoc paired comparison with Bonferroni correction showed that SRTs for all distances measured with the speech masker were significantly different from each other ($p < 0.005$). A paired comparison (t-test) indicated no significant difference ($p = 0.21$) between the speech-modulated noise masker thresholds at the two distances.

In Fig. 2.4 the results of varying masker distance are shown for the LISN-S corpus together with the corresponding CRM results replotted from Fig. 2.3. Overall, the decrease in SRTs (or increase in intelligibility) with increased target-masker separation observed with the CRM was replicated with the LISN-S sentences. SRTs decreased by about 2 dB in the colocated condition between the same- and different-talker masker, but did not change in the spatially separated condition. This behavior resulted in a slightly reduced spatial advantage with the different-talker masker compared to the same-talker masker. The difference in SRTs in the colocated condition can be explained by pitch cues, introduced by the different (female) talkers, resolving some of the talker confusions. This

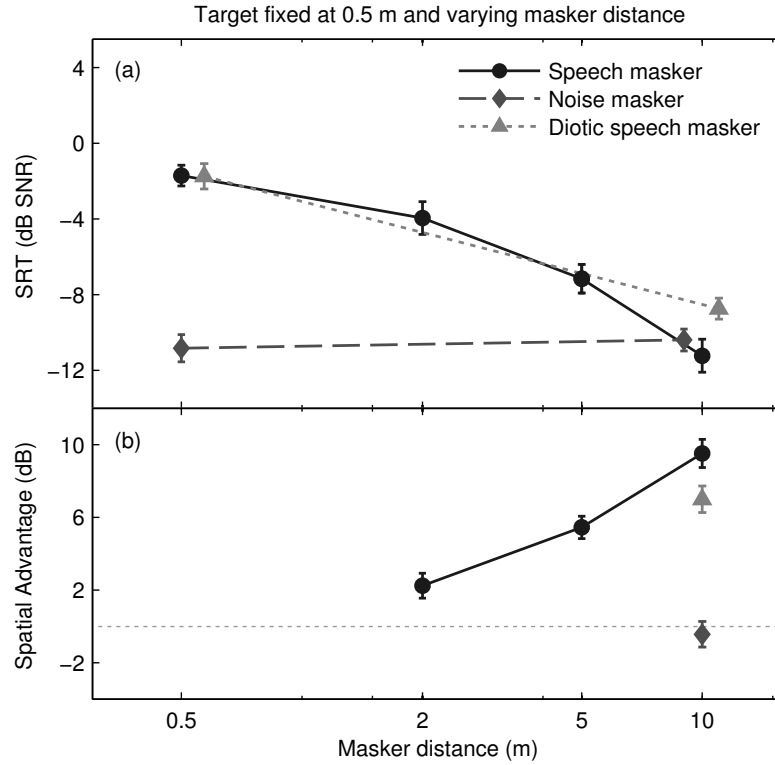


FIGURE 2.3: Top panel: Mean and across-subject 95% confidence intervals of the SRT (expressed per masker) for the CRM corpus with target fixed at 0.5 m distance and varying masker distance. Bottom panel: Mean and across-subject 95% confidence intervals of the spatial advantage (i.e. the difference between the related result and the $M_{s0.5}T_{s0.5}$ condition).

2 dB difference is consistent with that found by Cameron and Dillon (2007). A repeated measures ANOVA showed significance for condition [$F(1, 15) = 98.04, p < 0.001$] as well as talker [$F(1, 15) = 20.85, p < 0.001$] and for interaction [$F(1, 15) = 6.31, p < 0.05$]. A paired comparison (t-test) showed significant difference between talker-types in the colocated condition ($p < 0.005$) but not in the separated condition ($p = 0.09$). The fact that the distance related masking release observed with the CRM corpus is also present with the LISN-S corpus confirms that the effect of distance found with the CRM is not simply an artifact due to the temporal smearing introduced by room reverberation as discussed in Sec. 2.2.2.

2.3.2 Varying target distance

The mean SRT values and 95% confidence intervals across subjects when the target distance varied and the masker was fixed at 0.5 m for the CRM corpus are shown in Fig. 2.5. The results with the two-talker speech masker is indicated by the filled circles and the speech-modulated noise masker by the diamonds. For the speech masker

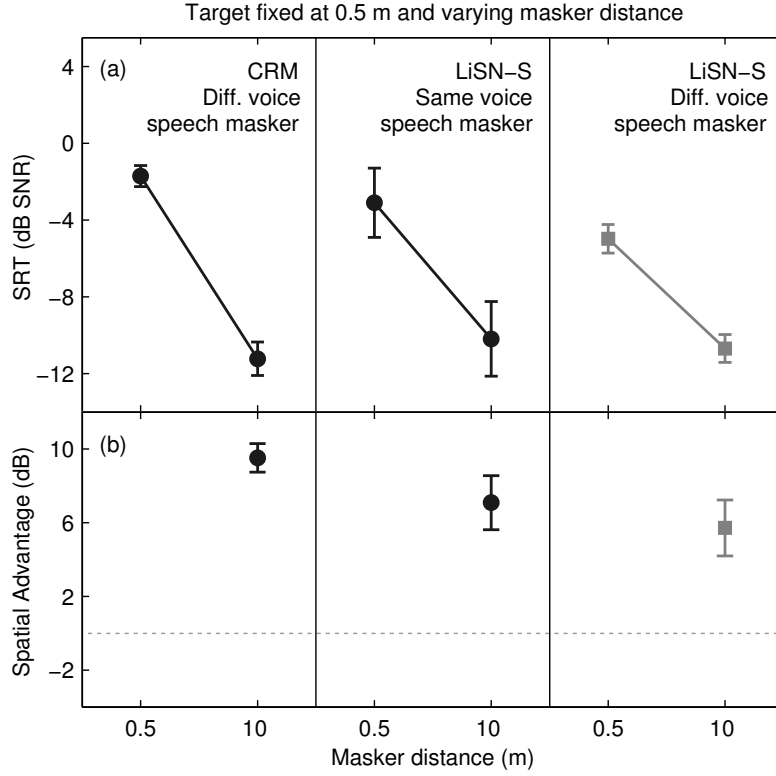


FIGURE 2.4: Top panels: Mean and across-subject 95% confidence intervals of the SRTs (expressed per masker) for the CRM and LISN-S corpus with target fixed at 0.5 m distance and masker either at 0.5 or 10 m distance. The left panels replot the CRM results (also found in Fig. 2.3) for reference. The middle panels show results with the same masking talker as the target. The right panels show LISN-S results using maskers with a talker different than the target talker. Bottom panels: Mean and across-subject 95% confidence intervals of the spatial advantage (i.e. the difference in SRT from the colocated, i.e. $M_{s0.5}T_{s0.5}$, condition).

condition, individual data are additionally shown and indicated by the open circles. For both masker conditions only a small effect of distance on the mean SRT is observed. However, when considering individual data for the speech masker, moving the masker to 10 m decreased SRTs substantially for some subjects (up to 10 dB) while increasing SRTs by several dB for other subjects. This resulted in a large variability across subjects, especially with the target at 5 m and 10 m distance. A repeated measures two-way ANOVA showed significance for repetition [$F(1, 15) = 11.54, p < 0.005$] and condition [$F(1.69, 25.22) = 4.29, p < 0.05$] but not for interaction [$F(1.76, 26.45) = 2.76, p = 0.09$]. Here the Greenhouse-Geisser correction factor was applied to ensure that sphericity violation did not influence the significance calculation. According to a paired comparison (t-test) the speech-modulated noise masker provides no significant advantage ($p = 0.06$) when changing the target from 0.5 m to 10 m distance. The fact that the SRT does not change with distance in the speech-modulated noise masker condition for all test subjects indicates that the higher SRTs observed in some subjects with close speech

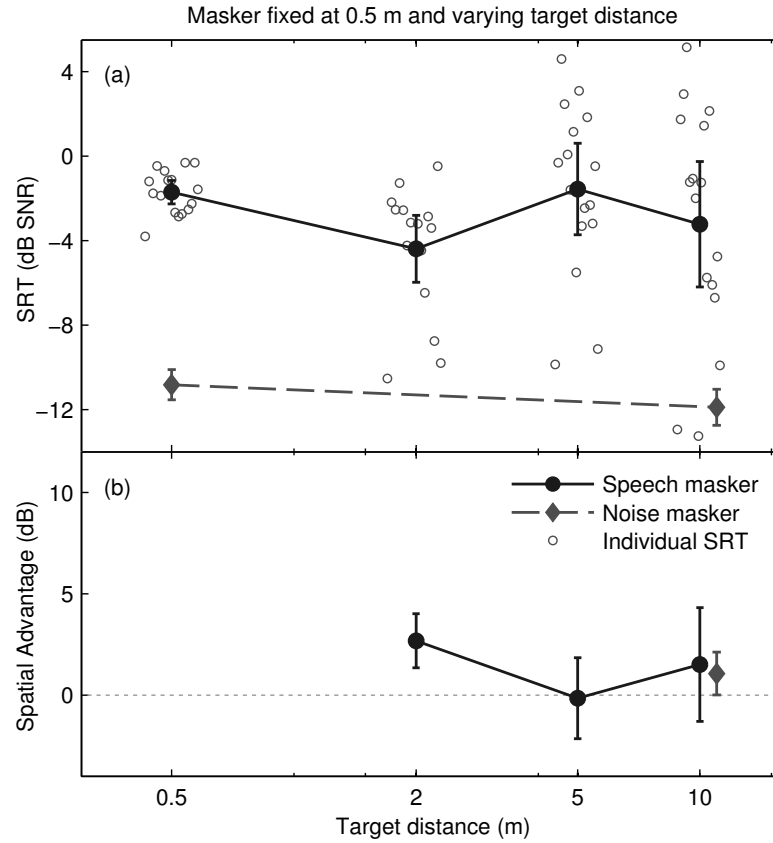


FIGURE 2.5: Top panel: Mean and across-subject 95% confidence intervals TMR at SRT (expressed per masker) for the CRM corpus with masker fixed at 0.5 m distance and varying target distance. Bottom panel: Mean and across-subject 95% confidence intervals of the spatial advantage (i.e. the difference between the related result and the $M_{s0.5}T_{s0.5}$ condition).

masker and distant target is not simply caused by a degradation of target intelligibility due to the significantly increased amount of reverberation.

2.3.3 Diotic versus dichotic presentation

In Fig. 2.3 the effect of diotic stimulus presentation is shown as the triangles measured with the CRM corpus with the target fixed at 0.5 m distance and the masker at 0.5 m and 10 m. Subject were divided into two equally sized groups which were presented with diotic versions of the right or left signal. A paired comparison (t-test) found no significant difference between the SRTs measured with the two groups ($p = 0.20$) and therefore the groups are combined in Fig. 2.3. A repeated measures three-way ANOVA showed significance for condition [$F(1, 15) = 1172.77$, $p < 0.001$] as well as diotic/dichotic presentation [$F(1, 15) = 23.35$, $p < 0.001$] but not for repetition [$F(1, 15) = 1.40$, $p = 0.26$]. The only significant interaction was between condition and diotic/dichotic presentation [$F(1, 15) = 24.17$, $p < 0.001$]. Hence, for the colocated condition no notable

difference between diotic and dichotic presentation was observed. However, when the masker is moved to 10 m distance the SRT is 3 dB higher for the diotic presentation. This suggests that the effect of distance on SRM is mainly a monaural process and the binaural benefit is limited to about 3 dB.

2.4 Discussion

2.4.1 Objective signal-based measures

In this section different signal properties are investigated which might have contributed to the increase in speech intelligibility that was observed in Sec. 2.3 when target and masker signals were separated in distance. These properties include short-term SNR fluctuations, cross-ear glimpsing and changes to the envelope domain SNR. The analysis only considers the colocated ($M_{s0.5}T_{s0.5}$) and maximally separated condition ($M_{s10}T_{s0.5}$) for the LISN-S corpus as an example. However, the principles and conclusions apply similarly to all other stimulus conditions of the LISN-S as well as the CRM corpus.

2.4.1.1 Short-term SNR improvements

Since the masker spectra were equalized, there was no long-term SNR effect of changes to target-masker distance. This was confirmed by applying the intelligibility-weighted long-term SNR benefit (Greenberg *et al.*, 1993). However, increasing the distance of a sound source inside a room will increase the relative energy of the reverberation (i.e., decrease the DRR) and will thereby smear both the temporal and spectral characteristics of the signal. Hence, even though the long-term spectra of the maskers were equalized in the present experiments (see Sec 2.2.3), the reverberation still affected the short-term behavior of the stimuli and may have created regions with improved SNR.

To test if the short-term SNR can explain the SRT improvement observed in the spatially separated conditions, a simple short-term SNR model was implemented similar to the one used by Glyde *et al.* (2013) and Brungart (2012). The left and right ear signals were filtered using a bandpass filterbank consisting of 18 fourth-order gammatone filters with 1/3-octave spacing covering the range from 160 Hz to 8 kHz (Glasberg and Moore, 1990). The output of each filter was segmented with a sliding 20 ms long rectangular window and for each segment the RMS value was calculated. Time-frequency segments

where the RMS level of the target signal was below the audibility threshold (taken from ISO 389-7, 1996) were discarded. The derived short-term SNR values in each individual frequency channel were collected in a histogram. Histograms in two example channels (1000 Hz and 4000 Hz) calculated for the entire LISN-S corpus are shown in the top panels of Fig. 2.6 for the colocated ($M_{s0.5}T_{s0.5}$, black line with median value) and spatially separated condition ($M_{s10}T_{s0.5}$, gray line with median value). The median value decreased from the colocated condition ($M_{s0.5}T_{s0.5}$) to the separated condition ($M_{s10}T_{s0.5}$) by 6.4 dB and 3.7 dB in the 1000 Hz and 4000 Hz channel respectively. This means that the overall SNR of the audible time-frequency frames in these channels is reduced by spatial separation. In addition to a decreased median SNR value, the spatial separation sharpens the distribution and especially reduces the occurrence of frames with very positive SNR values. Hence, the increased reverberation when the masker is moved from 0.5 m to 10 m distance effectively reduces the fluctuations in short-term SNR and removes time-frequency frames with very high SNR. To better quantify the reduction of short-term SNR due to increased target-masker separation, the intelligibility-weighted segmental SNR benefit was applied (Greenberg *et al.*, 1993; Hansen and Pellom, 1998). This resulted in a 2.92 dB SNR reduction when moving the masker from 0.5 to 10 m while keeping the target at 0.5 m, suggesting that a separation in distance reduces, rather than improves, speech intelligibility. Hence, the SRM observed in Sec. 2.3 when spatially separating the target and masker in distance can not be explained by short-term SNR effects. This general conclusion was confirmed for all other stimulus conditions.

2.4.1.2 Cross-ear glimpsing

In Sec. 2.3.3 it was shown that 3 dB out of the 10 dB in SRM measured in the $M_{s10}T_{s0.5}$ condition were attributed to binaural auditory processes. A significant amount of SRM associated with separating sources in their azimuth (horizontal) angle has previously been attributed to fluctuating SNRs across ears caused by head-shadow. Brungart (2012) and Glyde *et al.* (2013) implemented a cross-ear glimpsing model which combines signals across ears in order to create a better-ear representation of the signal. The model could account for a considerable amount of their SRM data. To test if a cross-ear glimpsing mechanism can also account for the binaural advantage observed in Sec. 2.3.3, a similar model was adopted by applying the short-term SNR model described above as a monaural front-end to the signals arriving at the left and right ear of a listener. In

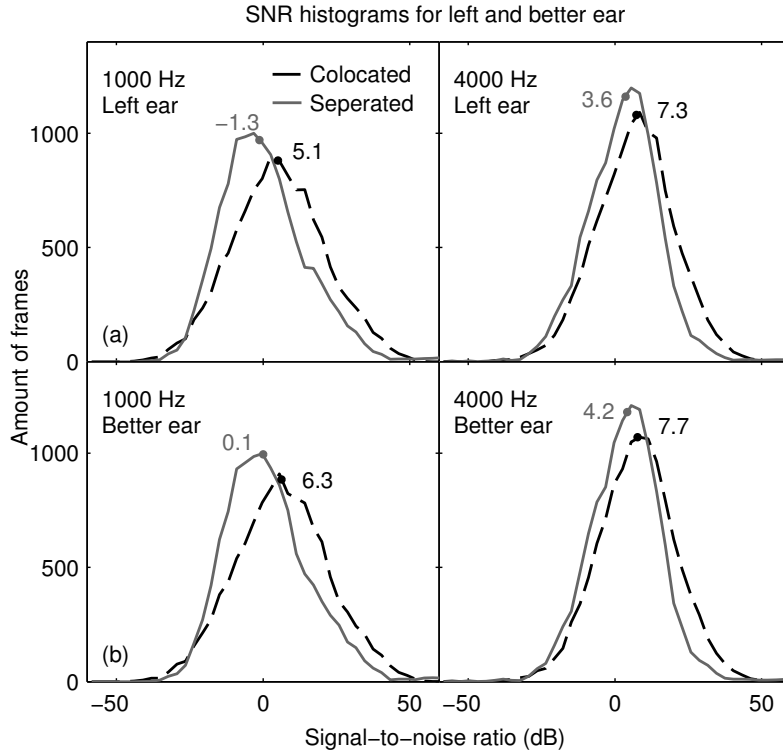


FIGURE 2.6: Histograms of SNR in 20 ms audible segments in two different gammatone frequency channels for either colocated ($M_{s0.5}T_{s0.5}$, black line) or separated presentation ($M_{s10}T_{s0.5}$, gray line). The upper panels display histograms for the left channel only whereas in the lower panels only include the time-frequency segment with better SNR across ears (cross-ear glimpsing).

each frequency channel, short-term SNRs were compared between the left and right ear and only the higher SNRs were collected in a better-ear histogram. Example better-ear histograms are shown in the bottom panels of Fig. 2.6 for the colocated ($M_{s0.5}T_{s0.5}$, black line with median value) and spatially separated condition ($M_{s10}T_{s0.5}$, gray line with median value) in the 1000 Hz and 4000 Hz frequency channel. There is only a small SNR difference between the left-ear signal and the better-ear signal as the median value shifted between 0.4 dB and 1.4 dB. Comparing the better-ear benefit between the colocated and separated condition very little benefit can be observed in these frequency channels. Since the placement of the sound sources in this study were all directly in front of the listener, the main source for any cross-ear glimpsing benefit would have been due to strong lateral reflections. However, these are less substantial than the direct sound component, especially since the nearest surfaces in the auditorium were far from the HATS.

To better quantify the overall potential contribution of cross-ear glimpsing to SRM in distance, again the intelligibility-weighted segmental SNR benefit was applied, resulting

in a broad-band cross-ear glimpsing benefit of 0.54 dB in the colocated condition and 0.61 dB in the spatially separated condition. Hence, a cross-ear glimpsing mechanism is not very likely to explain the binaural advantage of 3 dB observed in Sec. 2.3.3 when moving the masker further away than the target. A further analysis of alternative binaural mechanisms that may provide a reduction in either energetic or informational masking is out of the scope of this study.

2.4.1.3 Envelope domain signal to noise ratio

The effects of room acoustics on speech intelligibility has been studied extensively and incorporated in the speech transmission index (STI; IEC 60268-16, 2011) using the room modulation transfer function (MTF) (Kuttruff, 2000). Hence, the STI seems to be a promising approach to better understand or even predict the effect of target-masker distance changes on the SRT as described in Sec. 2.3. However, the STI only considers the MTF of the target signal and not the nonlinear interaction between target and masker modulations. The latter aspect seems to be essential for describing the SRM observed in Sec. 2.3.1 when the masker location is changed in distance but the target location is fixed. Therefore, the approach by Jørgensen and Dau (2011), which considers the SNR in the envelope domain (SNR_{env}) as a predictor for speech intelligibility, might be promising. Based on an auditory model, they compared the modulation power spectrum of the signal mixed with the speech-modulated noise masker and of the signal alone and successfully applied the SNR calculated in this domain. With reference to the stimuli described in Sec. 2.2, increasing the distance of a sound source increases the reverberant energy (or decreases the DRR) which reduces fast temporal fluctuations of the target speech and thereby effectively low-pass filters its modulation spectrum. To investigate if the changes in SRT observed in Sec. 2.3 and Figs. 2.3-2.5 for varying target and masker distance can be explained by changes in the SNR in the envelope domain, the “modulation-excitation pattern” approach proposed by Jørgensen and Dau (2011) was applied. Modulation-excitation patterns were derived for two example auditory frequency channels at 1000 Hz and 4000 Hz by integrating envelope power at the output of a modulation filterbank with bandpass filters realized as second order Butterworth filters from 1 to 64 Hz. The resulting modulation excitation patterns are shown in Fig. 2.7. For the top panels, the target and masker were colocated at 0.5 m. In this case, the SNR in the modulation domain, as quantified by the area between the “masker

only” (diamonds) and “masker and target” (squares) condition, is not notable for both the 1000 Hz and 4000 Hz frequency channel. However, when the masker is moved to 10 m distance (bottom panels) there is a noticeable SNR difference between the “masker only” condition and the “masker and target” mixture. The masker alone exhibits substantially less modulation power when moved further away, which corresponds to temporal gaps being filled and an overall decrease in signal fluctuations. Comparing the top and bottom panels indicates that the “masker and target” mixture also decreases in modulation power when the masker is moved further away, but to a lesser extent than the masker alone condition. This suggests that the target is more easily detected in the spatially separated condition than in the colocated condition. Hence, a model that measures the SNR in modulation-domain, such as proposed by Jørgensen and Dau (2011), would be able to, at least, qualitatively predict the SRT data shown in Figs. 2.3 and 2.4. This was not possible with a model that only relies on a measure of the SNR in the frequency domain as described in Sec. 2.4.1.1.

However, an intelligibility model that is purely based on the the modulation domain SNR can not describe any informational masking effects. Since both the CRM and LISN-S speech corpus involve a significant amount of informational masking (Brungart *et al.*, 2001; Glyde *et al.*, 2013), it is not expected that such model can fully predict the SRT data given in Sec. 2.3. In particular, it will fail to predict the SRT data for the case that the target is moved further away (Sec. 2.3.2, Fig. 2.5), i.e., where most likely distraction-based informational masking is the dominant effect (see Sec. 2.4.3 for a detailed discussion). Even though a modulation-domain SNR model might be the most promising approach to predict at least some of the behavior of the measured SRT data, a further analysis is out of the scope of the present study.

2.4.2 Effect of equalization

Throughout this study, both the level and long-term spectra of the applied stimuli were equalized. This removed some of the acoustic properties related to changes in source-distance and may have increased the likelihood of target-masker confusions. But at the same time it allowed for better comparison between the applied conditions and emphasized on changes in the DRR. Ultimately, an adaptive intelligibility measure (as used to estimate the SRT) manipulates/removes level differences to vary the SNR. Naturally, in realistic environments where multiple talkers are present, level differences related

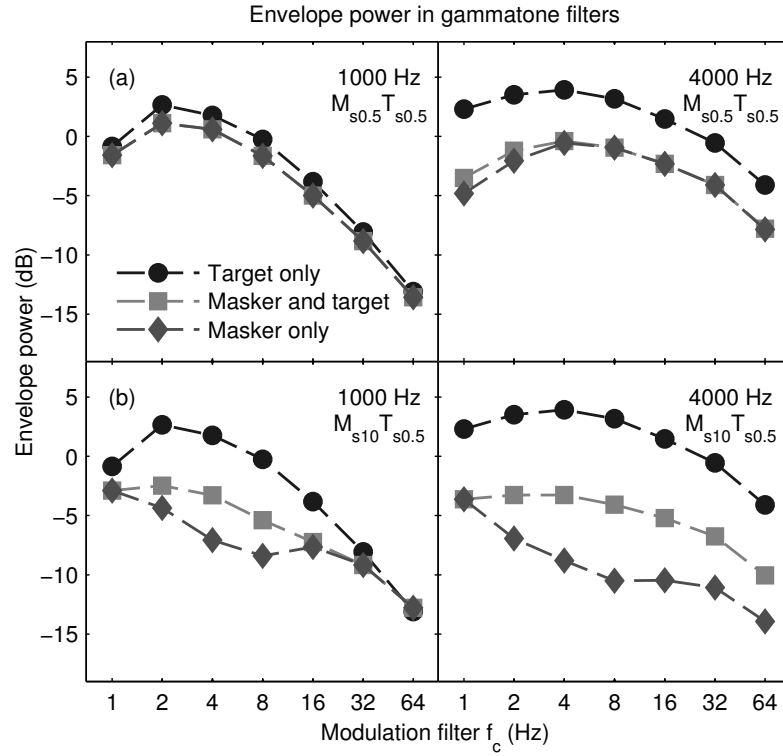


FIGURE 2.7: Envelope excitation patterns for either colocated ($M_{s0.5}T_{s0.5}$, top panels) or separated presentation ($M_{s10}T_{s0.5}$, bottom panels) in gammatone filters with center frequencies of 1000 Hz (left panels) or 4000 Hz (right panels). The excitation patterns are shown for the target alone (circles), mixture of target and masker at 0 dB SNR (squares) and for the masker alone (diamonds). The envelope domain SNR, SNR_{env} , is defined as the difference between the mixture and masker alone.

to the distance of talkers will affect energetic masking, because of differences in SNR, and informational masking because of loudness cues which facilitate talker segregation (Brungart *et al.*, 2001). Furthermore since the stimuli were equalized in level, the perceived loudness of sources at each distance could have been different. Most likely the far away sources might have sounded louder than the nearby sources. But if this was the case, then the SRTs should increase with increasing masker distance, which is in clear contradiction with the data shown in Fig. 2.3. In the end, there are many ways that the signals could have been equalized, e.g. with respect to the direct sound alone, direct sound and early reflections or loudness. Even though such investigation would be very interesting, it is out of the scope of the current study.

The applied spectral equalization filters described in Sec. 2.2.3, removed long-term spectral differences between the separated target and masker. The filters mainly addressed the low-pass filter that is introduced by the room for far sound source. This is due to room reverberation dominating the overall signal power, which at high frequencies

is affected by increased air absorption as well as decreased reflectivity of the used wall materials. It was assumed here that this processing has no significant effect on the measured results and conclusions, but ensured that SRTs were comparable in terms of long-term SNR. To confirm that this assumption is appropriate, an additional experiment with eight NH listeners (five of whom participated in the original experiment) was conducted. This experiment measured condition $M_{s0.5}T_{s0.5}$, $M_{s10}T_{s0.5}$ and $M_{s0.5}T_{s10}$ with and without frequency equalization using the CRM corpus and was repeated once. A t-test with Bonferroni correction showed no significant difference ($p > 0.05$) between SRTs measured with and without the equalization. Moreover, the data was very similar to the corresponding data shown in Fig. 2.3 and Fig. 2.5 and thus, is not explicitly shown here.

It can be further discussed whether the effects observed in this study are actually due to a perceived difference in distance or due to a change in the DRR. Since the feeling of distance is generally hard to evoke without visual cues or externalized signals (i.e. sources are heard as being outside the listeners head), it is quite possible that the difference in DRR is the main cue for talker segregation. But since these cues originate from the difference in distance (and the room) it is not detrimental for the study, just worth considering if one would try to make quantitative estimates from these results. Furthermore, the study only considered maskers that were directly in front of the listener - hence the binaural differences were very small. It is very likely that the binaural effect (shown in Fig. 2.3) would have been more substantial if the maskers had been changed in their angular direction. In the end, hopefully this work will inspire further investigation in which, particularly the interaction and individual contribution of distance and directional cues are studied systematically.

2.4.3 Role of informational masking

Throughout literature the detrimental effects of reverberation on speech intelligibility are well documented (Bronkhorst and Plomp, 1990; Nábelek and Robinson, 1982) and modeled (IEC 60268-16, 2011). Since the main difference between the different spatial configurations in this study are changes to the DRR, the results demonstrate certain conditions where reverberation actually improves intelligibility. While the previous sections explored the signal-related causes to the observed effect, this section considers the potential role of informational masking and perceptual segregation of sound sources.

2.4.3.1 Varying masker distance

Considering conditions where the target is kept at 0.5 m in Figs. 2.3 and 2.4, the experimental data revealed that the SRT decreased when the speech masker was moved further away (e.g., moving from $M_{s0.5}T_{s0.5}$ to $M_{s10}T_{s0.5}$). However, when the speech-modulated noise masker was used, very low SRTs were observed independent of masker distance. This effect may be explained by informational masking, as the speech-modulated noise masker does not cause target/masker confusions. Considering the speech masker, the condition where target and masker are colocated ($M_{s0.5}T_{s0.5}$) is known to provide the highest amount of informational masking (Freyman *et al.*, 2001; Kidd *et al.*, 2007). The decreasing thresholds as speech maskers move further away, indicate that differences in distance aids perceptual segregation of target and masker, just like angular separation in angular SRM. One could say that the confusions between target and masker are resolved by the perceived difference in distance (or the difference in the DRR). The observation that there is no difference between the SRT results for the far speech masker ($M_{s10}T_{s0.5}$) and the far speech-modulated noise masker ($M_{n10}T_{s0.5}$) suggests that the largest separation in distance measured here (10 m) fully removes informational masking and the threshold is limited by energetic masking. This follows the hypothesis of Best *et al.* (2012), arguing energetic masking limits and defines the maximum SRM that can be observed in a given experiment.

To further investigate the involvement of confusions between target and masker, Tab. 2.3 shows the percentage of masker errors for the measured CRM corpus. A masker error occurs when the listener response is a color/number combination belonging to one of the maskers. The percentage given here is calculated from the sum of the total number of errors (i.e. masker errors plus random errors). Masker errors are often associated with informational masking as they directly measure the amount of confusions. As expected, the colocated condition with the speech masker ($M_{s0.5}T_{s0.5}$) contains the largest amount of masker errors. This percentage decreases as the target-masker separation increases. This further highlights that spatial separation in distance results in perceptual segregation of target and masker thereby resolving confusions.

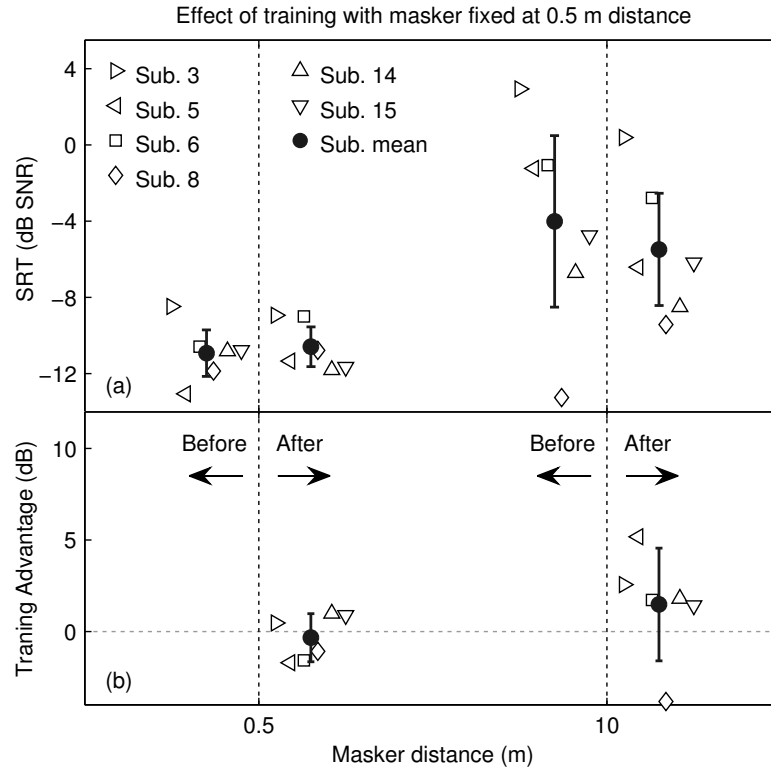
2.4.3.2 Varying target distance

In the condition where the masker was kept at 0.5 m ($M_{s0.5}T_{s10}$) and the distance of the target was changed (Figs. 2.5), a large variability across subjects was observed for the speech masker but not for the speech-modulated noise masker. With the speech masker at 0.5 m and the target at 10 m distance, some subjects even had higher SRTs than for the colocated condition. This behavior could have been caused by several factors. It is commonly acknowledged that strong reverberation leads to decreased intelligibility (Nábělek and Robinson, 1982). Hence, the increased reverberation with the target at 10 m could have caused the increase in SRTs. However, if this was true, replacing the speech masker by a speech-modulated noise masker ($M_{n0.5}T_{s10}$) should also result in an increased SRT, but this was not the case. Figure 2.5 shows that the SRT for the speech-modulated noise masker actually decreases slightly with increased target distance.

Alternatively, the large difference in SRTs between the speech and speech-modulated noise masker conditions for the far target and close masker may be due to informational masking. However, if informational masking is the main factor, why do listeners not benefit from the spatial separation as with the close target and far masker? Considering the clarity of the different speech signals involved, the far target is very blurred or “distorted” by the strong room reverberation, whereas the near masker is very clear or “undistorted”. Considering the masker errors in Tab. 2.3, the far target/near masker condition produces very few “confusion” errors. Hence, it may be assumed that the target-masker confusions are removed by moving the target further away than the masker, but at the same time, a different type of informational masking is introduced. It could be that the undistorted nearby maskers are highly distracting and thereby make it very hard for the listener to (selectively) suppress the masker and attend to the blurred target. In other words, the speech maskers often win the competition in attention over the target speech even if the listener knows they are not attending to the target. This effect might have been enlarged by the nature of the CRM corpus. Similar to the confusion-based informational masking, this distraction-based informational masking is removed when applying speech-modulated noise maskers. Indeed, some subjects reported that it was especially hard to ignore the nearby speech masker and to focus on the far target. The very large variation of the SRTs across listeners may further support the assumption that higher-level auditory mechanisms are involved rather than the low-level mechanisms related to

TABLE 2.3: Percentage of masker errors for the measured CRM results.

	Near target		Far target	
	Condition	masker err.	Condition	masker err.
Speech masker	$M_{s05}T_{s05}$	14.8%	$M_{s05}T_{s05}$	14.8%
	$M_{s2}T_{s05}$	11.5%	$M_{s05}T_{s2}$	8.0%
	$M_{s5}T_{s05}$	8.9%	$M_{s05}T_{s5}$	7.8%
	$M_{s10}T_{s05}$	7.0%	$M_{s05}T_{s10}$	5.7%

FIGURE 2.8: Individual SRTs for 6 subjects before and after training using the $M_{s0.5}T_{s10}$ and $M_{s10}T_{s0.5}$ conditions (upper panels) and the corresponding advantage achieved by training (lower panels).

energetic masking. A final reason for the large variability across subjects might have been that the target position in most of the experiment was at 0.5 m, which might have primed or confused subjects as they were not explicitly informed that the target changed location. Hence, some listeners might have built up a strategy to focus on the “clear” near speech rather than concentrating on the call-sign and were mislead when the masking speech was nearby and the target was far. However, this would presumably result in an increased amount of masker errors in the $M_{s05}T_{s10}$ condition, which is not the case.

2.4.3.3 The effect of listener training

To further examine the condition with the far target and near masker, an extensive training as well as modified experimental procedure were applied to retest 6 out of the original 16 subjects. The training was done in order to ensure that the listeners were not primed to the 0.5 m target and knew which of the sources to listen to. Thereby, this extra experiments is designed to further understand the results presented in Fig. 2.5. In this follow-up experiment only the $M_{s0.5}T_{s10}$ and $M_{s10}T_{s0.5}$ conditions were tested using the CRM corpus. The training consisted of a graphical representation of the test condition which visually indicated the target and masker position and the correct color/number combination. Sentences were taken from the CRM corpus and were always presented with the masker at 55 dB SPL and a fixed SNR of -3 dB. Subjects could switch between the conditions and took 10 minutes to familiarize themselves with the stimuli and relate those to the graphical representation. Afterwards, the two conditions were randomly presented with two repetitions following the methods described in Sec. 2.2. During the test the same visual representation from the training was used to indicate the location of the target and masker. Mean individual results before and after training are shown in Fig. 2.8 together with the across subject mean and related 95% confidence intervals. A paired t-test showed no significant difference between the pre- and post-training SRTs neither for the $M_{s10}T_{s0.5}$ condition ($p = 0.54$) nor the $M_{s0.5}T_{s10}$ condition ($p = 0.27$). However, the effect of training is significant if SRTs from the outlier subject 8 (diamonds) were removed ($p < 0.05$). This indicates that the spread of thresholds shown in Fig. 2.5 for the $M_{s0.5}T_{s10}$ condition could be reduced but not removed by training. However, in general the effect of priming subjects to the nearby (0.5 m) target position throughout most of the main experiment cannot explain the spread of thresholds in the spatially-separated far target conditions shown in Fig. 2.5. Hence, distraction-based informational masking seems to play the main role when faced with nearby maskers while trying to attend to a target that is further away.

2.5 Summary and conclusion

The study investigated the improvement in speech intelligibility that can be measured when a target talker is separated in distance from a masking source. BRIRs were applied to spatialize the sound sources and to vary their distance. The long-term spectra as well

as the RMS levels of the reverberant signals were equalized leaving differences in the DRR as the primary cue to differentiate target and masker. A SRM of up to 10 dB was measured with the CRM speech corpus, confirmed with the LISN-S speech corpus and shown to be independent of the applied normalization and equalization procedures. The improvement in intelligibility was particularly prominent when the target was kept close and the masker was moved further away. However, distance dependent SRM was only observed when the masker was realized by competing speech and no SRM was found with a speech-modulated noise masker. Moreover, when the applied dichotic stimulus presentation was replaced by a diotic presentation only a small reduction (< 3 dB) in SRM was observed, indicating that the distance-related SRM is mainly a monaural effect.

To better understand the auditory mechanisms that are potentially involved in the observed SRM effect, different objective signal-based measures were applied. Frequency-based measures such as the long-term SNR, segmental intelligibility-weighted SNR and cross-ear glimpsing all failed to explain the observed effect. However, the SNR in the modulation domain was found to correlate well with the measured SRT data, at least, when the target is close and the masker is varied in distance. Since the SRM was not observed with a speech-modulated noise masker, it was suggested that in addition to a change in the modulation-based SNR, differences in distance resolve speech masker confusions and thus reduce informational masking.

In the condition where a speech masker was close while the target talker was far away the SRM varied strongly across subjects, some showing a large SRM while others even showing a negative SRM. Since the masker confusions were rather low in this condition, it was argued that this effect could not simply be explained by confusion-based informational masking. It was hypothesized that a different type of informational masking was involved, in which the rather anechoic (clear) masker captures the attention of the listener and makes it hard to attend to the highly reverberant (blurred) target signal.

2.6 Acknowledgments

The authors would like to thank Dr. S. Jørgensen (Centre for Applied Hearing Research, Technical University of Denmark, Denmark) for his contribution to the envelope domain

analysis. This work was supported by an International Macquarie University Research Excellence Scholarship (iMQRES) and Widex A/S.

Chapter 3

The effect of a hearing impairment on source-distance dependent speech intelligibility in rooms¹

Westermann and Buchholz (2014a) found substantial improvements in speech reception thresholds (SRTs) for normal hearing listeners in a simulated auditorium when the target was separated in distance from a two-talker masker. This study applied similar methods, but tested hearing impaired (HI) listeners instead. The HI listeners received a 7 dB benefit when the target was fixed at 0.5 m and the masker was moved from 0.5 m to 10 m. But when the target was moved away the SRTs increased by 5 dB. This indicates that different to NH listeners, HI listeners have difficulties suppressing nearby maskers while focusing on a far target.

¹Aspects presented at the Meetings on Acoustics (2013). Chapter represents a manuscript to be submitted as an Express Letter to The Journal of the Acoustical Society of America.

3.1 Introduction

The auditory system employs different mechanisms to successfully understand speech in reverberant multi-talker environments. These auditory mechanisms are often disturbed in hearing impaired subjects (HI), which makes it hard (or even impossible) for them to communicate in such challenging "cocktail party scenarios" (e.g., Bronkhorst, 2000). Numerous studies have shown how (NH) listeners, and to some degree also HI listeners, can take advantage of the angular separation as well as the voice characteristics of the individual talkers (e.g., Brungart *et al.*, 2001). Moreover, it has been shown that in a reverberant environment early reflections, which arrive at the listener within about 50 ms after the direct sound, can further support intelligibility (e.g., Bradley *et al.*, 2003). Recently, Westermann and Buchholz (2014a) showed how NH listeners can effectively use distance-related cues, especially those related to changes in the direct-to-reverberant ratio (DRR), to better understand a target talker in a background of masking talkers (for an overview on auditory distance perception see Zahorik *et al.*, 2005). Using binaural room impulse responses (BRIRs) measured in a reverberant auditorium they presented a sentence test with the target and masker at different distances directly in front of the listener. Thereby, to focus on reverberation cues, distance-dependent level and spectral changes were equalized. They investigated both a scenario where the target was kept close (0.5 m) and the masker distance varied from 0.5 m to 10 m, and a scenario where the masker was kept close and the target distance varied. Measuring speech reception thresholds (SRTs), they found intelligibility improvements of up to 10 dB when the target was at 0.5 m distance and the masker was changed from 0.5 m to 10 m. When the masker was kept close and the target was moved away the mean SRT still improved, but the individual SRTs varied largely. Some listeners received a substantial benefit from the spatial separation whereas other listeners performed even slightly worse than in the colocated condition.

To better understand this observation, Westermann and Buchholz (2014a) applied different objective signal-based measures. They found that the improved intelligibility with target-masker separation could neither be explained by long-term (Greenberg *et al.*, 1993) or short-term signal to noise ratio (SNR) improvements (Hansen and Pellom, 1998) nor cross-ear glimpsing (Brungart, 2012; Glyde *et al.*, 2013). However, the improvement when maskers are moved further away than the target could at least be qualitatively

described by the SNR in the modulation domain (Jørgensen and Dau, 2011). The large variability in the SRT across subjects for the close-target and far-masker condition, and in particular, the reduced performance seen in some subjects, could not be explained by any of the applied objective measures. They hypothesized that informational masking (IM) might explain this behavior, which was supported by the observation that all subjects were able to receive a substantial benefit when the speech masker was replaced by a purely energetic, speech modulated noise masker. However, analyzing the masker errors for the speech masker the variability was not due to target-masker confusions as commonly observed in the IM dominated, colocated conditions (Ihfeldt and Shinn-Cunningham, 2008). They hypothesized that the nearby “clear” masker captured the attention of the listener over the “blurred” reverberant target and named this “distraction-based” IM, as in contrast to “confusion-based” IM. This separation into (at least) two types or aspects of IM is in agreement with discussions, for example, provided by Kidd *et al.* (2007).

Overall, the study by Westermann and Buchholz (2014a) indicated that the NH auditory system can utilize reverberation cues, as provided by differences in distance between sources to better understand speech in multi-talker reverberant environments. However, since previous studies have shown that HI listeners have severe deficits in utilizing reverberation cues for distance perception (Akeroyd *et al.*, 2007) it is important to investigate if HI listeners gain the same benefit as NH listeners when the distance between target and masker is varied. Thereby, it is of particular interest what the effect is of a hearing impairment on distraction-based IM when the target is further away than the masker.

3.2 Methods

3.2.1 Stimuli

As in Westermann and Buchholz (2014a), two speech corpora were used in this experiment: the coordinate response measure (CRM; Bolia *et al.*, 2000) corpus and the speech material of the Listening in Spatialized Noise-Sentences test (LISN-S) (Cameron and Dillon, 2007). In the CRM corpus each sentence has the structure: “Ready [call sign] go to [color] [number] now”, with eight call-signs, four colors (red, green, blue and white) and eight numbers (1 through 8), resulting in 256 sentences for each of eight different talkers. SRTs were measured using only the four male talkers. Subjects were

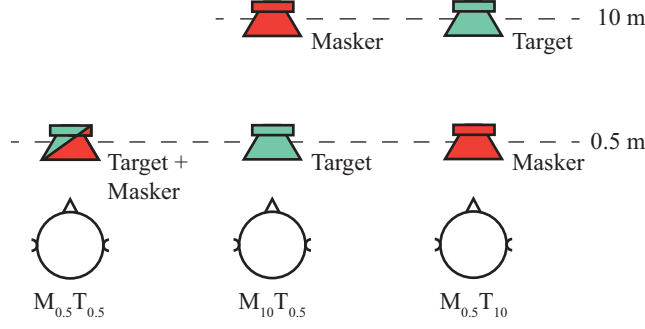


FIGURE 3.1: Schematic showing the spatial conditions depicting the target and masker distance as well as condition labels.

assigned the “Baron” call-sign, and asked to report the color/number corresponding to that speaker. Two different maskers were applied, a speech masker and a speech modulated noise masker. The speech masker consisted of two randomly chosen CRM sentences with different talker and color/number combination from the target. The speech modulated noise masker was realized by applying the Hilbert envelope of each of the speech maskers to noise with the same long-term spectrum as all of the speech maskers (for details see Best *et al.*, 2013b).

To allow conclusions with a more general validity, the target sentences and maskers from the LISN-S were also tested. With the LISN-S, SRTs are measured with a continuous two-talker masker (Cameron and Dillon, 2007). Both the target and masker talkers are female. The masker is either the same or a different-talker from the target. Here, only the different-talker masker was used. Since the LISN-S corpus only allows testing of four conditions without repeating sentences, only some of the conditions measured with the CRM corpus could be measured with the LISN-S.

All (anechoic) target and masker signals were convolved with binaural room impulse responses (BRIRs) measured at different distances in an auditorium using a B&K Head and Torso Simulator (HATS). The auditorium had a reverberation time of $T_{30} = 1.9$ s at 2 kHz and a volume of approximately 1150 m^3 . These were the same BRIRs used in Westermann and Buchholz (2014a). Three different spatial configurations were tested here as illustrated in Fig. 3.1. Note that the labels of the spatial conditions show the applied distances as well as the masker type, i.e. M_{s10} is a speech masker at 10 m distance and $M_{n0.5}$ is a (speech modulated) noise masker at 0.5 m distance.

To maintain the time-alignment of target and masker, which is only critical for the CRM corpus, the propagation delay introduced from the distance between sound sources

was removed by time-aligning the direct sound component of the measured BRIRs. Furthermore, in order to minimize intelligibility improvements directly resulting from distance-dependent changes in long-term spectrum and overall level the maskers were equalized. The equalization was designed so that the equalized long-term spectra of the masker always equaled the long-term spectrum of the masker colocated with the target (i.e. either $M_{s0.5}M_{s0.5}$ or $M_{s10}M_{s10}$). Finite Impulse Response (FIR) equalization filters with a length of 512 taps (at a sampling frequency of 44.1 kHz) were designed and applied using MATLAB. The equalization procedure was applied to both the CRM and LISN-S speech corpora.

3.2.2 Procedures

Experiments were carried out in a double-walled booth, using equalized, Sennheiser HD-215 circumaural headphones driven by a RME Hammerfall HDSPe AIO sound-card and a computer with a MATLAB GUI. Preceding each experiment, both air and bone conduction audiometric thresholds were measured in octave bands from 250 Hz to 8000 Hz. For both the CRM and LISN-S the masker level was kept at a root mean square (RMS) level of 60 dB SPL, measured in a B&K type 4153 artificial ear before compensation for hearing loss. The target level was initially set to 67 dB SPL and varied relative to the masker following a one-up one-down rule to adaptively estimating the SRT. In order to (partly) compensate for audibility, linear amplification was applied according to the NAL-RP scheme (Dillon, 2001). The individually prescribed insertion gains were realized using 512-tap long FIR filters designed in Matlab and were applied to the stimuli before presentation via headphones.

First the LISN-S test was measured and then the CRM. Within each test the order of presentation was randomized and all conditions measured with the CRM corpus were repeated once. Testing each subject required a single session of 1.5 hours. After the initial hearing screening and audiometry, the subjects were given verbal instruction read by the experimenter. Before the CRM experiment started, training was performed using one random condition to ensure familiarity with the GUI and understanding of the task. No training was applied before the LISN-S test.

3.2.3 Subjects

Nine HI subjects (three females and six males) participated aged 44-77 years (mean 67). All subjects had symmetrical sloping sensorineural hearing losses, and all were native Australian English speakers and experienced hearing aid users. The individual audiograms, as well as their mean value are shown in Fig. 3.2. All subjects were active participants from the National Acoustic Laboratories database, and had significant experience with speech intelligibility tests. In order to allow a direct comparison between the derived HI data and NH data, results from 16 normal hearing (NH) listeners (< 20 HL) were taken from Westermann and Buchholz, 2014a.

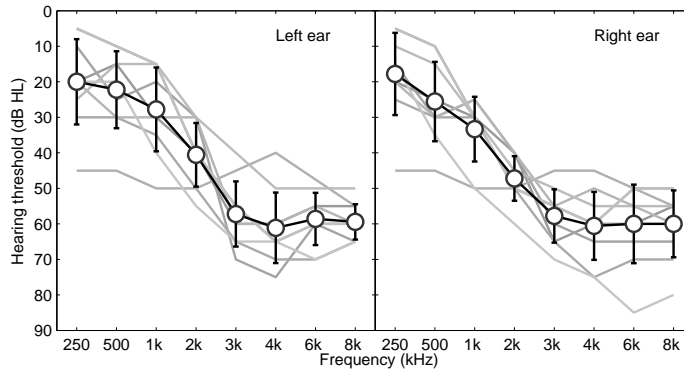


FIGURE 3.2: Mean and standard deviation of hearing thresholds of the nine subjects.

3.3 Results

The left and right panels of Fig. 3.3 show the SRTs measured for the NH and HI listeners, respectively, using the CRM speech corpus. The filled black symbols denote the speech masker and the open symbols the speech-modulated noise masker. The lower panels show the corresponding spatial advantage, calculated as the difference between the individual per subject SRT in the colocated condition and the individual SRT per subject in the separated condition. When moving the speech masker from 0.5 m (i.e., colocated condition $M_{s0.5}T_{s0.5}$) to 10 m (i.e., spatially separated condition $M_{s10}T_{s0.5}$) the SRT decreased on average by about 7 dB for the HI listeners, i.e. the listener's performance is strongly improved. However, this improvement is smaller than the average improvement of 10 dB observed with the NH listeners. For the speech-modulated noise masker, the SRT in the colocated condition decreased to the same value as in the far-masker ($M_{n10}T_{s0.5}$) condition and thus, no spatial advantage was observed. This was the same

for NH and HI subjects. When the masker was kept at 0.5 m and the target distance was increased from 0.5 m to 10 m (i.e. from $M_{s0.5}T_{s0.5}$ to $M_{s0.5}T_{s10}$) the mean SRT increased (intelligibility decreased) by 5 dB for the HI listeners. This decrease in performance is in qualitative agreement with some of the NH subjects, but a significant number of NH subjects still showed a clear improvement as illustrated by the large spread of the NH data (and discussed in Westermann and Buchholz, 2014a).

Because of the unbalanced data ($M_{n0.5}T_{s10}$ was not measured for the HI listeners), the statistical analysis was split into two components. Firstly, a three-way repeated measures analysis of variance (ANOVA) was performed on the conditions with colocated presentation and conditions with the target at 0.5 m and masker at 10 m, with masker type and spatial condition as within-subject variables and hearing loss (NH or HI) as a categorical between-subject variable. Significance were found for all three factors: masker type [$F(1, 23) = 232, p < 0.001$], spatial condition [$F(1, 23) = 443, p < 0.001$] and hearing loss [$F(1, 23) = 1658, p < 0.001$]. Additionally, significant interactions were found between masker type and spatial condition [$F(1, 23) = 465, p < 0.001$] and between spatial condition and hearing loss [$F(1, 23) = 5, p < 0.05$], but not between masker type and hearing loss [$F(1, 23) = 0, p = 0.9$]. Finally, there was a three way significant interaction between all of the factors [$F = 14, p < 0.01$]. In a second statistical analysis, a one-way repeated measures ANOVA was performed on the HI data in conditions with the speech masker thereby including $M_{s0.5}T_{s10}$ and applying spatial condition as the within-subject factor. A strong significance was found for the spatial condition [$F(1.2, 9.7) = 115, p < 0.001$]. Here, the Greenhouse-Geisser correction was applied because a significance effect in Mauchly's test of sphericity.

Figure 3.4 shows the measured SRTs and the corresponding spatial advantage using the LISN-S corpus. Similar to the CRM data for the HI listeners as well as to the NH data in the LISN-S, SRTs decreased by approximately 5 dB when the speech masker was moved to 10 m (i.e. from $M_{s0.5}T_{s0.5}$ to $M_{s10}T_{s0.5}$) and SRTs increased by about 4 dB when the target was moved to 10 m distance ($M_{s0.5}T_{s10}$).

Overall, the HI data shows the same tendencies as the NH data when varying masker distance for both speech corpora, but the NH perform better in all conditions, especially when the target is further away than the masker ($M_{s0.5}T_{s10}$).

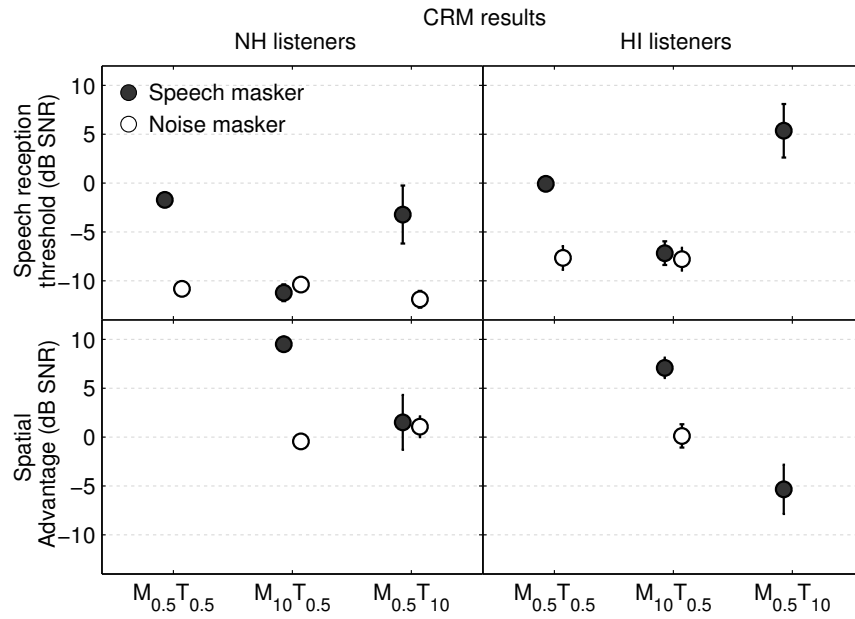


FIGURE 3.3: Top panels: Mean and across-subject 95% confidence intervals of SRTs measured with the CRM corpus. Bottom panels: Mean and 95% confidence intervals of the spatial benefit.

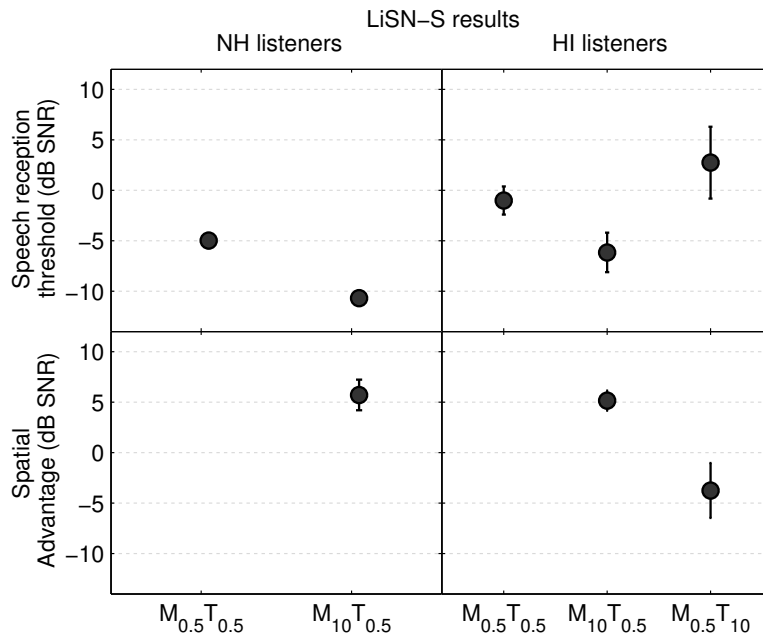


FIGURE 3.4: As Fig. 3.3, but measured with the LISN-S corpus.

3.4 Discussion and conclusions

The present study investigated the effect of distance-related reverberation cues on speech intelligibility in HI listeners. Generally, both the CRM and LISN-S results show that increasing the distance of a masker results in an improvement of mean SRTs of about

TABLE 3.1: Percentage of masker errors for the measured CRM results.

	Near target			Far target		
	Condition	HI	NH	Condition	HI	NH
Speech masker	$M_{s05}T_{s05}$	13.2%	14.8%	$M_{s05}T_{s05}$	13.2%	14.8%
	$M_{s10}T_{s05}$	7.5%	7.0%	$M_{s05}T_{s10}$	3.9%	5.7%
Noise masker	$M_{n05}T_{s05}$	2.0%	0.2%	$M_{n05}T_{s05}$	2.0%	0.2%
	$M_{n10}T_{s05}$	2.0%	0.5%	$M_{n05}T_{s10}$	0.5%	-

5-7 dB, whereas the speech modulated noise maskers were unaffected by the spatial separation. This difference can be explained by considering the concepts of energetic masking (EM) and IM (for a review see Kidd *et al.*, 2007). Whereas speech maskers in the colocated condition create substantial target-masker confusions, and thus involve IM (Freyman *et al.*, 2001), the noise masker does not involve any target-masker confusions. This is confirmed in Tab. 3.1, where in the colocated condition 13.2% of all errors for the speech-masker are target-masker confusions (or masker errors; Ihlefeld and Shinn-Cunningham, 2008), but only 2.2% are masker errors for the speech-modulated noise masker.

The observation that the SRTs for the spatially separated condition ($M_{s10}T_{s0.5}$) are equal for the speech masker and the speech-modulated noise masker suggests that also for HI listeners the spatial separation aids the perceptual segregation of target and masker and thereby removes target-masker confusions, and thus, removes IM. Again this is supported by the data in Tab 3.1, where masker errors are significantly reduced (halved) to 7.5% when the speech masker is moved further away. Considering the case when target and masking talkers are presented from different directions, Best *et al.* (2013b) argued that SRTs are limited by EM. In the same way it may be argued here that the decrease in SRTs due to the spatial separation in distance is also limited by EM, which is approximated by the SRT of the speech-modulated noise masker.

Comparing the data for the NH and HI listeners illustrates that all SRTs are increased for the HI subjects, which is more pronounced in the spatially separated conditions. As a consequence, the spatial advantage achieved by moving the target away from the speech-masker is reduced in HI subjects. The increase in colocated thresholds is usually explained by decreased sensitivity to loudness cues, which typically provides the main cue for segregating the target from the speech masker in this highly IM dominated condition (Brungart *et al.*, 2001). The increase in SRTs in the spatially separated condition may

be explained by increased EM (see Best *et al.*, 2013b) due to reduced target audibility as well as reduced temporal and spectral resolution, and maybe distorted spatial cues. This is supported by the observation that the SRTs for the purely energetic, speech-modulated noise masker are increased by the same amount as the SRTs for the speech masker.

When the speech masker was kept at 0.5 m and the target was moved further away ($M_{s0.5}T_{s10}$), SRTs significantly increased for all HI subjects. This was different to the NH group, which also showed a large inter-subject variability, but with some subject receiving a large benefit from moving the target away and others showing a small detrimental effect. Since in this spatial condition a strong and consistent benefit was observed for the speech-modulated noise masker for both the NH subjects, the subject-dependent behavior observed for the speech masker is most likely linked to IM effects. However, unfortunately this condition was not measured with the HI listeners.

One reason for the behavior condition $M_{s0.5}T_{s10}$, could be that the near-by masker is confused with the target, but this would lead to an increased amount of masker errors, which is not the case. According to Tab. 3.1, this condition provides the lowest number of masker errors for all speech-masker conditions (i.e., 3.9% for HI and 5.7% for NH subjects). Hence “confusion-based” IM is most likely not involved. Westermann and Buchholz (2014a) argued that the close and “clear” masker in this condition distracts the attention of the listeners from the far and “blurred” target, causing so-called “distraction-based” IM. If this the case, then the results would indicate that the ability to selectively attend to the target, and thereby to suppress the distractors, is highly subject dependent and largely reduced in HI subjects. This might be linked to cognitive factors as well as auditory factors, which due to the hearing loss as well as the increased age, may be both reduced in HI subjects. Therefore, future studies should consider if in particular cognitive factors or abilities (such as the executive function) can explain the large differences between subjects for the far-target and close-masker condition with speech maskers.

3.5 Acknowledgments

This work was funded by an International Macquarie University Research Excellence Scholarship (iMQRES) and Widex A/S.

Chapter 4

The influence of informational masking in reverberant, multi-talker environments¹

The relevance of informational masking (IM) in real-world listening is not well understood. In literature, IM effects of up to 10 dB in measured speech reception thresholds (SRTs) are reported. However, these experiments typically employed simplified spatial configurations and speech corpora that magnified confusions. In this study, SRTs were measured with normal hearing subjects in a simulated cafeteria environment. The environment was reproduced by a 41-channel 3D-loudspeaker array. The target talker was 2 m in front of the listener and masking talkers were either spread throughout the room or colocated with the target. Three types of maskers were realized: one with the same talker as the target (maximum IM), one with talkers different from the target, and one with unintelligible, noise-vocoded talkers (minimal IM). Overall, SRTs improved for the spatially distributed conditions compared to the colocated conditions. Within the spatially distributed conditions, there was no significant difference between thresholds with the different-talker and vocoded-talker masker. Conditions with the same-talker masker were the only conditions with substantially higher thresholds, especially in the colocated conditions.

¹Manuscript submitted for publication to The Journal of the Acoustical Society of America.

4.1 Introduction

Many studies have attempted to uncover the factors involved when listening in reverberant multi-talker environments, often labeled as the “cocktail party effect” (Bronkhorst, 2000; Cherry, 1953). The cocktail party is a highly complex problem covering acoustical phenomena, auditory masking, attention, binaural processing and spatial processing. Masking is often divided between *energetic masking* (EM) and *informational masking* (IM) (Brungart *et al.*, 2001; Freyman *et al.*, 1999; Watson, 2005). Historically, the need for such a segregation arose from studies like Carhart *et al.* (1969), where higher intelligibility thresholds were reported in presence of a speech-masker in comparison to a noise-masker. Usually, EM is defined as masking which degrades the peripheral representation of the target signal (Cherry, 1953; Kidd *et al.*, 2007). Various auditory models have been successfully applied to account of the effect of EM (Dau *et al.*, 1996; Durlach *et al.*, 1986). On the other hand, there are no comprehensive frameworks explaining how the peripheral representation of a mixture with multiple sound sources is successfully, or unsuccessfully, converted to a perception of discrete auditory objects and which account for how attention is steered to listen to only one specific source. This has lead to IM, which describes failures in such domains, being loosely defined as everything that can not be accounted for by EM. To abrogate such vague explanations, several authors have proposed definitions of IM (Durlach *et al.*, 2003; Shinn-Cunningham, 2008; Watson, 2005). Watson (2005) specifically splits IM between effects attributed to uncertainty and similarity. Uncertainty is caused by the listener not knowing where to listen and is often linked to experiments with tone-complexes (e.g. Watson *et al.*, 1976). IM due to similarities between target and masker is caused by failures to segregate the target and masker and is often associated with speech-on-speech masking (e.g. Carhart *et al.*, 1969). The theory of auditory scene analysis (ASA) (Bregman, 1994), in which the auditory system segments and integrates auditory elements into basic objects from which streams are segregated and selected, was related to IM in Shinn-Cunningham (2008). Mainly, she introduces a conceptual model between bottom-up salience and top-down attention and argues that IM is due to failures in either auditory object formation or object selection. Failures in object formation are caused by target-masker similarities hindering basic bottom-up grouping and streaming. Failures in object selection are caused by both similarity and uncertainty, where similarities can interfere with the correct selection of properly segregated streams and uncertainty can either inhibit direction

of top-down attention or draw exogenous attention (e.g. a person saying your name). This study leans towards Shinn-Cunningham’s work dealing with IM which presumably occurs because of (1) confusions due to the presence of maskers and target similarities, or (2) distractions due to the presence of maskers that compete with and capture the exogenous attention of the listener.

In a reverberant multi-talker environment, sound sources will often be spread out in a room. Many studies have shown that speech intelligibility increases when a target and masking talker are spatially separated rather than colocated (i.e. on top of each other), resulting in a spatial advantage. This advantage, or spatial release from masking (SRM), can result in differences in speech reception thresholds (SRTs) of up to 20 dB and is observed for changes in direction (Bronkhorst, 2000; Freyman *et al.*, 1999), distance (Shinn-Cunningham *et al.*, 2001; Westermann and Buchholz, 2014a), and head orientation of the masking talkers (Strelcyk *et al.*, 2014). The degree of masking release is influenced by factors such as the spatial configuration, the number of interfering maskers and the applied speech corpus (Bronkhorst, 2000; Brungart *et al.*, 2001). To explain SRM, the concepts of EM and IM are often applied (e.g., Freyman *et al.*, 1999; Glyde *et al.*, 2013). Whereas the release from EM may be linked to binaural auditory mechanisms that provides an improvement in “effective” signal-to-noise ratio (SNR), such as better-ear glimpsing or equalization-cancellation (Durlach, 1963), release from IM is linked to a perceptual segregation of target and masker signals (e.g., Freyman *et al.*, 1999), which does not necessarily involve an SNR advantage. Such a perceptual segregation could facilitate both the auditory object formation and selection stage in Shinn-Cunningham (2008)’s framework. However, in relation to real-life listening, the colocated reference condition, with its “spatial similarities”, is unnatural so while the reported binaural mechanisms are likely important for everyday listening it is hard to comment on their quantitative effect.

In addition to the colocated reference condition, a majority of SRM studies use speech corpora which result in a substantial amount of IM, such as the Coordinate Response Measure (CRM, e.g. Best *et al.*, 2013b; Bolia *et al.*, 2000; Brungart *et al.*, 2001) or corpora with the same target and masking speaker (e.g. same-voice condition in Listening in Spatialized Noise-Sentences test; Cameron and Dillon, 2007). Generally for these speech corpora, is an unrealistic amount of confusions, from either similarity between

talkers or context which are expected to inhibit auditory object formation and selection. Together, the spatial and speech corpora (both talker and contextual) similarities provides substantial amounts of EM and IM that are (somewhat) resolved by spatial separation causing the large SRM effect reported in many studies. This leaves the obvious question about the strength and importance of SRM in real-life listening. While the influence and consequence of reduced EM from spatial separation might be clear, the effect of IM is hard to quantify. One can only wonder how much IM is left when masking talkers have different voices, are located in different places and room acoustics provide additional cues for speaker segregation such as distance (direct-to-reverberation ratio), coloration and changes in speech modulations.

Many studies have tried to separate EM and IM by measuring the effect of each individually. This requires reference maskers which result in only one type of masking. Arbogast *et al.* (2002) and later Ihlefeld and Shinn-Cunningham (2008) quantified IM in the CRM corpus by eliminating spectral overlap between target and masker thereby removing EM. The former study showed that spatial separation improved thresholds for an IM-only masker by 18 dB whereas the EM-only masker improved by 7 dB. In other studies, EM has been isolated and IM removed by reversing the masking speech (Freyman *et al.*, 2001), mixing genders (Brungart *et al.*, 2001) or constructing unintelligible noise with similar peripheral masking characteristics (Best *et al.*, 2013b; Brungart *et al.*, 2001). Similarly to the goals of this study, Culling (2013) investigated IM in an environment with multiple talkers in a simulated reverberant environment over headphones. He measured SRTs for a varying amount of either same-talker, different-talker or speech-shaped noise maskers and found only very little differences between the same-talker and different-talker masker. However, the speech-shaped noise maskers were stationary and did therefore not try to capture the EM contribution of the different-talker masker, but rather was compared to babble conditions. As a result, it is difficult to segregate if the reported differences in SRTs between noise and speech maskers are due to dip-listening or IM.

Generally, the effectiveness of these different types of EM-reference maskers can be discussed. Firstly, the EM, or peripheral representation, is often different between the EM-reference masker and mixed EM and IM condition, e.g. from different temporal behavior from speech reversal or from broadband spectral smoothing with speech-modulated noise maskers. Secondly, IM caused by failures in bottom-up formation of low-level auditory

objects is still likely to be present. Furthermore, as it seems impossible to address each and every aspect of IM discussed above in an EM-reference masker, it is (as always) pertinent to have better definitions of IM, so that scientist doing studies in the area can clearly declare the nature of IM they are studying, as it has been done for years in the field of EM. In the end, the EM-reference maskers will only be an approximation of the EM in the original signal and this should be recognized when considering outcomes of such studies.

To better understand the involvement and relevance of IM in the reverberant multi-talker environment, the current study implemented a speech intelligibility test in a simulated cafeteria where ongoing background conversations masked the target. The cafeteria was simulated using a room-acoustical model and auralized using a three-dimensional loudspeaker array. A number of conditions were designed to investigate (1) the strength of IM in a realistic cafeteria setting by comparing intelligibility scores between a speech masker and an unintelligible vocoded masker, (2) the effect of spatially distributing sound sources throughout the room as opposed to presenting them in the same location, i.e. taking into account the distance, direction, and head-orientation of the masking talkers, (3) the effect of talker similarity, i.e., considering masking talkers with the same and different voices to the target talker, and (4) the upper and lower SNR boundaries of the encountered IM effects.

As IM is the central topic of this study, it is pertinent to clarify the nature of the IM that is considered here. Since the isolated EM condition is realized by unintelligible vocoded speech, this study mainly examines confusion-based IM. In other words, IM related to target-masker similarities which causes failures in streaming and object selection. Hence, this study follows the lines of literature on IM encountered in SRM studies such as Brungart *et al.* (2001) and Best *et al.* (2013b). In turn this also means that IM which interferes with basic auditory object formation is still present in the EM control condition and IM caused by failures in exogenous attention is not considered.

4.2 Methods

4.2.1 Subjects

In this study seventeen (thirteen female, four male) subjects with Australian English as their first language participated, all with normal hearing (thresholds ≤ 20 dB hearing loss at audiometric frequencies from 250 Hz to 8000 kHz). Mean age was 30 years (ages from 18 to 42 years), and all subjects reported normal cognitive function. Subjects were either employed at the National Acoustic Laboratories, or they were students at Macquarie University. Those that were not employed at the National Acoustic Laboratories were given a gratuity for their participation. All subjects gave written consent before participating in the study.

4.2.2 Stimuli

A sentence test was implemented using the Bamford-Kowal-Bench (BKB) sentence material (Bench *et al.*, 1979). The corpus contains 336 sentences, organized in 21 lists, spoken by a native Australian-English male speaker and sampled at 44.1 kHz. The sentences have a simple syntactical structure (e.g. "The angry man shouted,") and an average length of about 1.5 s. The original BKB material is filtered so that its long-term spectra matches the "universal" long-term average speech spectrum (LTASS) defined by Byrne *et al.* (1994). However, this filtering made the sentences sound unnatural when presented inside the simulated cafeteria environment (Sec. 4.2.3). Therefore, the unfiltered long-term spectrum of the monologues used for the same-talker speech masker (described below) recorded with the original BKB talker were used to construct a 512-tap inverse finite impulse response (FIR) filter. This filter was then applied to all of the BKB sentences. After filtering, the sentences sounded more natural and cohesive with the cafeteria background.

The BKB target sentences were used to measure speech intelligibility in 12 different masker conditions. These masker conditions were realized by three versions of the four spatial configurations shown in Fig. 4.2. The four spatial configurations were all realized in a simulated cafeteria environment using either two or seven two-talker dialogues in either a colocated or a spatially separated configuration. The three masker versions differed in the way the individual talker signals were generated:

- i. **Different-talker speech masker** - The different two-talker dialogues were realized using anechoic recordings of seven scripted dialogues taken from published examinations of the International English Language Testing System (IELTS). The recordings were made in the anechoic chamber of the National Acoustic Laboratories, were about 5 minutes long and were spoken by eight female and six male talkers. The recordings were post-processed so that root mean square (RMS) levels were equal during speech segments, following the procedure outlined in IEC 60268-16 (2011).
- ii. **Vocoded-talker masker** - In order to create a EM-reference masker, a noise vocoder was implemented and applied to the different anechoic recordings used in the different-talker speech masker described above. The aim of the vocoder was to make speech unintelligible while maintaining the EM components of the different-talker dialogues over time and frequency. In addition, the vocoding process was designed so that it would not destroy localization cues, namely interaural time and level differences (ITDs and ILDs), and therefore, the spatial percept of discrete sources in a room would be maintained. In order to accomplish this, the short-time Fourier transform (20 ms windows and 75% overlap) was used to convert each of the anechoic speech maskers described above to the time-frequency domain. The individual time-frequency representations were then spectrally smoothed across rectangular windows with a width of either one octave for the seven-dialogue condition or two octaves for the two-dialogue condition. The additional smoothing in the two-dialogue condition was applied to ensure that the combined masker signal was unintelligible even with fewer concurrent talkers. By using very short temporal-windows and broad frequency smearing, the transients and inherent level fluctuations were preserved as well as possible. The smoothed magnitude spectrum of the individual talkers was combined with the phase-spectrum from white noise, and the vocoded signal was reconstructed using the inverse short-time Fourier transform. In order to ensure that the EM content was similar to that of the different-talker masker, a spectral matching filter was applied. The filter was implemented as a 512-tap FIR filter using the critical band smoothed spectrum of the different-talker and vocoded-talker masker measured with a Brüel & Kjær (B&K) 4134 condenser microphone in the center of the loudspeaker array.

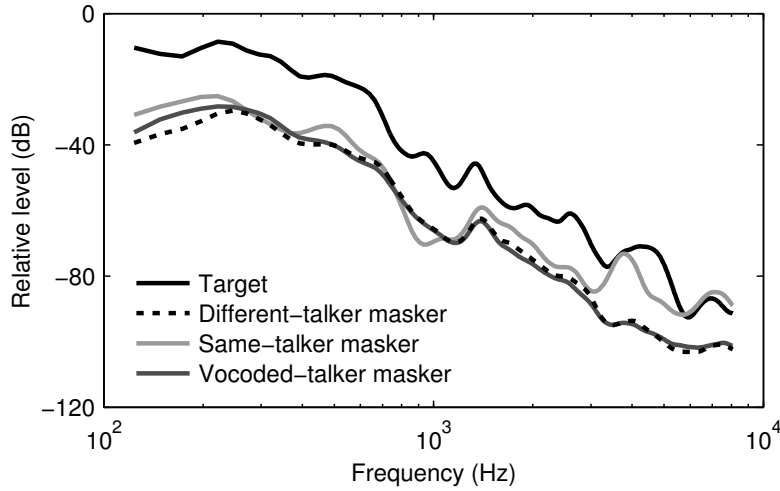


FIGURE 4.1: Long-term spectra in critical bands each of the three applied speech maskers and the target sentences.

- iii. **Same-talker speech masker** - To maximize the effect of IM (or target-masker confusions), a same-talker condition was implemented. In this case all the different two-talker dialogues shown in Fig. 4.2 were realized by monologues recorded with the same talker as used for the BKB target sentences. The monologues were based on scripts taken from published IELTS examinations. To create two-talker dialogues, each of the monologues was segmented to form dialogues with the same approximate temporal pattern as the different-talker dialogues. The segmentation was done by hand to ensure that the breaks did not occur in the middle of words. Compared to the different-talker dialogues, the created same-talker dialogues did not contain a clear semantical stream, i.e. the dialogues did not make sense. However, since participants were not continuously following the background dialogues the lack of semantical validity was not expected to affect target intelligibility.

Figure 4.1 shows the long-term spectra in critical bands of the three different types of seven dialogue maskers together with the BKB-sentence material measured with a B&K 4134 condenser microphone in the center of the loudspeaker array. Note that the different-talker and vocoded-talker masker have very similar spectra because of the spectral matching. While it is hard to make quantitative approximations of EM content, the spectrum of the same-talker masker does appear to be marginally more similar to the target sentences than the other maskers.

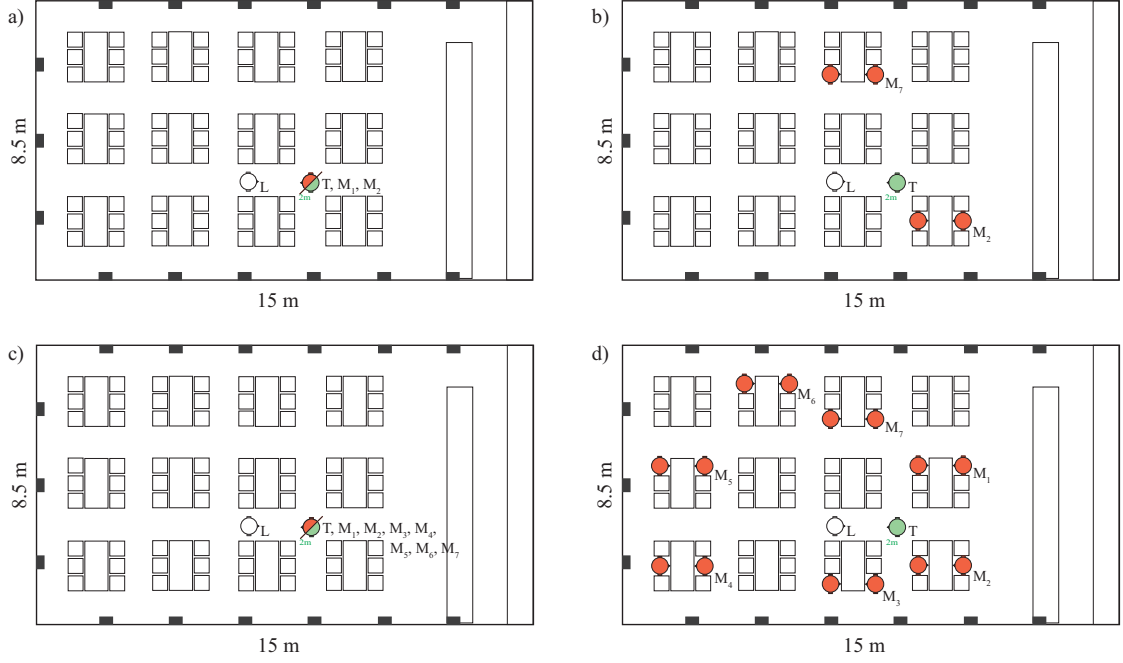


FIGURE 4.2: Top-down view of cafeteria simulated in ODEON. The listener position (L) faces the target (T) at 2 m distance. The two-talker maskers were either distributed in the room ($M_1 - M_7$) or colocated with the target (T).

4.2.3 Spatialization of sounds

A cafeteria scene with multiple masking talkers and a fairly long reverberation time was chosen as it represents a complex scene often encountered in real-life listening. However, environmental non-speech sounds like foot steps, moving of chairs, noise generated by cutlery and plates were not considered. The acoustic scene was created using the room simulation software ODEON (Rindel, 2000) and subsequently processed with the loudspeaker-based room auralization (LoRA) toolbox (Favrot and Buchholz, 2010). The resulting stimuli were presented to the test subjects using the 41-channel loudspeaker system available in the anechoic chamber of the National Acoustic Laboratories.

The simulated cafeteria (shown in Fig. 4.2a-d) was 15 m long by 8.5 m wide by 2.8 m high and had a reverberation time of $T_{30} \approx 0.6$ s. The model included windows, tables and chairs, and all talkers were simulated by sound sources with a directivity measured on a real talker (i.e., applying ODEON's directivity file `Tlknorm_natural.so8`). The target speaker was always placed 2 m in front of the listener (corresponding to T in Fig. 4.2a-d).

For each of the 15 source-receiver pairs (Fig. 4.2), reflectograms and decay curves were computed with ODEON. These were converted to room impulse responses (RIRs) and

auralized using the LoRA toolbox to create filters for each of the 41 loudspeaker-channels. The toolbox divides the RIRs into three parts accounting for the direct sound, specular early reflections (here up to third order) and late reverberation. The direct sound and early reflections were mapped to the nearest loudspeaker in the array. The late reverberation was realized by applying uncorrelated noise to the calculated frequency- and direction-dependent decay curves. Details can be found in Favrot and Buchholz (2010). The final target and masker signals were then spatialized by convolving the derived 41-channel filters of the corresponding talker position with the anechoic source signals described in Sec. 4.2.2. The combined multi-talker masker scenarios shown in Fig. 4.2 were then derived by simply adding the individual 41-channel masker signals.

4.2.4 Procedures

Subjects were seated at the center of a spherical loudspeaker array inside an anechoic chamber. The loudspeaker array consisted of 41 Tannoy V8 loudspeakers arranged in multiple rings covering a sphere with a radius of 1.85 m. The loudspeaker responses were individually equalized from the critical band smoothed response measured with a B&K 4134 condenser microphone placed in the center of the array by applying a 1024-tap FIR equalization filter. A height-adjustable chair ensured that the head of the subject was in the exact center of the array. The signal path originated outside the chamber with a PC running MATLAB fitted with an RME MADI sound card. This was connected to two RME M-32 D/A converters which feed eleven four-channel Yamaha XM4180 amplifiers connected to the loudspeakers through an acoustically dampened passage. In order to communicate with the operator, the subjects wore a lavalier microphone connected to the RME MADI sound card via a RME M-16 A/D converter. Additionally, a video camera was used to monitor the subjects.

The experiment was conducted in one visit lasting approximately one hour. Following an audiometric screening, the listeners were seated in the loudspeaker array, and the height of the chair was adjusted. The listeners were told to imagine being in a cafeteria environment and instructed to repeat the BKB sentence following a beep. Head movements were allowed, but subjects were instructed to sit still to ensure that they stayed in the “sweet-spot” of the reconstructed sound fields. The SNR resulting in 50% correct performance was adaptively measured using an one-up one-down staircase method (Keidser *et al.*, 2013) varying the level of the target and keeping the masker level fixed

at 65 dB(A) for all masker conditions. Sound pressure levels (SPLs) were measured and calibrated *in situ* with a B&K 2250 sound level meter with a 1/4-inch microphone placed in the center of the loudspeaker array and using an integration time of 30 seconds. The level of the target sentences was initially set to 70dB(A), which was calculated by first concatenating the entire BKB speech material and then applying the speech level calculation described in IEC 60268-16 (2011), which excludes speech pauses.

Performance was scored morphemically, i.e. for each morpheme in a given sentence. The SRTs were calculated using the algorithm presented in Keidser *et al.* (2013). This requires a minimum of 16 presentations with decreasing step sizes of 5, 2 and 1 dB. A run completed when the standard error, estimated by two times the standard deviation of the SNRs over the root of the number of presented sentences, fell below 0.8 dB or the maximum number of 32 sentences was reached (for further details see Keidser *et al.*, 2013). The BKB-sentence material contains 21 lists of 16 sentences. One list was used for training purposes and results were discarded. The masker in the training was always the different-talker speech masker (Sec. 4.2.2) using seven simultaneous dialogues. For each SRT two randomly chosen lists were combined to 32 sentences. Since the experiment measured 12 SRTs and there were only enough lists for ten conditions, two of the SRTs for each subject reused sentences that the subjects had already been exposed to. To minimize learning effects on overall mean results, the SRTs where sentences had to be reused were balanced over all conditions (i.e. each condition had the same number of SRTs where sentences were heard before). Throughout the experiment, the order of presentation and list/masker combination was randomized. No feedback was provided during testing.

4.3 Results

The mean SRTs and corresponding 95% confidence intervals measured when applying the two-dialogue and seven-dialogue maskers are shown in the top panel of Fig. 4.3 and Fig. 4.4, respectively. The results are grouped based on the spatial configuration, i.e. colocated or spatially separated. The three masker types containing either different-talker, vocoded-talker or same-talker maskers are marked by the circles, squares and diamonds, respectively. The difference between the colocated SRTs and spatially distributed SRTs, or spatial advantage, was calculated individually for each subject. The mean and 95% confidence intervals of the spatial advantage are shown in the lower panels

of Fig. 4.3 and Fig. 4.4. A three-way repeated measures analysis of variance (ANOVA) was applied to all measured results. It showed significance for type of masker applied [$F(2,16) = 61.7, p < 0.001$], spatial configuration [$F(1,16) = 316.5, p < 0.001$] and the number of dialogues in the cafeteria [$F(1,16) = 337.9, p < 0.001$]. The ANOVA also showed significance for all types of interaction effects, namely between type of masker and spatial configuration [$F(2,32) = 29.0, p < 0.001$], type of masker and number of dialogues [$F(2,32) = 11.7, p < 0.001$], spatial configuration and number of dialogues [$F(1,16) = 10.8, p < 0.005$] and finally for interaction between all three dependent variables [$F(2,32) = 15.2, p < 0.001$].

4.3.1 Two-dialogue cafeteria

For the two-dialogue masker results shown in Fig. 4.3, a t -test with Bonferroni correction did not reveal a significant difference between the different-talker and vocoded-talker maskers in the colocated condition ($p = 0.24$) but found weak significance in the spatially distributed condition ($p < 0.05$). However, the same-talker masker was significantly different from both the different- and vocoded-masker in the colocated ($p < 0.001$) but not in the spatially separated condition ($p = 0.11$ and $p = 0.88$ for each masker, respectively).

The lack of significant difference between the SRTs measured with the different- and vocoded-talker masker in the colocated condition suggests that IM has little relevance in the different-talker masker. A similar conclusion can be drawn for the spatially separated condition, where the SRT for the different-talker masker was even lower (by 1 dB) than for the vocoded-talker masker. The small difference may be linked to an increase in EM due to the temporal and spectral smearing applied in the vocoding process, which may reduce dip-listening cues. Reconsidering the colocated results, if such dip-listening cues are indeed reduced with the vocoded-talker masker, they might counteract a small amount of IM in the different-talker masker and consequently produce the non-significant difference in this condition.

In the colocated condition, the SRT for the same-talker masker was approximately 6 dB larger than for both the vocoded-talker and different-talker masker. This significant difference indicates that the same-talker masker produced more EM due to more spectral overlap (see Fig. 4.1) and a substantial amount of IM. When the spatial separation was

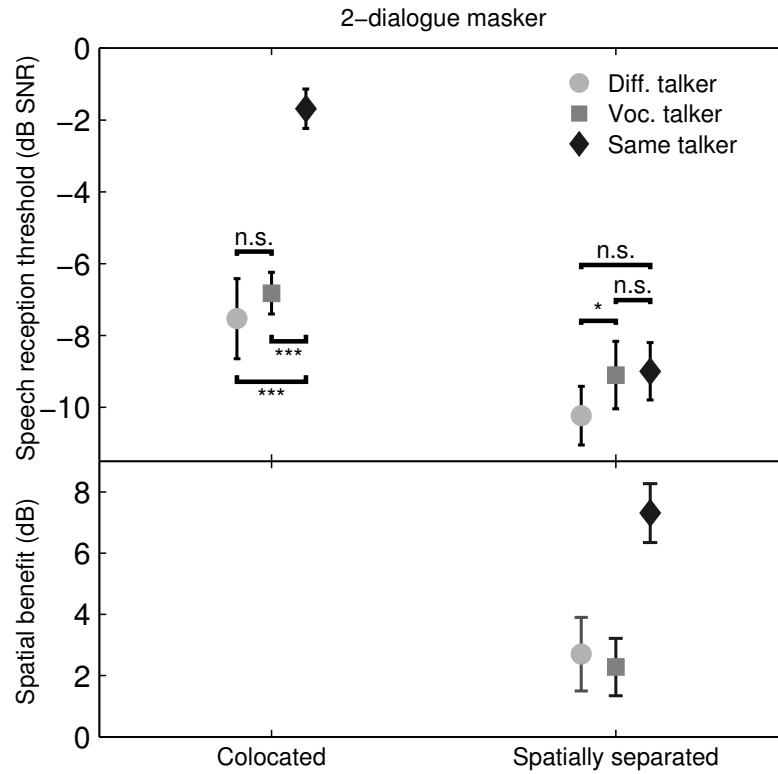


FIGURE 4.3: Top panel: Mean and across-subject 95 % confidence interval of SRTs (divided between colocated and spatially separated) for different-, vocoded- and same-talker maskers (circles, squares and diamonds, respectively). Bottom panel: Mean and across-subject 95% confidence intervals of the spatial advantage, i.e. the difference between the spatially separated SRT and the colocated SRT calculated individually for each subject. Stars indicate level of significance between conditions (i.e. *, ** and *** correspond to $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively).

introduced the SRTs for the same-talker masker improved by almost 8 dB to a level that was similar to that of the spatially separated different-talker and vocoded-talker masker. The fact that the SRTs measured with the same-talker masker were similar to the SRTs measured with the vocoded-talker masker in the spatially separated condition suggests that the contribution of EM is equivalent. Furthermore, in this condition the slightly higher SRTs measured with the different-talker masker (albeit not significantly) could indicate that the same-talker masker provides slightly more EM than the different-talker masker. Hence, the spatial separation substantially improved intelligibility with the same-talker masker and effectively removed the IM observed in the colocated condition.

In general, the SRTs decreased in all conditions when shifting from colocated to spatially separated masker presentation. For the different-talker and vocoded-talker masker, this spatial benefit, or SRM, was around 2.5 dB and for the same-talker masker, it was approximately 8 dB.

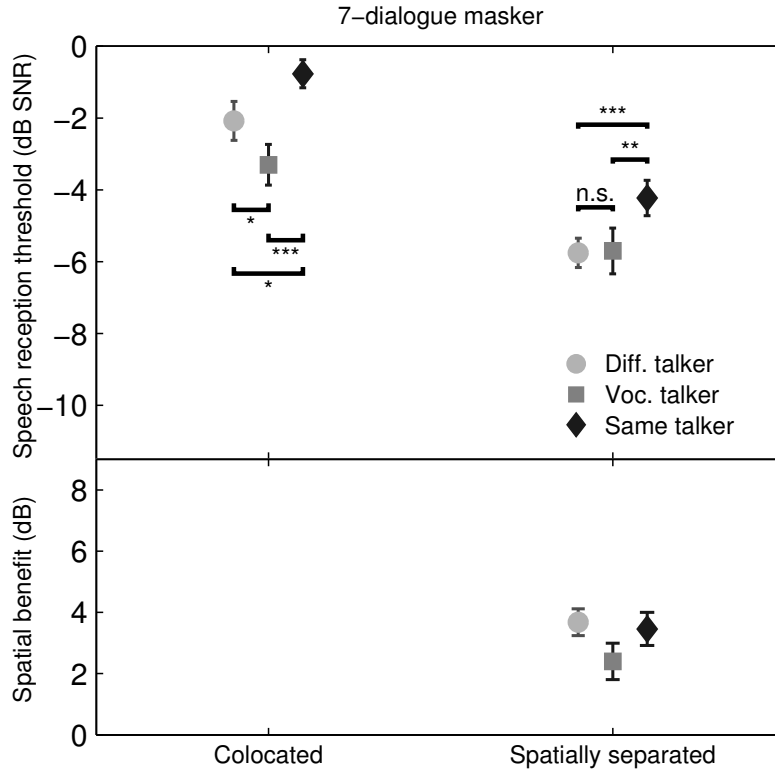


FIGURE 4.4: Same as for Fig. 4.3, but for the seven-dialogue masker.

4.3.2 Seven-dialogue cafeteria

For the results of the seven-dialogue masker shown in Fig. 4.4, a paired t -test with Bonferroni correction revealed that the different-talker and vocoded-talker SRTs were significantly different in the colocated condition ($p < 0.05$), but not in the spatially separated condition ($p = 0.29$). The same-talker condition was significantly different from both the different-talker and vocoded-talker masker in the colocated ($p < 0.05$ and $p < 0.001$, respectively) as well as the spatially distributed condition ($p < 0.01$ and $p < 0.001$, respectively).

In the colocated condition the SRT for the different-talker masker was about 1.2 dB higher than for the vocoded-talker masker. This difference was not observed in the two-dialogue masker condition (Sec. 4.3.1) and might indicate a minor involvement of IM. However, the difference in SRT between the vocoded-talker and different-talker masker was removed by spatially separating the masking talkers from the target talker.

Comparing the individual SRTs in the two- and seven-dialogue conditions, SRTs in the seven-dialogue condition with the different- and vocoded-talker maskers are substantially

higher both in the colocated and spatially separated condition. This may be explained by the fact that there are fewer temporal and spectral fluctuations when more masking talkers are present, which results in a decreased number of available spectro-temporal gaps with high SNR. Hence, increasing the number of maskers decreases the possibility of dip-listening. However, the spatial benefit was very similar (of about 2.5 dB) for the two- and seven-dialogue background for both the different- and vocoded-talker masker.

The same-talker masker, as in the two-dialogue conditions, resulted always in the highest SRTs. However, here the difference between the same-talker and the other two maskers in the colocated condition was reduced to about 2 dB as opposed to the 6 dB observed in the two-dialogue condition. This substantial reduction in SRM was mainly due to an increase in SRTs in the spatially separated condition, whereas the SRTs in the colocated condition were very similar between the two- and seven-dialogue maskers.

4.4 Discussion

Throughout literature it has been shown that differences in talker characteristics as well as spatial location resolve target-masker confusions and thereby reduce, or even completely remove IM (Bronkhorst, 2000; Brungart *et al.*, 2001). In these studies, spatial cues mainly due to angular differences between target and masking talkers were considered, but Westermann and Buchholz (2014a) have additionally shown that room-reverberation cues provided by a separation in distance can similarly reduce IM. In Sec. 2.3, both spatial separation in angle and distance, as naturally occurring in the real world, was applied to spatially separate the maskers from the target. This resulted in SRTs which for the different-talker condition were equal or even lower than for the vocoded-talker condition. This was true for the two-dialogue as well as seven-dialogue masker condition and indicates that the spatially separated different-talker conditions are dominated by EM. The interaction, or relative importance of differences either in talker or location, is exemplified with the two-dialogue masker (Fig. 4.3) where the different-talker masker shows no or at least very little IM in the colocated condition and spatial separation removes IM from the same-talker masker. However, for the seven-dialogue masker (Sec. 4.3) it was found that talker cues alone could significantly reduce but not fully remove IM. The SRT for the different-talker masker in the colocated condition was about 1.3 dB lower than for the same-talker masker, but still 1.2 dB higher than for the vocoded-talker masker. This difference was removed by providing spatial

cues. However, it should be noted that the seven-dialogue masker is expected to exert less IM than the two-dialogue masker. This has been shown on several occasions in literature (Freyman *et al.*, 2004). However, in comparison to other studies looking in the relationship between number of talkers and IM the masking talkers were all at the same level. Since the talkers in this experiment were at different distances it was pertinent to explore condition where less IM was expected. Overall from the results it can be said that, in realistic scenarios where both spatial and talker cues are available, listeners can rely on both cues to severely reduce or even completely remove IM. Throughout the experiment, while some IM effects might have been present and could have been teased apart by substantially increasing the amount of test subjects, the current study indicates that such effects would be minor in terms of dB. In addition, listeners in real-life communication would also often have access to visual cues, which have been shown to be even more effective in reducing IM with speech maskers (Helfer and Freyman, 2005). Thus, confusion-based IM, as it is discussed here and in numerous other studies (Sec. 4.1), seems to have a negligible effect on speech intelligibility in real-life scenarios when one or more of these cues are available.

However, these findings rely on two assumptions: (1) that the vocoded-talker masker does not result in IM and (2) that the EM contribution of the different-talker masker and vocoded-talker masker are the same. According to the definition of IM set fourth in Sec. 4.1 (and several other studies: Ihlefeld and Shinn-Cunningham, 2008; Kidd *et al.*, 2007) to successfully remove confusion-based IM from the vocoder-talker masker, the vocoder processing needs to ensure that no target speech segments are confused with the masker. An informal listening test found that the masker was unintelligible and the vocoding method made the target clearly stand out, but no formal testing was done. Other studies have applied similar processing to construct EM-only maskers, but only applied the broadband envelope to the noise with the masker long-term spectra (Best *et al.*, 2013b; Brungart *et al.*, 2001). However, this type of processing does not capture any of the fluctuations of speech within each frequency band, thereby likely violating the second assumption. The vocoder used in this study applied spectral smoothing (one or two octaves) that was wider than the auditory critical bands in order to ensure that the masking speech was unintelligible while maintaining most within frequency band fluctuations. However, there is no clear solution to how to best solve this trade-off when designing a vocoder for creating EM-only stimuli that does not violate the two

assumptions and thus the parameters were chosen heuristically. Furthermore, as noted in Sec. 2.1 this type of EM-only reference does not address IM occurring in the basic auditory object formation stage.

For the two-dialogue cafeteria conditions, the different-talker SRT was significantly lower than the vocoded-talker SRT ($p < 0.05$) in the spatially distributed condition. This further confirms that this different condition is not affected by IM, but also suggests that the vocoded-talker masker provides more EM than the different-talker masker. If this is the case, then this effect might have been counteracted in the colocated condition by a small amount of IM resulting in the insignificant difference between SRTs measured with the different- and vocoded-talker masker. Several studies have shown that SRTs measured with modulated speech-shaped noise (similar to single-channel vocoders) maskers are higher than SRTs measured with “irrelevant” different-talker maskers (Bernstein and Grant, 2009; Festen and Plomp, 1990; Qin and Oxenham, 2003). The difference in SRTs found in these studies are between 1 and 5 dB, corresponding with the difference observed in this study. In the present study, the discrepancy between vocoded- and different-talker SRTs might be explained by the vocoding process smearing the signal over time and frequency and thereby, reducing dip-listening as well as spatial cues, in particular cross-ear glimpsing cues. Glyde *et al.* (2013) showed how reduced spectral resolution, following a (moderate) hearing loss, significantly reduces cross-ear glimpsing cues. This resulted in increases in SRTs of about 1-2 dB in their spatially separated condition (± 90 degrees). The reduction in spectral resolution following a hearing loss is similar to the effect of spectral smoothing applied during the vocoding process and thus, a similar increase in SRT may be expected here.

In the colocated condition, the SRTs for the same-talker masker were substantially higher than for the vocoded-talker, clearly indicating involvement of IM. This is expected, since the target-masker similarities both in terms of voice and spatial location cause failures in streaming and auditory object selection. For the two-dialogue cafeteria condition, the measured spatial benefit was approximately 8 dB with the same-talker masker. This benefit is comparable or slightly lower than benefits reported in other SRM studies with two masking talkers (e.g., Bronkhorst, 2000; Freyman *et al.*, 1999; Glyde *et al.*, 2013). However, these others studies compared colocated masker conditions with two maskers that were spatially separated to ± 90 degrees in an anechoic environment. In the current study, the two-dialogue maskers were not spaced symmetrically (see Fig. 4.2) as

one masker (M_2) was positioned considerably closer to the target than the other (M_7) masker. In addition, room reverberation was included here, which has been shown to reduce the effect of SRM (Kidd and Mason, 2005). Finally, the maskers employed here highly supported listening in dips which is significantly reduced in corpora such as the CRM because of time-aligned target and maskers.

The measured SRM for the different-talker and vocoded-talker maskers was approximately 3 dB, both in the seven- and two-dialogue condition. Other studies that applied maskers which only resulted in EM found spatial benefits of up to 8 dB (Best *et al.*, 2013b; Kidd and Mason, 2005). However, Kidd and Mason (2005) measured the SRM with several different levels of reverberation, and in their most reverberant condition the spatial benefit was only 3 dB. Plomp (1976) conducted similar experiments and found spatial benefits of 3.2 dB in a room with a reverberation time $T_{30} = 0.4$ s. Generally, they explained the smaller spatial benefit by a reduction in interaural fluctuations caused by reverberation. The results found in this study are in good agreement with those found in the studies that considered room reverberation.

Generally, the SRTs for the two-dialogue masker were lower than for the corresponding seven-dialogue masker conditions. Other studies that measured the effect of number of masking talkers have found similar behavior (Brungart *et al.*, 2001; Freyman *et al.*, 2004). This effect is commonly linked to the increasing advantage of dip listening as the masker fluctuations increase with decreasing number of masking talkers. However, this is only the case when EM is dominant, but when IM is additionally involved this increase in SRT with increasing number of talkers is counteracted by at least two additional mechanisms. As the number of masking talkers increases, the masker becomes more noise-like. In consequence, stream formation of individual maskers that can be confused with the target speech is less likely to occur and the maskers tend to form a single, fused background stream instead. In addition, if the number of masking talkers increases and the total masker SPL is kept constant, the level of each talker decreases in relation to the target and, as a result, the difference in loudness between target and individual masking speakers increases. This difference in loudness provides a strong segregation cue that very much limits the occurrence of IM at high SNRs or more accurate, at high target-to-masker energy ratios (TMR) (Agus *et al.*, 2009).

The observation that the applied SNR has an effect on the occurrence or strength of IM

has already been discussed by other studies. Best *et al.* (2013b) for instance, argued that in the case of spatially separated target and masker, the auditory system is able to fully segregate the target from the masker signals and as a consequence, SRTs are dominated by EM effects. Hence, when sufficient speech segregation cues are available, SRTs seem to be limited by EM and no IM can be observed. Similarly, Brungart *et al.* (2001) and Agus *et al.* (2009) showed that when the level of the target in reference to each individual masker, as defined by the TMR, exceeds 0 dB TMR the effect of IM dramatically drops. Naturally, above 0 dB TMR the target is louder than each individual masker and loudness cues can be used for talker segregation. Hence, IM seems to have a limited dynamic range, which is limited at low SNRs by EM and above 0 dB TMR by loudness cues. The exact details are complicated and will depend on a large number of scene-related factors (e.g., the number, sex, and spatial configuration of masking talkers) as well as subject-related factors (i.e., hearing ability). Since in real-life the scene-related factors are dictated by the encountered acoustic scene, this leaves the obvious question: which TMRs are encountered in real-life listening? TMRs applied in typical anechoic SRM experiments are easily estimated, but this estimation becomes harder when including room acoustics and masking sources at varying distances. For the scenario applied in this study (Fig. 4.2), ODEON supplies the predicted levels of each individual source. Assuming normal vocal effort for the target and masking talkers, the predicted TMR in the spatially distributed cafeteria scene with seven-dialogues (Fig. 4.2d) ranges between 1.8 (masker M_3 facing towards the listener) and 6.5 dB (masker M_1 facing away from the listener). When combining each of the masking dialogues by averaging their predicted levels, the predicted SNR is -4.5 dB. Hence, even at negative SNRs the TMRs are still positive.

Smeds *et al.* (2014) investigated the SNRs that listeners typically experience in their daily life and found most of the relevant SNRs observed in multi-talker, cafeteria-like environments to be positive (around 3 dB SNR). Hence, TMRs in such conditions would be very positive and loudness cues abundant for target segregation. Hence, in most challenging, multi-talker environments encountered in real-life, the involvement of IM is even more unlikely than in the experiments considered here.

4.4.1 Perspectives

It can be argued that morphemic sentence tests are a poor representation of real-life communication. Not only do they lack conversational dynamics and listener involvement, they also ignore any form of comprehension. In addition, they neglect attention switching between different target sources and do not take into account listening effort. Several studies have looked into increasing the amount of realism in speech tests. In addition to applying more true-to-life SNRs, speech material and comprehensions tasks which mimic real-life conversations are desirable. Best *et al.* (2013a) compared sentence recall and comprehension in the same complex cafeteria environment as applied in this study (seven-dialogue cafeteria; Fig. 4.2d). They measured comprehension in 18 normal-hearing and 28 hearing-impaired listeners by conducting an on-going questionnaire assessing the listeners understanding of the monologues presented in the cafeteria. Although, they found a strong correlation ($r = 0.77$, $p < 0.001$) between the comprehension scores and the SRT measured with a sentence test (same as applied here) the comprehension test revealed additional information on the cognitive abilities of the subjects.

Cognition and IM have often been linked (Kidd *et al.*, 2007), but as far as the authors are aware, no studies have shown a significant correlation between the individual susceptibility to IM and cognitive measures (i.e. Glyde *et al.*, 2012). However, some studies have applied a dual-task paradigm, and found a relation between working memory capacity and SRM (Helfer *et al.*, 2010). It could be of interest to apply a similar methodology to the study presented here, especially, if it was possible to create scenarios with distraction-based IM, thus investigating other types of IM as defined in Sec. 2.1. This type of IM could be driven by the salience or novelty of a stimulus (e.g., Knudsen, 2007), adding cues in the masker familiar to the subject (such as a their name e.g., Wood and Cowan, 1995) or by including maskers closer to the listener than the target (e.g., Westermann and Buchholz, 2014a). Attention- or distraction-related IM may well be observed in real-world environments, but this needs to be further investigated.

Overall, further investigations will be required to generalize the contribution of IM to other real-life listening environments. In particular, it would be interesting to increase the possibility of talker confusions by using a less reverberant room, maskers more similar to the target (e.g. all male talkers) or other spatial configurations. The reproduction

method limits the minimum distance of simulated sound sources to the distance of the loudspeakers in the array (here 1.85 m), although real-world listening often occurs at nearer distances, which could significantly change the applied TMRs. In order to generalize the results, more conditions with both close target and maskers are needed. In addition, the study could be expanded to include subjects with a hearing impairment. Since hearing loss is often accompanied by reduced temporal, spectral, and spatial resolution, these subjects may be more susceptible to IM or show an increased dynamic range of IM. If this group exhibits significant susceptibility to IM, it might be possible to design algorithms for hearing devices which reduce IM.

4.5 Summary and conclusion

This study investigated the role of IM in a simulated, reverberant cafeteria environment by systematically varying the similarity between target and masking talkers as well as their spatial configuration. The results demonstrated the following:

1. Significant IM was only observed in the colocated condition with the same-talker masker (i.e. when the target and masking talkers were all the same person).
2. Differences in either location or talker resolved target-masker confusions and effectively removed IM. It was further argued that the involvement of IM is even less likely when visual cues are additionally included and increased SNRs are considered, as in many real-world environments the SNR is slightly higher than the SNRs considered in this study.
3. The SRM observed in the simulated cafeteria environment is considerably smaller than the SRM typically reported in literature, where the effect of room reverberation is typically excluded and symmetrical two-talker masker conditions are considered. The SRM with the two-dialogue same-talker masker was about 8 dB, but only approximately 3 dB for all other maskers. The diminished spatial benefit was explained by reduced interaural differences and lack of IM.

Overall, this study suggests that IM is negligible in most realistic multi-talker environments and that IM is often exaggerated in psychoacoustic experiments by colocated target and masker conditions and copora with excessive confusions. However, it should

be noted that this study mainly focussed on confusion-based IM, which is most commonly studied, but did not explicitly consider aspects such as attention switching or distraction-related IM.

4.6 Acknowledgments

This work was supported by an International Macquarie University Research Excellence Scholarship (iMQRES) and Widex A/S.

Chapter 5

The effect of nearby maskers in reverberant, multi-talker environments¹

The extent to which informational masking (IM) is involved in real-life listening is not well understood. In literature, IM effects of more than 8 dB are reported, but these experiments typically used simplified spatial configurations and speech corpora with exaggerated confusions. Westermann and Buchholz (2014b; Chap. 4) considered a more realistic cafeteria environment and only found substantial involvement of IM when the target and maskers were colocated and the same talker. The present study further investigates the practical relevance of IM in real-world environments IM, specifically considering the effect of hearing impairment and distractions by nearby maskers. Speech reception thresholds (SRTs) were measured with normal hearing (NH) and sensorineural hearing impaired (HI) listeners in a simulated cafeteria environment. Three different masker configurations were considered: (1) seven dialogues distributed in the cafeteria (2) two monologues presented close to the listener with varying angular separation and (3) a combination of (1) and (2). The contribution of IM was measured as the difference in SRT between speech maskers and unintelligible vocoded maskers. No significant IM was found with the dialogues alone. However, including nearby maskers resulted in substantial IM for NH and HI listeners. These results suggest a distance-dependant prioritization of sound sources in complex scenes and that this prioritization of nearby maskers is especially important when considering IM.

¹Aspects presented at the Association for Research in Otolaryngology MidWinter meeting (2014). Chapter represents a manuscript to be submitted to The Journal of the Acoustical Society of America.

5.1 Introduction

For years, researchers have investigated the auditory mechanisms related to understanding speech in reverberant multi-talker environments (e.g. Bronkhorst, 2000; Cherry, 1953). In such conditions, it has been shown that the auditory system can take advantage of talker characteristics, such as differences in fundamental frequency, spatial location and fluctuations in maskers to better understand a target talker (Brungart *et al.*, 2001; Festen and Plomp, 1990; Freyman *et al.*, 1999). When speech is masked by speech the concepts of informational masking (IM) and energetic masking (EM) are often applied (for a review see (Kidd *et al.*, 2007)). While EM describes masking effects that occur because of overlap in the auditory periphery, IM is often related to more central, cognition-based masking effects. However, no conclusive definition exists for IM and its boundary to EM. In this study, IM is defined as a result of (1) *confusions* introduced by target and masker similarities and (2) *distraction* occurring as a result of the masker capturing the attention of the listener. This definition is line with Kidd *et al.* (2007) as well as (Westermann and Buchholz, 2014b).

Several studies have tried to quantify the influence of IM. Generally, these studies employ a reference condition with a high level of target-masker confusions and a method that enables segregation of the target and masker signals in order to measure the reduction, or release from, IM. Such reference conditions normally include target and masker in the same location (colocated) (Freyman *et al.*, 1999), and speech corpora with many inherent confusions (Bolia *et al.*, 2000). Thereby, confusions are often created by using the same talker to realize target and masker speech as well as using speech material that has a very similar, synchronized sentence structure for the target and maskers. Perceptual segregation has been introduced from changes in spatial location (Freyman *et al.*, 1999), gender and talker characteristics (Brungart *et al.*, 2001) or source-receiver distance (Westermann and Buchholz, 2014a). Other studies have estimated the influence of IM by comparing intelligibility with a speech masker and a speech-modulated noise-masker (Best *et al.*, 2013b; Brungart *et al.*, 2001). Across studies it has been suggested that IM effects can result in differences of up to 8 dB in measured speech reception thresholds (SRTs). However, the reference condition applied in all of these studies relies on speech corpora and spatial configurations with exaggerated confusions.

Westermann and Buchholz (2014b) investigated the influence of IM in a simulated cafeteria environment presented via a three-dimensional loudspeaker array. They measured SRTs in a background of dialogues consisting of the same talker as the target, different talkers or unintelligible noise vocoded talkers that were either colocated with the target or distributed throughout the simulated room. Overall, they found no contribution of IM in conditions that were representative of real-life, i.e. when the masking talkers were spatially distributed and different from the target. Furthermore, they argued that in conditions where limited cues were available to segregate the target from the maskers the contribution of IM was dependent on the level of the target talker compared to the level of each individual masker, known as the target-to-masker ratio (TMR). Since the TMRs were predominantly positive in their simulated cafeteria (as in most real-world environments: Smeds *et al.*, 2014), they concluded that the influence of IM is low in realistic environments. However, they mainly considered confusion-based IM and did not consider the effect of maskers close to the listener, which typically provide rather low TMRs. They also did not consider the effect of a hearing impairment, which may decrease the salience of auditory cues as well as cognitive performance and thus, may affect the susceptibility to either form of IM.

Few studies have looked into the effect of IM with masker that are closer to the listener than the target. Lavandier and Culling (2007) employed a number of conditions with nearby maskers to measure the effect of the direct-to-reverberant ratio on spatial release from masking (SRM), but they always kept target and masker with 65° angular separation which likely resolved IM. Westermann and Buchholz (2014a) investigated the effect of differences in distance on SRM when target and masker were directly in front of the listener with normal hearing (NH), and later, with hearing impaired (HI) listeners (Chap. 3). They found that placing a masker further away in distance from the target resolves IM and leads to improved SRTs; however, in the opposite case when the masker was closer to the listener than the target, they observed a substantial IM effect. This effect was especially pronounced with HI listeners. Analyzing the errors the listeners made when the masker was closer to the target, they found very little target-masker confusion and thereby concluded that the involved IM may be related to the distraction by the maskers (i.e., affecting selective attention) rather than confusions. However, their experiment applied highly confusing speech corpora and a colocated reference condition, both limiting the ecological validity of their results.

The current study presents a speech intelligibility test in a simulated cafeteria that combines the increased ecological validity of Westermann and Buchholz (2014b) with the nearby maskers of Westermann and Buchholz (2014a). The test was specifically designed to estimate IM effects with nearby maskers in realistic environments, and it was conducted on both NH and HI listeners. Furthermore, cognitive testing was conducted on the HI listeners in order to estimate the relationship between cognitive performance and susceptibility to IM from nearby maskers.

5.2 Methods

5.2.1 Subjects

In this study 16 NH (12 female, six male) and 16 HI (six female, ten male) native Australian English speaking subjects participated. The mean age of the NH subjects was 29.2 years, and the subjects had pure-tone audiometric thresholds ≤ 20 dB hearing loss (HL) at audiometric frequencies from 250 Hz to 8000 kHz. These subjects were either employees of the National Acoustic Laboratories or students at Macquarie University. The HI listeners had a mean age of 72.5 years and all had symmetric (threshold differences between ears of < 10 dB), mild to moderate sensorineural hearing losses. Individual audiograms and mean and standard deviation of the audiometric thresholds are shown in Fig. 5.1. All HI subjects had extensive experience with psychoacoustic experiments. Before the study, all participants gave written consent and subjects not associated with National Acoustic Laboratories (NAL) were given a gratuity for their participation.

5.2.2 Stimuli

These experiments used sentences from the Bamford-Kowal-Bench (BKB) corpus (Bench *et al.*, 1979). This speech corpus contains 336 sentences with an approximate length of 1.5 s and simple syntactical structure (e.g. “The girl lost her doll”). Sentences are spoken by a native Australian-English male speaker, sampled at 44.1 kHz and divided into 21 lists. As in Westermann and Buchholz (2014b) a 512-tap inverse finite impulse response (FIR) filter is applied to the original BKB material to make its long-term spectra match the long-term spectrum of a long (65 minutes) anechoically recorded monologue spoken by the original speaker. This filtering stage removed the original spectral shaping which

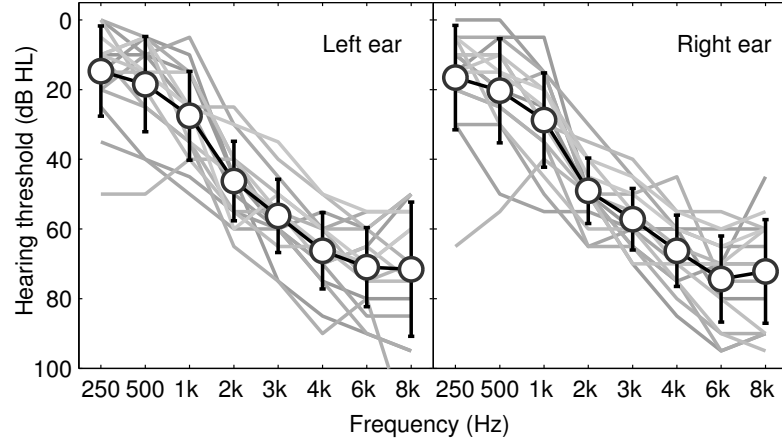


FIGURE 5.1: Mean and standard deviation of the pure-tone audiometry for the 16 HI participants.

was applied to match the long-term average speech spectrum (LTASS) defined in Byrne *et al.* (1994) and made the sentences sound more natural in the cafeteria background described in Sec. 5.2.4.

Two different maskers were realized to estimate the influence of IM: (1) a *speech masker* comprised of talkers different than the target talker (2) a *vocoded masker* which was a vocoded version of the speech masker. The influence of IM was defined as the difference in SRTs between the two maskers. The maskers were approximately 5 minutes in length and were continuously looped throughout the experiment.

The speech masker consisted of a mixture of seven background two-talker conversations (or dialogues; note dialogues are all two-talker conversations as used in Westermann and Buchholz, 2014b) and two nearby maskers containing monologues. The dialogues and monologues were taken from the International English Language Testing System (IELTS) and recorded in the anechoic chamber at the National Acoustic Laboratories with six male and eight female, native Australian English speakers. The level of each individual talker was equalized using the speech level calculation methodology of the Speech Transmission Index (STI; IEC 60268-16). This processing was mainly applied to disregard the long speech pauses due to the turn-taking in the dialogues. Due to the limited number of recorded talkers, the monologues were spoken by two male talkers that also appeared in the dialogues. It has been shown on numerous occasions that the amount of IM that can be expected is proportionate with the number of masking talkers, with the maximum around two talkers (Freyman *et al.*, 2004). Hence, the seven-dialogue background masker is not expected provide substantial IM which was also confirmed in

Westermann and Buchholz, 2014b. On the other hand, this study mainly addresses the effect of the nearby maskers on IM.

To separate the effects of EM and IM, a vocoded version of the speech masker was implemented. The aim of the vocoder processing was to make sure that the combined, multi-talker background speech was completely unintelligible while maintaining the spatial percept of multiple noise-sources distributed around the listener. This was realized by preserving a high temporal resolution, thereby maintaining transients (as much as possible), and in turn, strongly smoothing the spectra. The short-time Fourier transform (STFT) was used to convert each of the anechoic speech maskers to the time-frequency domain. The applied window length was 20 ms with 75% overlap. The resulting time-frequency representation was spectrally smoothed across either one octave for the seven background dialogues or two octaves for the nearby maskers. The additional smoothing of the nearby masker was applied to assure that the vocoded speech was unintelligible.

5.2.3 Equipment

The speech testing was conducted in a spherical loudspeaker array available in the anechoic chamber at the National Acoustic Laboratories. Outside the anechoic chamber, a PC running MATLAB generated and played the sound files. The PC was fitted with a RME MADI sound card connected to two RME M-32 D/A converters. The analog output of the converters was amplified by 11 4-channel Yamaha XM4180 amplifiers whose output was fed into the anechoic chamber through an acoustically dampened passage and connected to each individual loudspeaker in the array. The loudspeaker array consisted of 41 Tannoy V8 loudspeakers arranged on a sphere with a radius of 1.85 m. The subject were seated on a height adjustable chair such that their head was in the center of the loudspeaker array. To reproduce the direct sound component of the nearby maskers, four 8080 Genelec monitor loudspeakers were suspended inside the array in level with the primary ring of 16 loudspeakers at a distance of 0.85 m from the center of the array. These small speakers were placed between the array loudspeakers at $\pm 11.25^\circ$ and $\pm 55.25^\circ$ and hung only with thin strings to minimize acoustical shadow. Since the Genelec monitors contained amplifiers, they were connected directly to the RME M-32 D/A converter through a balanced cable.

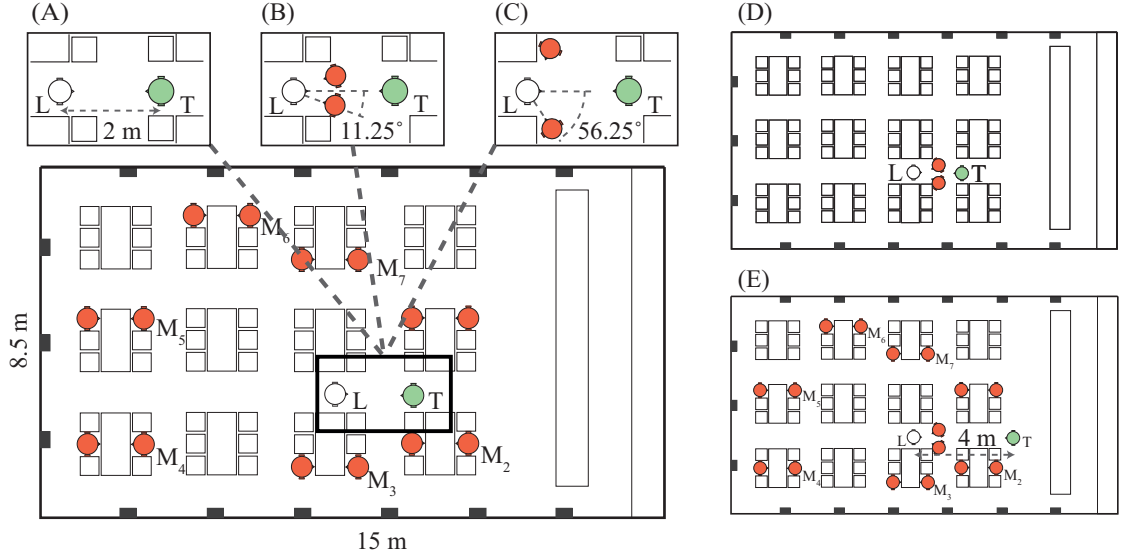


FIGURE 5.2: Top-down view of cafeteria simulated in ODEON for each of the measured conditions (A through E). The listener, L, faces the target, T, at either 2 m (A through D) or 4 m (condition E) distance. The masking dialogues were distributed in the room ($M_1 - M_7$) and the nearby maskers were separated from the target by either $\pm 11.25^\circ$ or $\pm 55.25^\circ$.

5.2.4 Spatialization of sounds

The loudspeaker array described in Sec. 5.2.3 was used to reproduce the cafeteria environment shown in Fig. 5.2. Firstly, a model of the room was created in ODEON (Rindel, 2000), which included various surfaces such as tables, chairs and windows, each with their individual inherent absorption coefficients. The room was 15 m long by 8.5 m wide by 2.8 m high and had a reverberation time of $T_{30} \approx 0.6$ s. The sources were placed as shown in Fig. 5.2 and realistic talker directivity was included by applying ODEON's directivity file `Tlknorm_natural.so8`). For each source the acoustic path to the listener, as captured by the room impulse response (RIR), was calculated by ODEON providing reflectograms and decay curves which were then processed with the loudspeaker-based room auralization (LoRA) toolbox (Favrot and Buchholz, 2010). Within the LoRA toolbox, the reflectogram is used to map the direct sound and specular early reflections (here up to third order) to the nearest loudspeaker in the playback loudspeaker array. The late reverberation was added by applying the frequency- and direction-dependent decay envelope to uncorrelated noise. This resulted in 41 impulse responses (IRs), corresponding to each channel in the loudspeaker array, for each sound source (excluding the nearby maskers).

Four of the conditions included nearby maskers (Fig. 5.2, B through E) which were

substantially closer to the listener than the loudspeakers in the array. Therefore, four small loudspeakers were suspended inside the spherical loudspeaker array at positions corresponding to each of these nearby maskers. The direct sound component of the nearby maskers was presented from these additional loudspeakers, whereas the remaining part of the RIR was reproduced using the 41-channel loudspeaker array. Thereby for each of the nearby maskers, a 42-channel IR was derived.

In order to spatialize each sound source in the simulated cafeteria (Fig. 5.2), the anechoic speech recordings described in Sec. 5.2.2 were convolved with the corresponding multi-channel IRs. To reduce individual variation of loudspeaker sensitivity and to compensate for the difference in arrival time of the near loudspeakers, equalization filters were designed for each loudspeaker and applied to all stimuli.

The maskers were fixed to an ecologically appropriate sound pressure level (SPL). In addition to room acoustical information about each source, ODEON also estimates its level at the receiver, assuming a given vocal effort. Here normal vocal effort was applied, and the overall level of the different cafeteria backgrounds was calculated by summing the predicted power of all involved sources. For the conditions shown in Fig. 5.2 the SPLs were 59.4, 65.4, 65.4, 62.3 and 65.4 dB(A) for condition A, B, C, D and E, respectively.

In order to (partially) restore audibility for the HI participants, the NAL-RP linear amplification scheme was adopted (Dillon, 2001). To remove effects and variability related to hearing aid processing and to ensure externalization of sound sources (i.e. localizing sources outside the head, rather than inside) the hearing loss compensation was performed at the loudspeakers rather than at the listeners' ears. However, this removed the ability to apply different prescriptions across ears, and as a result, symmetrical hearing loss was a recruitment requirement. The measured HL was entered in the NAL-RP formula and an insertion gain in third-octave channels was calculated. This insertion gain was limited to 0 dB for negative values (i.e., attenuation from the compensation at low frequencies was removed) and applied to a series of 20 third-octave linear-phase FIR filters covering a frequency range from 100 Hz to 8 kHz. The weighted third-octave filters were summed to provide a single, subject-specific FIR filter which was then applied to all 45 loudspeakers (including the four nearby loudspeakers) to realize individual amplification according to NAL-RP.

5.2.5 Cognitive measures

Two common cognitive tests were included in this study to measure cognitive performance of the HI listeners: (1) a computerized reading span test (RST) and (2) a Stroop test.

The RST assesses effective working memory capacity (Daneman and Carpenter, 1980). Participants were seated in front of a computer screen and presented an increasing number of sentences (from two to six). Subjects were given two tasks: (1) to determine whether the sentence was semantically meaningful or not and (2) after each set of sentences, to recall either the first or last word of each sentence. Without the participants knowledge, performance was only scored on the latter task.

The Stroop test aims to capture executive function abilities (Golden and Freshwater, 2002). It has three components, each of which are scored by the number of colors correctly read from a sheet in 45 seconds. The sheets comprised of 100 words printed either in black or colored ink (e.g. “green”, “red”, “blue” or “purple”). The first sheet named colors in black ink, the second consisted of four X’s in colored ink and the final sheet contained the names of colors written in ink with a different color. In the first task the participant had to read the color name, and in the second two the participants had to name the color of the ink (while ignoring the word). From the test an interference score (i.e. effect of the word naming another color) was deducted according to Golden and Freshwater (2002).

5.2.6 Procedures

Testing of each participant was completed in one appointment. For both NH and HI listeners, an audiometric screening was first conducted in a double-walled booth. While in the booth, the HI listeners then completed the computerized RST and Stroop test (Sec. 5.2.5). These tests took approximately 15 minutes. Subsequently, participants were taken to the anechoic chamber and seated in the center of the loudspeaker array. The height of the chair was adjusted to ensure that the subject’s head was at level with the center of the array. The participants were instructed on the task and fitted with a lavalier microphone to communicate with the test administrator outside the chamber. For the HI listeners, the measured audiogram was entered in a MATLAB script which in turn computed and applied the NAL-RP equivalent loudspeaker gain.

Preceding the main speech test, a short training session was conducted to familiarize the participants with the task and interaction with the test administrator outside the chamber. Here one of the 21 lists was presented in the cafeteria background without the nearby maskers (Fig. 5.2, condition A). The results obtained from the training were discarded. During the test 10 SRTs were measured, i.e. condition A through E (Fig. 5.2) each with the speech and vocoded masker (Sec. 5.2.2). The masker level was kept constant while the target level varied according to an adaptive, one-up one-down staircase method to estimate the signal-to-noise ratio (SNR) which yielded 50% correct performance. The testing procedure was implemented following Keidser *et al.* (2013), requiring for each SRT a minimum of 16 presentations and ended either when the standard error fell below 1 dB or when 32 sentences were presented. Since each list of the BKB material contained 16 sentences, two randomly selected lists were combined for each measured SRT. Across participants the order of presentation was randomized. The entire test took approximately 1.5 hours, and halfway through the participants were given a short break.

5.3 Results

5.3.1 Speech intelligibility measures

The top panels of Fig. 5.3 shows the mean and 95% confidence intervals of the measured SRTs. The triangles and squares show the results for the NH and HI listeners, respectively, and the gray symbols denote the speech masker and black symbols denote the vocoded masker. The bottom panels of Fig. 5.3 show the estimated IM, which is the difference between the speech and vocoded masker SRT, calculated individually per subject. Two-way repeated measures analysis of variance (ANOVA) were applied separately to the NH and HI data. For the NH listeners, it showed significance for both the condition [$F(2.48, 37.26) = 79.29$], type of masker [$F(1, 15) = 226.07$] and interaction between the two [$F(4, 60) = 13.52$]. Note, the Greenhouse-Geisser correction was applied to the condition effects to ensure that a violation of the sphericity assumption did not influence the significance of the observed effects. For the HI listeners the two-way ANOVA also showed significance for condition [$F(3, 45) = 9.74, p < 0.001$], masker type [$F(1, 45) = 36.93, p < 0.001$] and interaction [$F(3, 45) = 6.62, p < 0.001$]. Post-hoc *t*-tests with Bonferroni correction were applied to compare the speech and vocoded masker

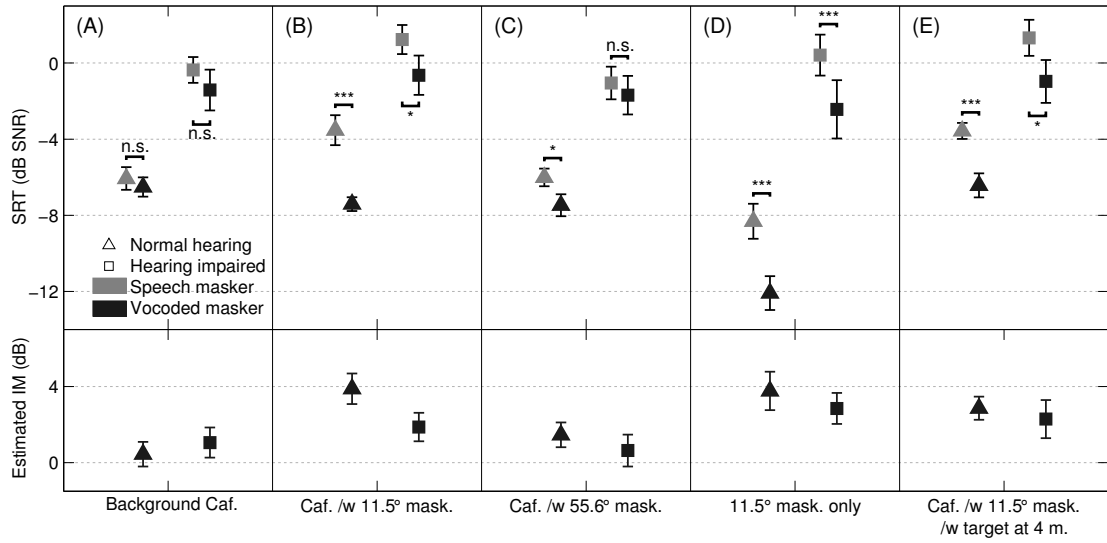


FIGURE 5.3: Top panels: Mean and across-subject 95 % confidence interval of SRTs for conditions A through E according to Fig. 5.2. Speech and vocoded maskers are indicated by gray and black symbols, respectively, and the NH participants by triangles and HI by squares. Stars indicate level of significance between conditions (i.e. *, ** and *** correspond to $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively). Bottom panels: Mean and across-subject 95% confidence intervals of the estimated IM (i.e. the difference between the speech masker SRT and the vocoded masker SRT).

SRT and results are indicated in Fig. 5.3 (i.e. *, ** and *** correspond to significance of $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively).

5.3.1.1 Normal hearing listeners

Overall, with only the background cafeteria (condition A) the difference in SRTs between the speech and vocoded masker was not significant for the NH listeners. This is in agreement with and confirms the findings presented by Westermann and Buchholz (2014b).

When the nearby maskers are introduced, substantial and significant differences between speech masker and vocoded masker SRTs of up to 4 dB can be observed across conditions B, C, D and E. In contrast to the study by Westermann and Buchholz (2014b), which did not find any IM in more realistic conditions, this suggests that IM effects can indeed be observed in realistic environments, but only when nearby maskers are present. The increased separation when the nearby maskers were shifted from $\pm 11.25^\circ$ to $\pm 56.25^\circ$ (condition B and C) substantially reduced the speech masker SRTs, but did not change the vocoded SRTs. This indicates a spatial release from IM as angular separation increases.

In the condition with only the masker at $\pm 11.25^\circ$ and no background maskers (D), the SRTs were considerably lower than when the background maskers were included. This indicates that considerable dip-listening is available to the NH listeners. However, the estimated IM for the two conditions is very similar, confirming that IM is introduced by the two nearby maskers and not disturbed by the cafeteria background. However, this will depend on the level difference between the nearby and background maskers. When this difference is decreased the nearby maskers will be less dominant and most likely the observed IM effect will diminish.

Overall, there was not a big change when the target was moved from 2 m to 4 m distance (between condition B and E). Only the SRT for the vocoded masker increased slightly, which resulted in a small reduction in IM of less than 1 dB.

5.3.1.2 Hearing impaired listeners

The overall behavior of the results for the HI subjects shown as squares in Fig. 5.3 is very similar to the NH listeners described in Sec. 5.3.1.1, except that overall SRTs are about 4 dB higher on average for the HI listeners. Higher threshold were expected because of the reduced audibility, frequency selectivity and temporal resolution of the impaired auditory system. In addition, the overall benefit of removing the background cafeteria dialogues (between condition A and D) was smaller for the HI than the NH listeners. However, a reduced ability to effectively use masker fluctuations (or “listen in the dips”) following a hearing impairment has been reported in multiple studies (e.g., Bernstein and Grant, 2009; Festen and Plomp, 1990).

Similar to NH listeners, also no significant involvement of IM was observed for the HI listeners in the cafeteria background without nearby maskers (condition A), although the SRT for the speech masker showed a tendency towards higher values than the vocoded masker. This could indicate that HI listeners are more susceptible to IM than NH listeners when multiple, partially-intelligible masking talkers are involved. Also similar to NH subjects, when the nearby maskers was moved from $\pm 11.25^\circ$ to $\pm 56.25^\circ$ (comparing condition B and C), the SRT for the speech masker decreased significantly but not for the vocoded masker and thus, highlighting a substantial spatial release from IM.

Somewhat surprising, the HI listeners showed smaller differences in SRTs between speech and vocoded maskers when the nearby maskers were present (conditions B, C, D, and E),

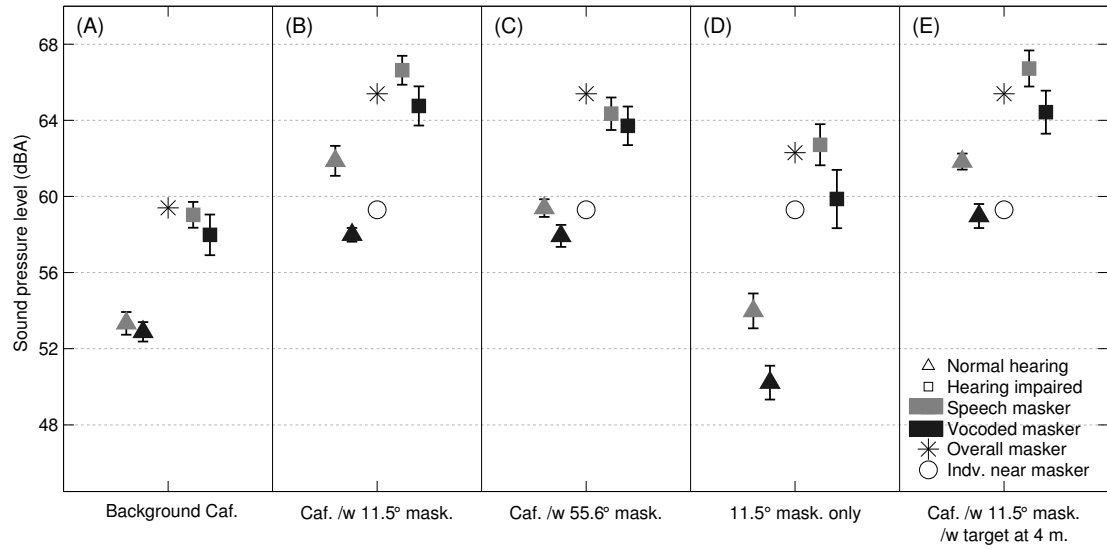


FIGURE 5.4: As Fig. 5.3. However, now the value shows the SPLs at the measured SRTs. In addition, the overall SPL of the masker and the level of each nearby masker is shown by the star and circle, respectively.

demonstrating that the involved IM is still significant but by about 1-2 dB smaller than in NH listeners. Generally, it has been argued that IM effects are independent of hearing loss (Agus *et al.*, 2009; Helfer and Freyman, 2008), and as far as the authors are aware no studies have suggested reduced IM following a hearing impairment. Westermann and Buchholz (2014b) argued that the occurrence of IM is dependent on the TMR, and diminishes at positive TMRs. The latter conclusion is further supported by a number of related studies, including Helfer and Freyman (2008) and Brungart *et al.* (2001). To illustrate the TMRs involved in this experiment, in particular between the target and the nearby maskers, the SRT data shown in Fig. 5.3 is replotted in Fig. 5.4, but this time the SPL of the target at the SRT is shown instead of the SNR at the SRT. Additionally, the overall level of the masker is shown for each condition by an asterisk (*), and in the conditions with a nearby masker, the level of each individual nearby masker is shown by a circle (○). The effective TMR in relation to each of the nearby maskers is the difference between the target SPL value and the SPL of the nearby masker. In conditions where the masker was present (B, C and E), the NH TMRs were close to 0 dB, while the HI TMRs were around +5 dB. In the nearby masker alone condition (D), the TMRs were around -5 to -8 dB for NH and around +1 to +4 dB for HI subjects. Hence, the NH and HI subjects are tested at different TMRs, which might explain the difference in the observed IM.

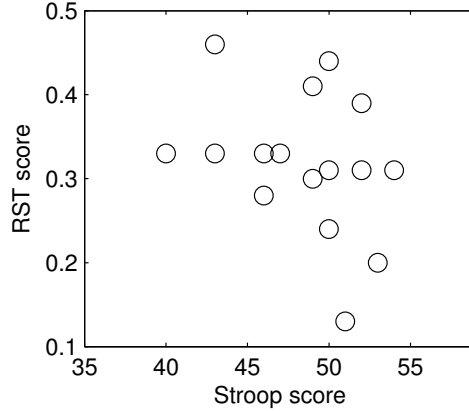


FIGURE 5.5: RST and Stroop score measured with the HI listeners. Each \circ represents one participant.

5.3.2 Cognitive measures

In Fig. 5.5 the results of the RST and Stroop tests are plotted against each other for the individual listeners (indicated as circles). For both the RST and Stroop test, higher scores indicate better performance. A Pearson's linear regression found no significant correlation between the two scores ($r^2 = 0.1$ and $p = 0.2$).

Furthermore, Tab. 5.1 summarizes results of a linear regression analysis between the four-frequency average hearing loss (4FAHL), age, RST score and Stroop test score on SRTs and SRT differences between speech and noise maskers, i.e. the estimated amount of IM. Here only the data for the conditions A and B are presented, but the other conditions with nearby maskers showed very similar results to condition B. Overall, there was no correlation between RST scores or age on any of the measures. There was a weakly significant correlation between the 4FAHL and the measured SRTs but not for the estimated amount of IM. This correlation between HL and intelligibility in noise is in agreement with many other studies (e.g., Agus *et al.*, 2009; Glyde *et al.*, 2012). Moreover, in condition A there was a weakly significant correlation between the Stroop score and the estimated amount of IM, indicating that the susceptibility to IM might be linked to executive function ability. However, this dependency was not observed in any of the other conditions.

5.4 Discussion

5.4.1 Effect of informational masking

TABLE 5.1

Measure	Predictor	Regression results	
		r^2	p
(A) Vocoded SRT	4FAHL	0.6	< 0.001
	Age	0.03	0.5
	RST	0.01	0.8
	Stroop	0.00	0.9
(A) Estimated IM	4FAHL	0.00	1
	Age	0.01	0.7
	RST	0.00	0.9
	Stroop	0.3	< 0.05
(B) Vocoded SRT	4FAHL	0.4	< 0.001
	Age	0.09	0.3
	RST	0.01	0.8
	Stroop	0.01	0.7
(B) Estimated IM	4FAHL	0.05	0.4
	Age	0.00	0.9
	RST	0	1
	Stroop	0.03	0.5

Westermann and Buchholz (2014b) could not find any IM in a simulated cafeteria environment when the maskers were different talkers from the target and spatially distributed. Their conclusion was confirmed in the present study, considering a very similar cafeteria background masker (Condition A). However, in all other conditions (B, C, D, and E), in which nearby maskers were introduced, a significant IM effect of up to 4 dB was observed.

Examples of such conditions in real-life could be class-rooms in which where a student is trying to listen to a teacher while nearby students are talking, or a dinner table scenario where everybody is talking and a person is trying to follow a conversation on the other side of the table. However, it is not known how the IM effect reported in this work translates to real-life listening; especially considering that the effect was only pronounced when the maskers located in a similar direction as the target, i.e. at $\pm 11.25^\circ$ with the target at 0° . Moreover, Helfer and Freyman (2005) showed that the inclusion of visual cues, allowing lip-reading, significantly reduces the contribution of IM. Although visual cues are often available in real life, they require the presence of sufficient light, which is not always the case, or the talker of interest is too far away or not facing the listener. Hence, the nearby masker conditions considered in this study may in general

be relevant for real-life listening, but it does not necessarily refer to a very common situation.

Westermann and Buchholz (2014b) found substantial IM effects when applying a speech masker that was closer to the listener than the target. Because of the properties of the Coordinate response measure (CRM) corpus that they applied, they were able to measure the number of occurring target-masker confusions. In this way, they argued that the observed IM with nearby maskers was not caused by confusions but by the target distracting the listener and thus, affecting selective attention. Since a similar setup is realized here, it is likely that also distraction- or attention-based IM is mainly involved. However, since the employed speech corpus does not allow counting target-masker confusions, it is difficult to draw definitive conclusions. When listening to the conditions with nearby maskers, the target and masker sounded very different and seemed rather difficult to confuse, but it was difficult to ignore the masker. This was supported by the subjects' comments, that the conditions with nearby maskers were the most "*annoying*" and that "*it was hard to block out the person in the hanging speaker*". Further testing with a closed-set speech corpus that can effectively measure target-masker confusions (such as the CRM corpus) could be applied; however, applying a corpus that exaggerates confusions contradicts the goal of this study to better understand the effect of IM in more realistic environments.

There was almost no difference in SRTs when the target was at 4 m (Fig. 5.3, condition E) compared to the similar configuration with the target at 2 m (Fig. 5.3, condition B). Since the target level is adjusted adaptively, the distance-dependent level changes are not included. Thereby, the main difference between the target signal at these two distances is a change in the direct-to-reverberant energy ratio (DRR). Zahorik (2002b) showed that DRR just-noticeable differences (JNDs) were around 5 – 6 dB for NH listeners, and Akeroyd *et al.* (2007) later demonstrated that such JNDs were much higher for HI listeners. As the difference in target position applied in this study represents a doubling of distance (maximally changing the DRR by 6 dB), and furthermore, as reverberation resulting from the target is masked by the background dialogues, it is likely that the subjects did not even perceive the change in target distance.

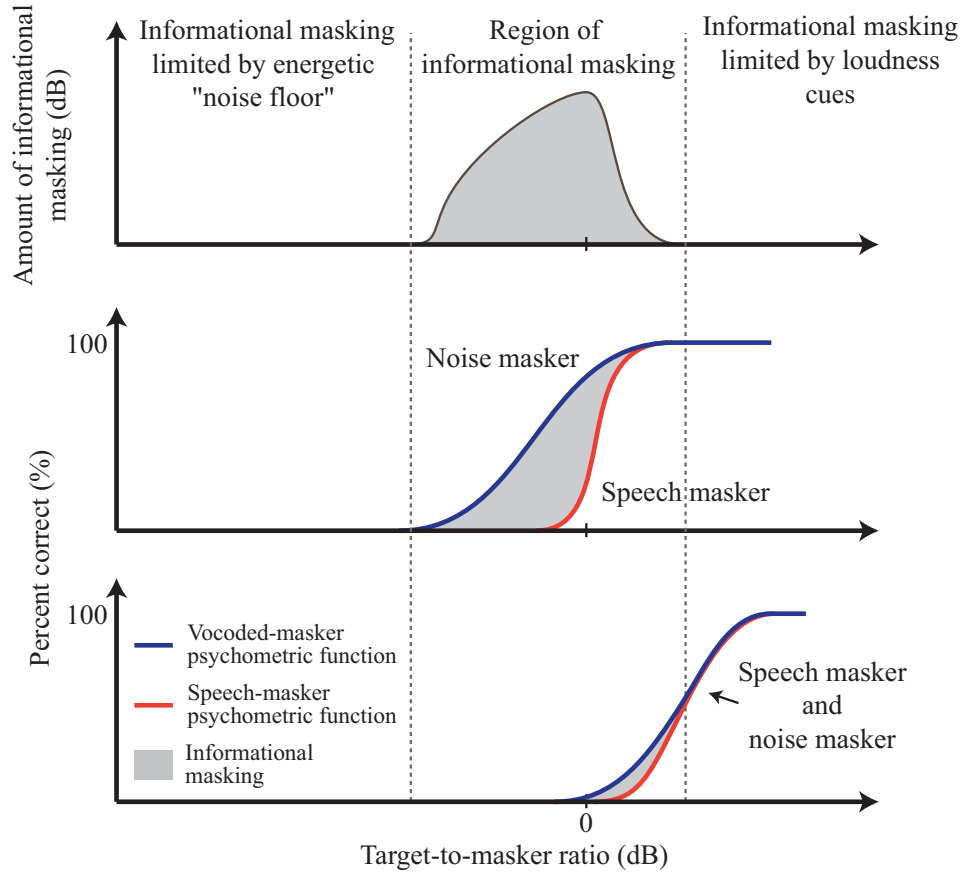


FIGURE 5.6: Schematic of the non-linear region of IM as a function of the TMR. TMRs outside this region are less likely to be affected by IM.

5.4.2 The region of informational masking

With reference to Agus *et al.* (2009) and Brungart *et al.* (2001), Westermann and Buchholz (2014b) argued that when the target SPL is higher than the SPL of each individual masking talker, i.e. the TMR is above 0 dB, IM is resolved by loudness cues.

Analyzing a cafeteria background masker similar to condition A, they showed that all TMRs at the measured SRTs are above 0 dB, which explains why no IM is observed in such condition (Fig. 5.3, panel A). In addition to this upper limit of IM, it has been shown that in conditions where sufficient cues are available to segregate the target from the masker, the SRTs are limited by EM. The IM that is observed in the colocated, same-talker reference condition is resolved when target and maskers are spatially separated

Hence, a “region of IM” can be defined with a lower boundary given by EM and an upper boundary given by loudness cues due to positive TMRs. This is illustrated in the upper panel of Fig. 5.6. The existence of a “region of IM” is further supported by

Agus *et al.* (2009), who showed that IM reaches a maximum at a TMR of just under 0 dB and falls off towards lower and higher TMRs. Even though the “region of IM” is a straight forward concept, the detailed behavior is rather difficult to understand and will depend on acoustic factors (e.g., type and number of masking talkers and their spatial distribution), auditory factors (e.g., hearing loss limiting temporal, spectral, and spatial cues as well as reduced sensitivity to loudness), the applied speech material (e.g., temporal and semantic structure; context information), and maybe cognitive factors (e.g., executive function ability).

Having defined such a region, two requirements can be formulated for IM to occur: (1) the target and masker must not be fully segregated perceptually (for further discussion see Ihlefeld and Shinn-Cunningham (2008)) and (2) the TMR must fall into the region of IM. However, these rules are based on confusion-based IM, and it has been shown that perceptually segregated (non-confusable) maskers can still cause distraction-based IM (Westermann and Buchholz, 2014a). Even though, the limitations of distraction-based IM are not well known, a concept such as a region of IM might still apply.

In relation to the results from this study, Fig. 5.4 illustrates that the nearby maskers in this study created TMRs around 0 dB, especially for the NH listeners. As illustrated in the middle panel of Fig. 5.6, conditions with TMRs around 0 dB fall into the middle of the region of IM. Thus, it is not surprising that NH listeners are especially affected by IM (Sec. 5.3.1.1).

The reduced IM observed with the HI listeners (Sec. 5.3.1.2) can also be explained by the concept of a region of informational masking. Fig. 5.4 shows that for SRTs measured with the HI listeners, the SPLs of the individual nearby maskers (shown as the \circ) are substantially lower than the target SPL. Hence for the HI listeners, the TMRs were greater than zero and thereby pushed outside the “region of IM”. This is illustrated in the bottom panel of Fig. 5.6.

The idea that the effect of IM decreases with increasingly positive TMR is further supported in Fig. 5.7, which shows the relationship and linear regression analysis between the SRT measured for the speech maskers (also thereby the applied TMR) and the observed amount of IM, i.e. the difference between the SRT for the speech and vocoded maskers. Here only conditions A and B (Fig. 5.2) are considered, but similar conclusions can be drawn for the other conditions. When only the background cafeteria (condition

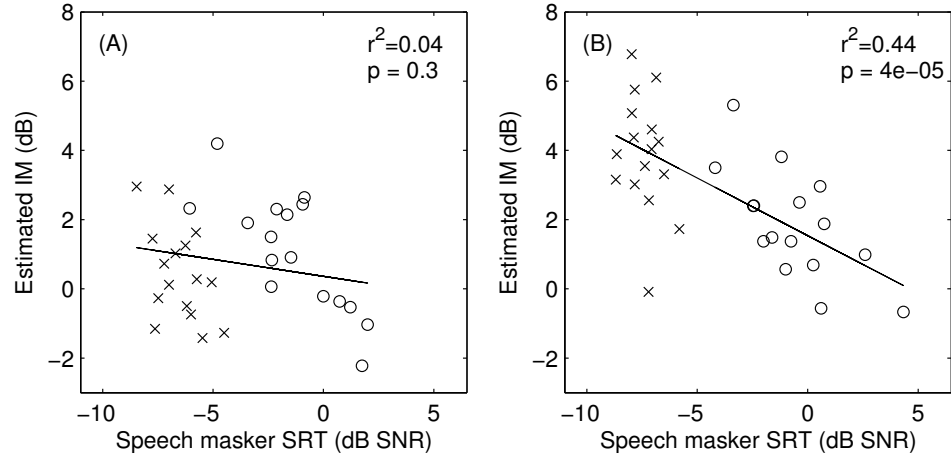


FIGURE 5.7: Relationship between individually measured speech masker SRTs and the estimated IM in the cafeteria with (right panel, condition B) and without (left panel, condition A) the nearby maskers. NH and HI listeners are shown as crosses and circles, respectively.

A) is present the TMRs are outside the region of IM (see Fig. 5.4) and, in average, no significant involvement of IM can be observed for NH subjects and a minor involvement for HI subjects (Sec. 5.3). However, a substantial spread in IM can be seen across the individual subjects (Fig. 5.7, left panel), but this spread is independent of the SRT ($r^2 = 0.04$ and $p = 0.3$). However, when the nearby maskers are included (condition B) a significant, though weak, linear relationship is found ($r^2 = 0.4$ and $p < 0.001$), supporting the idea that the involved IM effect decreases with increasingly positive TMR. This is consistent for the three other conditions with nearby maskers ($r^2 = 0.25 \dots 0.28$ and $p < 0.01$) and can explain the generally reduced amount of IM observed in HI subjects.

Overall, to further evaluate these TMR-related effects, conditions with fixed SNRs (and thereby also fixed TMRs) across the NH and HI listeners could be included. However, since NH and HI SRTs with the vocoded masker are approximately 6 dB apart it might be difficult to find an appropriate SNR which would be intelligible for all HI listeners while not reaching a ceiling of 100% intelligibility for NH listeners.

5.4.3 Cognition and informational masking

Informational masking is often linked with auditory cognition (Glyde *et al.*, 2012; Helfer and Freyman, 2008; Kidd *et al.*, 2007) and some even go so far as calling it “cognitive masking”. However, as far as the authors are aware no studies have successfully shown a strong relationship between cognitive measures and susceptibility to IM or release

from IM. Glyde *et al.* (2012) measured SRM with varying degrees of IM as a function of hearing loss and age and applied the COGSTAT questionnaire to measure individual cognitive ability, but found no correlation between the results.

For the cognitive measure used in this study, only the Stroop score correlated with the individual differences in IM and only in condition A. While these results are in no way conclusive, they suggest that mainly measures which include executive function and inhibition are of relevance to IM. This would also be in line with the observation that the IM involved in this study is most likely due to the nearby maskers distracting the subjects from attending to the target speech (Sec. 5.4.1). To establish further statistical power more subjects are needed and it could be of interest to compare the result with normative data from the young NH listeners, which unfortunately was not measured here.

Other studies have employed a dual-task paradigm, where participants are scored both on the speech experiment and on a secondary task (e.g., Helfer *et al.*, 2010). The hypothesis is that while intelligibility might be comparable between conditions, effort, or cognitive load, is different and can be measured by performance on a secondary task. It could be interesting to include a secondary and to measure its interaction with IM. This might even be different for confusion-based or distraction-based IM.

5.4.4 Perspectives

When only the background cafeteria was present (condition A), the HI listener showed a slight increase in SRTs, though not significant, when measured with speech maskers rather than vocoded maskers. In this background cafeteria condition Westermann and Buchholz (2014b) established that all TMRs at SRT are significantly above 0 dB and thus this condition should be out of the “region of IM”. The fact that HI still show signs of IM could indicate that HI listeners are either more susceptible to IM than NH listeners in general or their “region of IM” is extended to higher TMRs. Hence, they might simply be more distracted by the intelligible conversations around them.

The conditions and locations of the nearby maskers were chosen heuristically. To minimize long-term better ear SNR effects, the nearby maskers were placed symmetrically around the listener. Also to reduce fluctuations between conditions, only monologues were used for the nearby maskers. However, the effect of dialogues between nearby

maskers could be significant. In addition, other configurations might be of interest such as nearby maskers behind the listener, which is a more realistic condition as the frontal condition at $\pm 11.25^\circ$ but only provides very subtle differences in localization cues. The latter might be particularly interesting when the subjects wear hearing aids, which can introduce significant front-back confusions (Best *et al.*, 2010). While it is believed that the nearby maskers cause IM of a distracting rather than confusing nature, it has been shown that other factors can draw a listener's attention to the masker, such as hearing their own name or topics of interest (Wood and Cowan, 1995). Such features could be added to the masker to further enhance its distracting ability.

5.5 Summary and conclusion

This study investigated the ecological relevance of IM by considering a simulated cafeteria environment, thereby expanding the study of Westermann and Buchholz (2014b) by including nearby distracting maskers, HI listeners and cognitive measures. Generally the results showed:

1. In contrast to Westermann and Buchholz (2014b) who did not find any IM in conditions where target and masker were spatially separated and different talkers, this study demonstrated that IM can occur when near masking talkers are introduced. However, the resulting IM was most likely not due to target-masker confusions as most commonly considered, but rather due to the nearby maskers distracting the listeners from attending to the target speech.
2. As the nearby maskers were spatially separated from $\pm 11.25^\circ$ to $\pm 56.25^\circ$, the contribution of IM decreased. This spatial release from IM demonstrates that, even if nearby maskers are present, they need to be located in a similar direction as the target to introduce substantial IM effects.
3. The SRTs measured with the HI listeners were considerably higher than measured with the NH listener, especially in the conditions without the cafeteria background. Moreover, the HI listeners appeared to be less susceptible to IM. However, it was discussed that this was mainly a consequence of their higher SRTs which resulted in TMRs that were above 0 dB and significantly higher than for the NH listeners. These higher TMRs shifted the HI listeners out of the “region of IM” and thereby offered loudness cues that partially resolved IM.

4. Cognition has often been linked to IM. However, the cognitive measures applied here could not explain susceptibility to IM on an individual subject level. The RST showed no correlation with any of the data. And the Stroop test, which addresses executive function, was only slightly correlated with the the amount of IM measured individually, but only in the cafeteria background without nearby maskers. Generally, more work needs to be done to tie the potential link between cognition and IM.

Overall, this study suggests that real-life listening can involve IM when nearby maskers are present. However, in most other real-life listening conditions, IM seems to be of rather low relevance.

5.6 Acknowledgments

This work was supported by an International Macquarie University Research Excellence Scholarship (iMQRES) and Widex A/S. The authors would also like to thank Margot McLelland and Katrina Freeston for help with data collection.

Chapter 6

General summary and discussion

This thesis presented a series of experiments that studied and linked the role of auditory distance perception and informational masking (IM) when listening in complex acoustic environments. Initially, it was shown how both the normal hearing (NH) and hearing impaired (HI) auditory system can effectively use distance-related reverberation cues to segregate target and masking talkers. In addition, the results provided evidence for distraction-based, rather than confusion-based, IM effects. Secondly, a series of experiments in a simulated reverberant multi-talker environment found that confusion-based IM is to a large degree irrelevant when sound sources are spatially separated and the target talker is different from the maskers. However, with the inclusion of nearby maskers substantial distraction-based IM effects were observed both for NH and HI listeners. In order to explain differences observed in IM across subjects and acoustic scenarios, the concept of a “region of IM” was proposed, which at low speech reception thresholds (SRTs) is limited by energetic masking and at high target-to-masker ratios (TMRs) by loudness cues.

Chapter 2 presented a study outlining how NH listeners can effectively use differences in the direct-to-reverberant energy ratio (DRR), arising from spatial separation in distance, to improve the intelligibility in a two-talker background. Improvements of up to 10 dB were observed in the measured SRTs when the target was at 0.5 m and the masker was moved from 0.5 m to 10 m. These results were consistent when applying both the Coordinate response measure (CRM) and listening in spatialized noise sentences-test (LISN-S) speech corpora. Since this improvement decreased by less than 3 dB in a diotic condition, it was concluded that the effect was mainly monaural. A series of predictive signal-based measures were applied, but only the modulation domain signal-to-noise ratio (SNR) could predict qualitatively the improvement resulting from the spatial separation. It was argued that IM had a significant role when the target and

masker were colocated, but that spatial separation in distance aided perceptual segregation and thereby resolved confusions. This was further supported by comparing the results to SRTs measured with a speech-modulated noise masker, that showed no effect of spatial separation in distance. When the masker was fixed at 0.5 m distance and the target distance increased, the mean SRTs only slightly decreased, however, a large variability between subjects was observed. By analyzing the type of errors occurring in the colocated condition, it was shown that errors were not caused by target-masker confusions but rather by the masker competing for the listeners attention, causing distractions. This led to a distinction between confusion- and distraction-based IM, which was further investigated throughout the thesis.

The effect of spatial separation in distance on the intelligibility of speech for HI listeners was investigated in Chap. 3. Overall, intelligibility improvements were similar to those reported for NH listeners in Chap. 2. A small reduction in SRT improvements compared to NH listeners of about 3 dB, when the target was kept at 0.5 m distance and the masker was moved from 0.5 m to 10 m, could be explained by increased SRTs with the (purely energetic) speech-modulated noise masker. Increased effectiveness of energetic masking (EM) in HI listener is attributed to reduced frequency and temporal resolution of the impaired auditory system. In conditions where the masker was at 0.5 m distance and the target was moved from 0.5 m to 10 m distance, mean intelligibility decreased and SRTs were widely spread. Thereby it was suggested that the listeners in this study were more susceptible to distraction-based IM in comparison with the NH listeners in Chap. 2.

Chapter 4 presented a series of experiments that aimed to measure the contribution of IM, specifically related to confusions, in a simulated cafeteria. The environment was auralized in a 41 channel loudspeaker array in an anechoic chamber. Using this reproduction method allowed for a more realistic acoustical environment compared to headphone-based experiments. In general, IM effects were only present in the colocated condition and were only pronounced when the masker was the same talker as the target. Hence, spatial separation and talker differences resolved target-masker confusions. The experiment also calculated the effective spatial release from masking (SRM), which was around 3 dB, for maskers with different talkers than the target. This is much lower than found in idealized studies with anechoic presentation, “complete” angular separation (e.g., $\pm 90^\circ$) and speech corpora with exaggerated confusions. Overall, the prevalence of IM is related to the underlying TMR and it was argued that loudness cues in more

realistic environments, resulting from primarily positive TMRs, resolve target-masker confusions.

The concepts of the previous three chapters were combined in Chap. 5. Again a simulated cafeteria was used, but this time nearby maskers were included. In addition, this study also included HI listeners and two cognitive measures. It was shown how substantial IM effects occur once nearby maskers are included, both for NH and HI listeners. This connects to Chap. 2 and 3, where such conditions showed evidence for distraction-based IM. When the angular separation of the nearby maskers was increased from $\pm 11.25^\circ$ to $\pm 56.25^\circ$, the contribution of IM decreased. This both conforms with SRM literature but also suggests that distraction-based IM can be resolved by spatial separation. Comparing the results for the NH and HI listeners, the HI listeners were less affected by IM contradicting the results of Chap. 3. In order to explain this behaviour, the relationship between IM and the applied TMR was expanded to form the concept of a non-linear “region of IM”. The “region of IM” illustrates that because HI listeners generally had higher SRTs, they were tested at positive TMRs and, as a result, were less susceptible to IM. While the results of the cognitive measures in this study did not show strong correlations with the measured susceptibility to IM on an individual level, there was an indication that tests that focus on inhibition, such as the Stroop test, could be related to distraction-based IM.

6.1 Perspectives and limitations of this work

While this thesis addresses several aspects of auditory processing, it simultaneously raises many new questions. All of the studies presented here, were reliant on adaptive speech tests to estimate SRTs, often in a range from negative to very negative SNRs. While literature that reports effective SNRs in our day-to-day lives is very scarce, it has been shown that SNRs are mainly positive (Smeds *et al.*, 2014). The observations of Smeds *et al.* (2014), were supported by simulations in the cafeteria discussed in Chap. 4, where similar positive SNRs were found. Hence, all of the experiments in this thesis, including those that aim to represent a realistic environment, were effectively conducted at unrealistic SNRs. This is a confounding problem for the entire field and something that is currently being addressed by multiple research groups.

In addition to problematic SNRs, the experiments in the thesis relied on sentence-tests. While real-world communications are dynamic and involve listener engagement, rapport and simultaneous processing of what is being said and how to respond, sentence-tests are static and unengaging. As a result, performance measured with such tests might not correlate well with real-life performance. Researchers at the National Acoustic Laboratories as well as other research groups around the world, are aiming to develop tests which better represent communication in our daily lives. Once such tests have been established and possibly combined with similar reproduction methods as described in this thesis, we can hope to get a more accurate picture of auditory performance in the “cocktail party”.

The differentiation between confusion-based and distraction-based IM has been an underlying topic throughout this thesis. While the concept of distractions has been established before, research has mainly focused on confusion-based IM effects, as such effects have been easy to introduce and have shown reliable results (Kidd *et al.*, 2007). However, according to the latter part of this thesis (Chap. 4 and 5), it is mainly the distraction-based IM that is of relevance for many real-life environments. The entire concept of distractions raises many questions: What defines distractions and how can they be described acoustically or perceptually? How and why do they influence listeners differently? How can distractions be incorporated into our testing methodology in a controlled manner? What is their relevance in complex real-world environments? To answer such questions, would require coupling between cognitive, acoustical and psychoacoustical research fields.

In relation to the first two studies, it is evident that cues resulting from differences in distance can facilitate perceptual segregation of talkers. However, this study did not include the level differences that mainly define auditory distance perception. So while these studies are a starting point for distance-related SRM research, further work is still warranted. For one, it seems essential to investigate the relationship between DRR and level cues, as well as the relation between distance and angular separation. Finally, the results of these studies were obtained with the use of the same speech material and spatial configurations that exaggerate confusions which was shown in Chap. 4 and 5 to have limited relation to more realistic listening scenarios.

Throughout the thesis, there was evidence that nearby maskers increase the effect of IM.

From a theoretical perspective, it was argued that this was caused by the nearby signal being “cleared” than the signal that were further away, but this definition is relatively vague. It could be interesting expanding this research by means of exclusion, namely designing studies looking into which features are important. Besides distance itself, possible candidates to manipulate are cross ear coherence, DRR and coloration.

Furthermore, the conditions with nearby-maskers often resulted in large variation in SRTs between listener. Besides the investigation on the effect of training in Chap. 2 and cognition in Chap. 5, this thesis did not expressively discuss the individual variations between subjects. New research has suggested correlations between individual differences in SRTs measured in speech-on-speech masking as well as modulation sensitivity with periodicity coding in the auditory brainstem (Ruggles *et al.*, 2011). This is especially relevant with the nearby masker that, as discussed in Chap. 2 and 5, will have increased fluctuations or modulations compared to a masker that is further away where the room has a low-pass effect in the modulation domain. Undoubtedly, these large variations are of interest and should be studied further with multiple different approaches.

Each of the studies presented in this thesis employed a EM-only reference masker from which the contribution of EM was estimated. However, as discussed in Chap. 4, the validity of such maskers are not straight forward as they require a particular definition of IM. Firstly, all of the maskers presented in this work (as discussed in Chap. 4), do not address IM which is encountered at the basic auditory grouping stages. For such investigations, other more specialized EM-only references would need to be applied. Secondly, the EM-only reference maskers did aim at replicating any of the modulation masking of the original signals - only the short-term and long-term energy of the masker. There is evidence that such differences could have substantially skewed the results (Stone *et al.*, 2011). This leads back to the recurring difficulties with EM-only references, a field which needs more research.

The last two studies presented in Chap. 4 and 5, introduce an underlying relationship between the encountered IM, aptly named the “region of IM”. While such a concept is supported by literature (Agus *et al.*, 2009; Best *et al.*, 2013b; Brungart *et al.*, 2001), it needs to be investigated further. As mentioned, the exact shape and boundaries of such a region would be dependent on the spatial condition, hearing ability, and stimulus factors such as amount of interferes, sex of talker and other properties of the applied speech

corpus. Nevertheless, it could be of interest to explicitly measure this relationship and incorporate this knowledge into auditory models that aim to predict speech-on-speech masked intelligibility, but have failed thus far in certain scenarios because of IM (Glyde *et al.*, 2013; Jørgensen and Dau, 2013).

While two of the studies in this thesis incorporated HI listeners, only simple linear amplification was applied to partially restore audibility. Since hearing device processing, such as dynamic range compression and directional processing, will inevitably influence the observed effects, it could be worthwhile to expand the studies presented here to include listeners wearing a hearing device. Furthermore, for electrical hearing (i.e., cochlear implants), reduced spectral and spatial resolution combined with loss of temporal fine-structure (Wilson and Dorman, 2008) would likely greatly increase susceptibility to IM. One could imagine that confusion-based IM effects would be critical because of the cochlear implant's dynamic range limitations, greatly reducing loudness cues and the loss of fundamental frequency coding impairing talker segregation.

Appendix A

Binaural dereverberation based on interaural coherence histograms¹

A binaural dereverberation algorithm is presented which utilizes the properties of the interaural coherence (IC) inspired by the concepts introduced in Allen *et al.* (1977). The algorithm introduces a non-linear sigmoidal coherence-to-gain mapping which is controlled by an online estimate of the present coherence statistics. The algorithm automatically adapts to a given acoustic environment and provides a stronger dereverberation effect than the original method presented in Allen *et al.* (1977) in most acoustic conditions. The performance of the proposed algorithm was objectively and subjectively evaluated in terms of its impacts on the amount of reverberation and overall quality. For comparison, a binaural spectral subtraction method, based on Lebart *et al.* (2001), and a binaural version of Allen *et al.*'s original method were considered as reference systems. The results revealed that the proposed coherence-based approach is most successful in acoustic scenarios that exhibit a significant spread in the coherence distribution where direct sound and reverberation can be segregated. This dereverberation algorithm is thus particularly useful in large rooms for nearby source-receiver distances.

¹Based on Westermann *et al.* (2013)

A.1 Introduction

When communicating inside a room, the speech signal is accompanied by multiple reflections originating from the surrounding surfaces. The impulse response of the room is characterized by early reflections (first 50-80 ms of the room response) and late reflections or reverberation (Kuttruff, 2000).

In terms of auditory perception, early reflections mainly introduce coloration (Salomons, 1995), are beneficial for speech intelligibility (Bradley *et al.*, 2003) and are typically negligible with regard to sound localization (Blauert, 1996). In contrast, reverberation smears the temporal and spectral features of the signal which commonly deteriorates speech intelligibility (Moncur and Dirks, 1967), listening comfort (Ljung and Kjellberg, 2010) and localization performance. Some of the above negative effects are partly compensated for in normal-hearing listeners by auditory mechanisms such as the precedence effect (Litovsky *et al.*, 1999), monaural/binaural de-coloration and binaural dereverberation (e.g., Zurek, 1979; Blauert, 1996; Buchholz, 2007). However, in hearing-impaired listeners, reverberation can be detrimental because of reduced hearing sensitivity as well as decreased spectral and/or temporal resolution (e.g., Moore, 2012). In addition, a hearing impairment may affect the auditory processes that otherwise help listening in reverberant environments (e.g., Akeroyd and Guy, 2011; Goverts *et al.*, 2001). Thus, suppressing reverberation by utilizing a dereverberation algorithm, e.g. in hands-free devices, binaural telephone headsets and digital hearing aids, might improve speech intelligibility, localization performance and ease of listening.

Several dereverberation algorithms have been proposed in the literature. They address either early reflections or reverberation, are blind or non-blind, or use single or multiple input channels. Typical methods for suppressing early reflections include inverse filtering (e.g., Neely and Allen, 1979; Mourjopoulos, 1992) and linear prediction residual processing (e.g., Gillespie *et al.*, 2001; Yegnanarayana *et al.*, 1999). Processing methods for suppressing reverberation are typically based on spectral enhancement techniques which decompose the speech signal in time and frequency and suppress components which are estimated to be mainly reverberant. Different approaches have been proposed to realize this estimation.

Allen *et al.* (1977) proposed a binaural approach where gain factors are determined by the diffuseness of the sound field between two spatially separated microphones. They suggested two methods for calculating gain factors, one of which represented the coherence function of the two channels. However, because of a cophase-and-add stage, which combined the binaural channels, only a monaural output was provided. Kollmeier *et al.* (1993) extended the original approach of Allen *et al.* (1977) by applying the original coherence gain factor separately to both channels, thus providing a binaural output. Jeub and Vary (2010) demonstrated that synchronized spectral weighting across binaural channels is important for preserving binaural cues. In Simmer *et al.* (1994), a coherence-based Wiener filter was suggested which estimates the reverberation noise from a model of coherence between two points in a diffuse field. Their method was further refined in McCowan and Boulard (2003) and Jeub and Vary (2010) where acoustic shadow effects from a listener’s head and torso were included.

Single-channel spectral enhancement techniques employ different methods for reverberation noise estimation. Wu and Wang (2006) proposed that the reverberation noise can be estimated in the time-frequency domain from the power spectrum of preceding speech. Lebart *et al.* (2001) assumed an exponential decay of reverberation with time. In their model, the signal-to-reverberation noise ratio in each time-frame is determined by the energy in the current frame compared to that of the previous. Common problems with these methods are the so-called ”musical noise” effects and the suppression of signal onsets, both caused by an overestimation of the reverberation noise. Tsilfidis and Mourjopoulos (2009) introduced a gain-adaptation technique that incorporates knowledge of the auditory system to suppress musical noise. They also proposed a power relaxation criterion to maintain signal onsets. Alternative modifications based on the signal direct-to-reverberant energy ratio (DRR) have been proposed by Habets (2010). An overview of dereverberation methods can be found in Naylor and Gaubitch (2010).

In the present study, a binaural dereverberation algorithm is introduced, which utilizes the properties of the interaural coherence (IC), inspired by the concepts introduced in Allen *et al.* (1977). Applying the method of Allen *et al.* (1977) to different acoustic scenarios revealed that the dereverberation performance strongly varied between scenarios. In order to better understand this behavior, an investigation of the IC in different acoustic scenarios was performed, showing how IC distributions varied over frequency as a function of distance and reverberation time. Since the linear coherence-to-gain

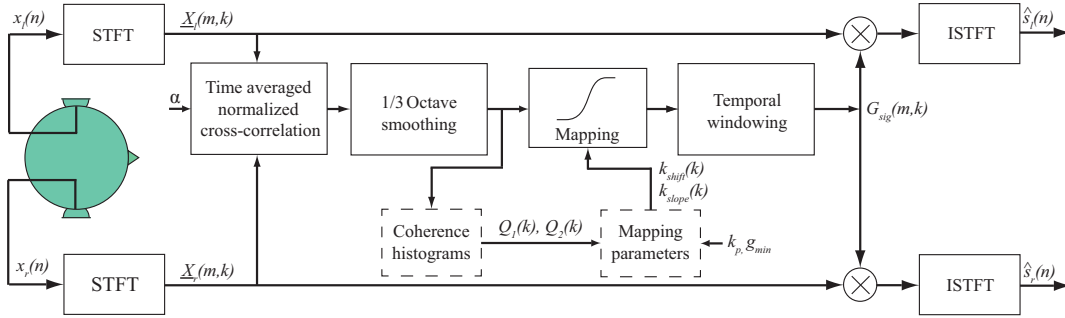


FIGURE A.1: Block diagram of the proposed signal processing method. The signals recorded at the ears, $x_l(n)$ and $x_r(n)$, are transformed via the STFT to the time-frequency domain, resulting in $X_l(m,k)$ and $X_r(m,k)$. The IC is calculated for each time-frequency bin and third-octave smoothing is applied. Statistical long-term properties of the IC are used to derive parameters of a sigmoidal mapping stage. The mapping is applied to the IC to realize a coherence-to-gain relationship and subsequent temporal windowing is performed. The derived gains (or weights) are applied to both channels $X_l(m,k)$ and $X_r(m,k)$. The dereverberated signals, $\hat{s}_l(n)$ and $\hat{s}_r(n)$, are reconstructed by applying an inverse SFTF.

mapping of the previous coherence-based methods (such as Allen *et al.* (1977)) can not account for this behavior, a non-linear sigmoidal coherence-to-gain mapping is proposed here, which is controlled by an online estimate of the inherent coherence statistics in a given acoustical environment. In this way, frequency-specific processing- and weighting characteristics are applied that result in an improved dereverberation performance, especially in acoustic scenarios where the coherence varies strongly over time and frequency. The performance of the proposed algorithm is evaluated objectively and subjectively, assessing the amount of reverberation and overall signal quality. The performance is compared to two reference systems, a binaural spectral subtraction method, inspired by Lebart *et al.* (2001), and a binaural version of the original method of Allen *et al.* (1977).

A.2 The coherence-based dereverberation algorithm

A.2.1 Signal processing

The signal processing of the proposed binaural dereverberation method is illustrated in Fig. A.1. Two reverberant time signals, recorded at the left and right ear of a person or a dummy head, $x_l(n)$ and $x_r(n)$, are transformed to the time-frequency domain using the Short-Time Fourier Transform (STFT) (Allen and Rabiner, 1977). This results in the complex-valued short-term spectra $\underline{X}_l(m,k)$ and $\underline{X}_r(m,k)$, where m denotes the

time frame and k the frequency band. For the STFT, a Hanning window of length L (including zero-padding of length $L/2$) and a 75 % overlap (i.e., applying a time shift of $L/4$ samples) between successive windows are used. For each time-frequency bin, the absolute value of the interaural coherence (IC or coherence from here) is calculated and third-octave smoothing is applied (Hatziantoniou and Mourjopoulos, 2000). A sigmoidal mapping stage is subsequently applied to the coherence estimates to realize a coherence-to-gain mapping. This mapping realizes a time-varying filter that attenuates time-frequency regions with a low IC (i.e., that are strongly affected by reverberation) and leaves regions untouched with high IC (i.e., where the direct sound is dominant). The parameters of the sigmoidal coherence-to-gain mapping are calculated based on an online estimate of the statistical properties of the IC (i.e., applying frequency-dependent coherence histograms). In order to suppress potential aliasing artifacts that may be introduced by applying this filtering process, temporal windowing is applied (Kates, 2008). This is realized by applying an inverse STFT to the derived filter gains and then truncating the resulting time-domain representation to a length of $L/2+1$. This filter response is then zero-padded to a length of L and another STFT is performed. The resulting filter gains are applied to both channels $\underline{X}_l(m, k)$ and $\underline{X}_r(m, k)$. The dereverberated signals, $\hat{s}_l(n)$ and $\hat{s}_r(n)$, are finally reconstructed by applying the inverse STFT and then adding the resulting (overlapping) signal segments (Allen and Rabiner, 1977).

A.2.2 Signal decomposition and coherence estimation

From the time-frequency signals $\underline{X}_l(m, k)$ and $\underline{X}_r(m, k)$, the IC is calculated as:

$$C_{lr}(m, k) = \frac{|\Phi_{lr}(m, k)|}{\sqrt{\Phi_{ll}(m, k)\Phi_{rr}(m, k)}} , \quad (\text{A.1})$$

with $\Phi_{ll}(m, k)$, $\Phi_{rr}(m, k)$ and $\Phi_{lr}(m, k)$ representing the exponentially-weighted short-term cross-correlation and auto-correlation functions:

$$\begin{aligned} \Phi_{ll}(m, k) = & \alpha |\underline{X}_l(m, (k-1))|^2 \\ & + |\underline{X}_l(m, k)|^2 \end{aligned} \quad (\text{A.2})$$

$$\begin{aligned} \Phi_{rr}(m, k) = & \alpha |\underline{X}_r(m, (k-1))|^2 \\ & + |\underline{X}_r(m, k)|^2 \end{aligned} \quad (\text{A.3})$$

$$\begin{aligned} \Phi_{lr}(m, k) = & \alpha \underline{X}_r(m, (k-1)) \cdot \underline{X}_l^*(m, (k-1)) \\ & + \underline{X}_r(m, k) \underline{X}_l^*(m, k) \end{aligned} \quad (\text{A.4})$$

where α is the recursion constant and $*$ indicates the complex conjugate. These coherence estimates yield values between 0 (for fully incoherent signals) and 1 (for fully coherent signals). If the time window applied in the STFT exceeds the duration of the room impulse responses (RIR) between a sound source and the two ears, the coherence approaches unity (Jacobsen and Roisin, 2000). When shorter time windows than the duration of the involved RIRs are applied in the STFT (which is typically the case), the estimated coherence is highly influenced by the used window length (Scharrer, 2010).

The recursion constant α determines the temporal integration time τ of the coherence estimate, which is given by:

$$\tau = -\frac{L}{4f_s \cdot \ln(\alpha)}, \quad (\text{A.5})$$

where f_s is the sampling frequency. The integration time needs to be short enough to follow the changes in the involved signals (i.e., speech), but long enough to provide reliable coherence estimates. In this study, an STFT window length of 6.4 ms (identical to that of Allen *et al.*, 1977 and corresponding to 282 samples) and a recursion constant of $\alpha = 0.97$ (corresponding to a time constant $\tau \approx 100\text{ms}$) are used. The applied time constant is similar to the ones used in previous work (e.g., Kollmeier *et al.*, 1993) and is able to follow syllabic changes.

A.2.3 Coherence-to-gain mapping

In order to cope with the different frequency-dependent distributions of the IC observed in different acoustic scenarios (see Sec. A.3), a coherence-distribution dependent

coherence-to-gain mapping is introduced. This is realized by a sigmoid function whose parameters are controlled by an (online) estimate of the statistical properties of the IC in each frequency channel. The resulting filter gains are:

$$G_{\text{sig}}(m, k) = \frac{(1 - g_{\text{min}})}{1 + e^{-k_{\text{slope}}(k)(C_{LR}(m, k) - k_{\text{shift}}(k))}} + g_{\text{min}}, \quad (\text{A.6})$$

where k_{slope} and k_{shift} control the sigmoidal slope and the position. The minimum gain g_{min} is introduced to limit signal processing artifacts associated with applying infinite attenuation.

In order to calculate the frequency-dependent parameters of the sigmoidal mapping function, coherence samples for a duration, defined by t_{sig} , are gathered in a histogram. For constant source-receiver location, t_{sig} of several seconds was found to provide a good compromise between stable parameter estimates and as short as possible adaptation time. For moving sources and changing acoustic environments, the method for updating the sigmoidal parameters might need revision.

A coherence histogram (shown as a Gaussian distribution for illustrative purposes) is exemplified in Fig. A.2 (gray curve) together with the corresponding 1st (Q_1) and 2nd (Q_2 or median) quartile. An example sigmoidal coherence-to-gain mapping function is represented by the black solid curve. The linear mapping function applied by Allen *et al.*, 1977 is indicated by the black dashed curve. When applying a linear mapping, the gain (given by C_{lr}) is smoothly turned down with decreasing IC (i.e., increasing amount of reverberation) and thus, almost all samples are attenuated to a certain degree. In contrast, the sigmoidal mapping strongly suppresses samples with low IC (which is only limited by g_{min}) and leaves samples with higher IC untouched. In this way a much stronger suppression of reverberation is achieved.

The degree of processing is determined by k_p which directly controls the slope of the sigmoidal mapping. The parameters k_{slope} and k_{shift} of the sigmoidal mapping are derived by inserting the two points $G_{\text{sig}|C_{lr}=Q_1} = g_{\text{min}} + k_p$ and $G_{\text{sig}|C_{lr}=Q_2} = 1 - k_p$ into Eq. A.6 and then solving the resulting two equations for k_{slope} and k_{shift} (see Fig. A.2),

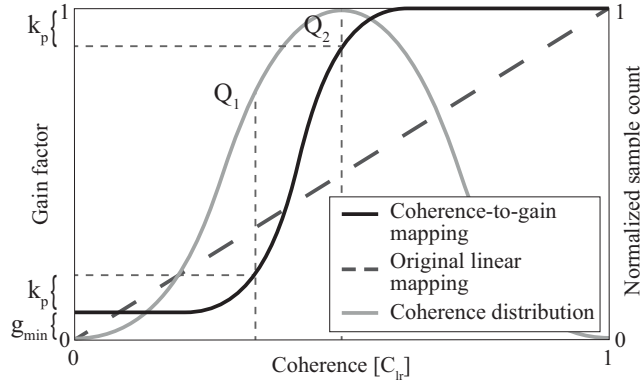


FIGURE A.2: Idealized IC histogram distribution in one frequency-channel (gray curve). The coherence-to-gain relationship in the specific channel is calculated to intersect $\varsigma(Q_1) = g_{min} + k_p$ and $\varsigma(Q_2) = 1 - k_p$. Thereby, G_{min} denotes the maximum attenuation and k_p determines the processing degree.

i.e.:

$$k_{\text{shift}}(k) = \left(\frac{\ln(G_{\text{sig}|C_{lr}=Q_1})^{-1}}{\ln(G_{\text{sig}|C_{lr}=Q_2})^{-1}} Q_2(k) + Q_1(k) \right) \cdot \left(1 - \frac{\ln(G_{\text{sig}|C_{lr}=Q_1})^{-1}}{\ln(G_{\text{sig}|C_{lr}=Q_2})^{-1}} \right)^{-1} \quad (\text{A.7})$$

$$k_{\text{slope}}(k) = \frac{\ln(G_{\text{sig}|C_{lr}=Q_1}) - 1}{Q_1(k) - k_{\text{shift}}} , \quad (\text{A.8})$$

where $Q_1(k)$ and $Q_2(k)$ are estimated in each frequency channel as the 1st and 2nd quartile of the measured coherence histograms and g_{min} and k_p are predetermined parameters. Following such approach, k_p provides the only free parameter, which directly controls the slope of the sigmoidal function and thus, determines the degree (or aggressiveness) of the dereverberation processing.

For speech presented in an auditorium with source-receiver distances of 0.5 m and 5 m (see Sec A.3), examples of sigmoidal mappings are shown in Fig. A.3 for different values of k_p in the 751.7 Hz frequency channel. It can be seen that the coherence-to-gain function steepens as k_p increases (i.e. as the processing degree increases). In addition, as the distribution broadens (from 5 m to 0.5 m) the slope of the coherence-to-gain function decreases. Hence, in contrast to the original coherence-based dereverberation approach in Allen *et al.* (1977), which considered a fixed linear coherence-to-gain mapping (Fig. A.2, dashed line), the proposed approach provides a flexible mapping function, which can be adjusted by the parameter k_p to any given acoustic condition.

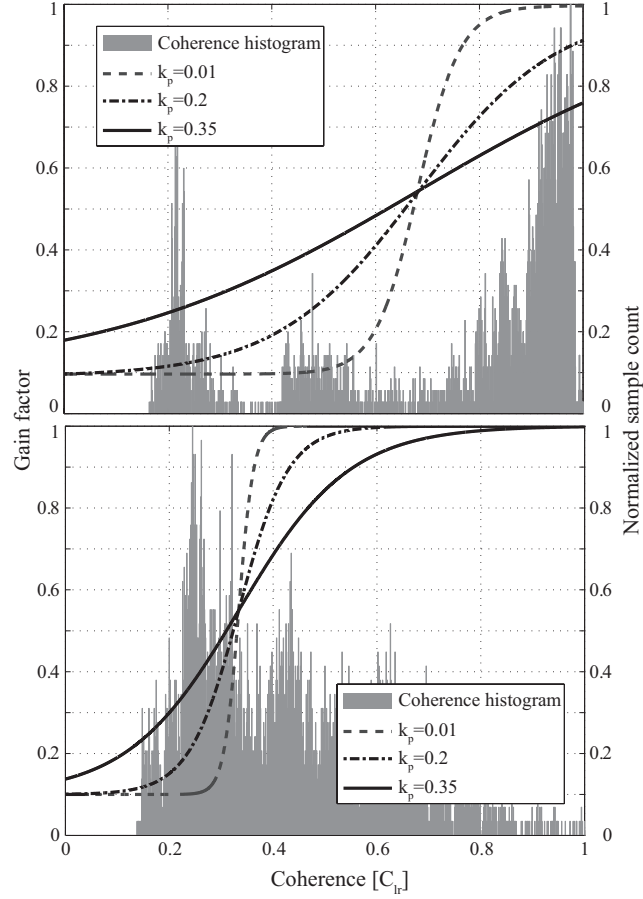


FIGURE A.3: IC histogram of speech presented in an auditorium with 0.5 m (top panel) and 5 m (bottom panel) source-receiver distance. Sigmoidal coherence-to-gain relationship for three different processing degrees of k_p are shown.

A.2.4 Reference systems

In order to compare the performance of the proposed algorithm to the state-of-the-art algorithms described in the relevant literature, two additional dereverberation methods were implemented: The IC-based algorithm proposed by Allen *et al.* (1977) and the spectral subtraction based algorithm described by Lebart *et al.* (2001). In order to allow a fair comparison, both methods were incorporated in the framework shown in Fig. A.1 and, thus, extended to providing a binaural output. Hence, the following three processing schemes were considered:

- i. The proposed coherence-based approach for three different values of k_p (see Tab. A.1 for processing parameters). The different values for k_p (i.e., the processing degree) were chosen to investigate the performance of the algorithm throughout the entire parameter range ($0 \leq k_p \leq (1 - g_{\min})/2$).

TABLE A.1: Processing parameters used for the proposed algorithm.

Parameter	Symbol	Value
Sampling frequency	f_s	44.1 kHz
Frame length	L	6.4 ms
Frame overlap		75%
Recursion constant	α	0.97
Gain threshold	G_{min}	0.1
Processing degrees	k_p	{0.01; 0.2; 0.35}
Sigmoidal updating time	t_{sig}	3 s

- ii. The method described by Allen *et al.* (1977) with a binaural extension according to Kollmeier *et al.* (1993). Hence, the IC (Eq. A.1) was directly applied as a weight to each time-frequency bin of the left and right channel. To allow a comparison with the proposed algorithm, third-octave smoothing and temporal windowing (Sec. A.2.1) were added. Hence, the same processing as shown Fig. A.1 was applied except that the sigmoidal coherence-to-gain mapping was replaced by a straight-line (linear) mapping (see Fig. A.3, dashed-dotted line). The same recursion constant and window length as in the first algorithm (i.) were used.
- iii. A binaural extension of the spectral subtraction approach described by Lebart *et al.* (2001). This approach relies on the estimation of reverberation noise in speech based on a model of the room impulse response (RIR). This model was derived from an estimation of the reverberation time. The binaural extension was realized by (i) averaging the reverberation time estimates for the left and right channel and (ii) synchronizing the spectral weighting in both channels. The latter was realized by calculating the weights for the left and right channel in each time-frequency bin and then applying the minimum value to both channels. The original processing parameters of Lebart *et al.* (2001) were used.

A.3 Evaluation methods

In order to evaluate the performance of the proposed dereverberation algorithm, objective as well as subjective measures were applied. Reverberant speech was created by convolving anechoic speech with binaural room impulse responses (BRIRs), recorded at 0.5 m and 5 m distances in an auditorium (see Appendix). The auditorium had a reverberation time of $T_{60} = 1.9$ s at 2 kHz and DRRs of -9.34 dB and -28 dB, respectively. Two anechoic sentences from the Danish speech database, recorded by Christiansen and

Henrichsen (2011), were used, each spoken by both a male and a female talker, resulting in two sentences for each position.

A.3.1 Objective evaluation methods

Several metrics have been suggested to predict the performance and quality of dereverberation algorithms (Kokkinakis and Loizou, 2011; Goetze *et al.*, 2010; Naylor and Gaubitch, 2010). Two commonly used objective measures were applied here to evaluate different aspects of the proposed dereverberation algorithm.

A.3.1.1 Signal-to-reverberation ratio

The segmental signal-to-reverberation (segSRR) ratio estimates the amount of direct signal energy compared to reverberant energy (e.g., Wu and Wang, 2006; Tsilfidis and Mourjopoulos, 2011) and was given by

$$\text{segSRR} = \frac{10}{K} \log_{10} \left[\frac{\sum_{n=kN}^{kN+N-1} (k_{\text{path}} s_d(n))^2}{\sum_{n=kN}^{kN+N-1} (k_{\text{path}} s_d(n) - \hat{s}(n))^2} \right], \quad (\text{A.9})$$

where $s_d(n)$ denotes the direct path signal, $\hat{s}(n)$ the (reverberant) test signal, k_{path} is a normalization constant, N the frame-length (here 10 ms), $k = 0 \dots W - 1$ and W the total number of frames. The direct sound was derived by convolving the anechoic speech signal with a modified (time-windowed) version of the applied BRIR, which only contained the direct sound component. The denominator provides an estimate of the reverberation energy by subtracting the waveform of the direct sound from the waveform of the tested signal (which includes the direct sound). The improvement in SRR was then calculated by:

$$\Delta \text{segSRR} = \text{segSRR}_{\text{proc}} - \text{segSRR}_{\text{ref}}. \quad (\text{A.10})$$

Thereby, $\text{segSRR}_{\text{ref}}$ was calculated from the original reverberant speech signal by convolving the anechoic speech with a given BRIR. The $\text{segSRR}_{\text{proc}}$ was calculated from the same reverberant speech signal but processed by the considered dereverberation algorithm. Hence, an algorithm that successfully suppresses reverberation should achieve SRR improvements of $\Delta \text{segSRR} > 0$ dB.

Since time-based quality measures, such as the segSRR, are sensitive to any applied normalization, all signals were normalized to equal root mean square (RMS) levels before the actual segSRR was calculated. In addition, the level of the direct path signal was multiplied by the factor k_{path} in such a way that the energy in the direct path was equal to the direct path component of the processed signal. The appropriate k_{path} was determined numerically by minimizing the denominator in Eq. A.9 for the case that the unprocessed (reference) reverberant signal was applied. Only frames with $\text{segSRR}_k < -10$ dB were included in calculating the total segSRR from Eq. A.9. This was done since the segSRR measure would otherwise be dominated by frames that mainly contain direct sound energy while frames that mainly contain reverberation energy provide only a minor contribution.

A.3.1.2 Noise-mask ratio

The noise-mask ratio (NMR) is often used as an objective measure for evaluating the sound quality produced by dereverberation methods (e.g., Furuya and Kataoka, 2007; Tsilfidis *et al.*, 2008). The measure is related to human auditory processing as only audible noise components (or artifacts) are considered. According to Brandenburg (1987), the NMR is defined as:

$$\text{NMR} = \frac{10}{W} \sum_{i=0}^{W-1} \log_{10} \frac{1}{B} \sum_{b=0}^{B-1} \frac{1}{C_b} \frac{\sum_{\omega=\omega_{lb}}^{\omega=\omega_{hb}} |R(\omega, m)|^2}{T_b(m)} , \quad (\text{A.11})$$

with W denoting the total number of frames, B the number of critical bands (or auditory frequency channels) and C_b the number of frequency bins inside the critical band with index b . The power spectrum of the reverberation, $|R(\omega, m)|^2$, was calculated by subtracting the power spectrum of the anechoic signal from that of the test signal where ω is the angular frequency and m is the time frame. The upper and lower cut-off frequencies were given by ω_{hb} and ω_{lb} , respectively, and the masked threshold by $T_b(m)$, which depends on the spectral magnitude in the b 'th critical band (for details see Brandenburg, 1987). The difference between the reverberant (reference) and processed NMR was then defined as:

$$\Delta\text{NMR} = \text{NMR}_{\text{proc}} - \text{NMR}_{\text{ref}} . \quad (\text{A.12})$$

As the amount of audible noise increases (i.e. NMR_{proc} decreases) the resulting ΔNMR decreases. Thus, smaller values of ΔNMR indicate a quality improvement.

A.3.2 Subjective evaluation methods

A subjective evaluation method similar to the Multiple Stimuli with Hidden reference test (MUSHRA) was applied to subjectively evaluate the performance of the different dereverberation algorithms (see RBS.1534-2001:, 2003). These types of experiments have been widely applied to efficiently extract specific signal features even in cases where differences are very subtle (e.g., Lorho, 2010). A graphical user interface (GUI) was presented to the subjects to judge the attributes "amount of reverberation" and "overall quality" on a scale from 0-100 with descriptive adjectives: "very little", "little", "medium", "much", and "very much". The subjects could switch between six different processing methods: the original IC-based method, the proposed IC-based method with $k_p = 0.01, 0.2$, and 0.35 , the spectral subtraction method, and an anchor. Anchors are an inherent trait of MUSHRA experiments to increase the reproducibility of the results and to prevent contraction bias (e.g., Bech and Zacharov, 2006). Additionally, subjects had access to the reference (unprocessed) stimulus via a "Reference button". Two different source-receiver positions (0.5 m and 5 m) were considered and each condition was repeated once. For an intuitive comparison with the objective evaluation results, the subjective scores were transformed to 100 - scores. The resulting scores were named "Strength of dereverberation" and "Overall loss of quality".

To evaluate the quality of speech, the anchor was realized by distorting the reference signal using an Adaptive Multi-Rate (AMR) speech coder (available from 3GPP TS26.073, 2008) with a bit-rate of 7.95 kbits/sec. The resulting distortions were similar to the artifacts produced by the different dereverberation methods. Anchors for judging the amount of reverberation were created by applying a temporal half cosine window with a length of 600 ms to the BRIRs and thereby artificially reducing the resulting reverberation while keeping direct sound and early reflections. The unprocessed reference stimulus was not included as a hidden anchor because pilot experiments showed that this resulted in a significant compression bias of the subjects' responses (for further details, see Bech and Zacharov, 2006). All experiments were carried out in a double-walled sound insulated booth, using a MATLAB GUI, Sennheiser HD-650 circumaural headphones and a computer with a RME DIGI96/8 PAD high-end sound card. The measurement setup

was calibrated to produce a sound pressure level of 65 dB, measured in an artificial ear coupler (B&K 4153).

Ten (self-reported) normal-hearing subjects participated in the experiment. All subjects were either Engineering Acoustics students or sound engineers and were considered as experienced listeners. An instruction sheet was handed out to all subjects. Prior to the test, a training session was carried out to introduce the GUI and the applied terminology. There was no time limit for the experiment but, on average, the subjects required 1 hour to complete the experiment.

A.4 Results

A.4.1 Effects of reverberation on speech in different acoustic environments

A.4.1.1 Spectrogram representations

The effects of reverberation on speech in a room are shown in the spectrograms in Fig. A.4. The anechoic speech sample for a male speaker is shown in panel (a). The anechoic signal, convolved with one channel of a BRIR recorded in an auditorium at a 0.5 m distance (see Sec A.3) is shown in panel (b). A comparison of panel (a) and (b) reveals that a large number of the dips in the anechoic speech representation are filled due to the reverberation, i.e., the reverberation leads to a smearing both in the temporal and spectral domain.

A.4.1.2 Interaural coherence

The lowest levels of coherence exist in an isotropic diffuse sound field, where the coherence measured between two points is given by a sinc-function:

$$C_{ideal} = \frac{\sin(2\pi f \frac{d_{mic}}{c})}{2\pi f \frac{d_{mic}}{c}} , \quad (A.13)$$

with c representing the speed of sound and d_{mic} the distance between the two points (Martin, 2001). In such a case, the coherence approaches unity at low frequencies and exhibits zero-crossings at frequencies corresponding to the distance between the two measurement points, as indicated by the solid curve in Fig. A.5. A similar behavior

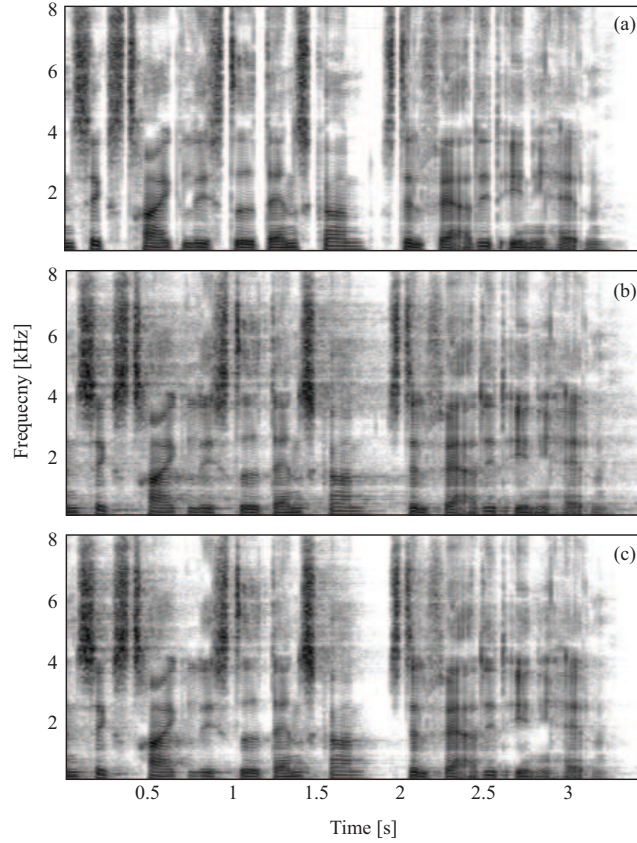


FIGURE A.4: Spectrograms illustrating the effects of reverberation and dereverberation on speech. Panel (a) shows the anechoic input signal. In panel (b), the speech is convolved with one channel of a BRIR measured in an auditorium at a distance of 0.5 m. Panel (c) shows the effects of the proposed dereverberation processing.

is found for the IC but altered by the interference of the torso, head, and pinna of a listener (Jeub *et al.*, 2009).

Figure A.5(a) shows IC histograms for speech presented in a reverberation chamber, calculated from the binaural recordings of Hansen and Munch (1991). The algorithm defined in Sec A.2.1 was first applied to describe the short-term (6.4 ms) IC of the binaural representation of an entire sentence spoken by a male talker. From the resulting IC values, the coherence histograms were derived. Gray scale reflects the number of occurrences (height of the histogram) in a given frequency channel. As expected from the ideal diffuse sound field, an increased coherence is observed below 1 kHz. Above 1 kHz, most coherence values are between 0.1 and 0.3. The lower limit of the obtained IC values and the IC spread of the distribution are caused by the non-stationarity of the input speech signal and the temporal resolution of the coherence estimation (i.e., the window length L and the recursion constant α).

Figure A.5(b) - A.5(d) shows example coherence histograms for 0.5 m, 5 m and 10 m source-receiver distances in an auditorium with a reverberation time of $T_{60} = 1.9$ s at 2 kHz and a volume of 1150 m³ (see Appendix for recording details). The overall coherence decreases with increasing distance between the source and the receiver. This results from the decreased direct-to-reverberant energy ratio at longer source-receiver distances. At very small distances (Figure A.5(b)), most coherence values are close to one indicating that mainly direct sound energy is present. In addition, the coherence arising from the diffuse field (with values between 0.1 and 0.3) is separable from that arising from the direct sound field. For the 5 m distance, substantially fewer frames with high coherence values are observed. This is because frames containing direct sound information are now affected by reverberation and there is no clear separability anymore between frames with direct and reverberant energy. At a distance of 10 m, this trend continues as the coherence values further drop and the distribution resembles that found in the diffuse field, i.e., very little direct sound is available.

For small source-receiver distances, where the direct sound is separable from the diffuse sound field, a dereverberation algorithm that directly applies the short-term coherence as a gain (i.e., applying a linear coherence-to-gain mapping as proposed by Allen *et al.*, 1977) should suppress reverberant time-frequency segments and preserve direct sound elements. However, with increasing source-receiver distance, the effectiveness of such an algorithm can be expected to decrease, since direct sound elements will be increasingly "contaminated" by diffuse reverberation. Moreover, the observed different coherence histograms suggest that the optimal coherence-to-gain mapping depends on frequency and the specific acoustic condition. Since the dereverberation algorithm proposed in Allen *et al.* (1977) applies a fixed coherence-to-gain mapping, it can only provide a significant suppression of reverberation in very specific acoustic conditions. In addition, because of the limited coherence range at lower frequencies (where all IC values are rather high), a linear coherence-to-gain relationship would result in a high gain at lower frequencies for all acoustical conditions and would effectively act as a low-pass filter.

A.4.2 Effects of dereverberation processing on speech

The spectrogram shown in Fig. A.4(c) illustrates the effect of dereverberation on speech. The proposed algorithm was applied with a moderate processing degree (i.e., $k_p = 0.2$).

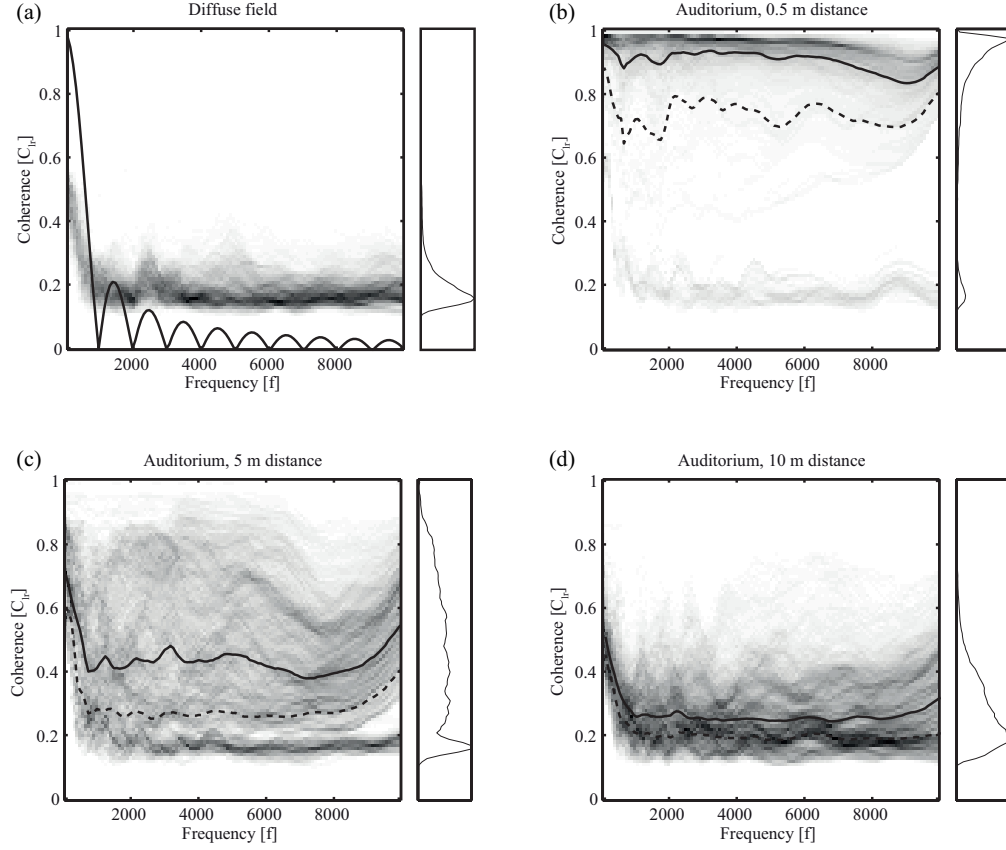


FIGURE A.5: (a) Coherence histograms of speech presented in a diffuse field as a function of frequency shown together with the ideal result as the black line. Sum across frequencies shown in side panel. (b-d) Similar histogram plots for an auditorium at different distances. The solid line indicates Q_2 or median and the dotted line Q_1 of the IC distribution.

It can be seen that a substantial amount of the smearing caused by the reverberation in the room (panel (b)) was reduced by the dereverberation processing.

A.4.2.1 Signal-to-reverberation ratio

Figure A.6 (gray bars) shows the signal-to-reverberation ratio, Δ_{segSRR} (Eq. A.10), for the different processing schemes. All algorithms show a significant reduction in the amount of reverberation (i.e., all exhibit positive values).

For the 0.5 m distance (left panel), the proposed algorithm (for $k_p = 0.2$) provides the best performance. For the lowest degrees of processing ($k_p = 0.35$), the performance is slightly below that attained for the spectral subtraction algorithm. For the 5 m distance (right panel), the proposed method for the highest processing degree ($k_p = 0.01$), performs comparably with the spectral subtraction method. As expected, the performance

of the proposed method generally drops with decreasing processing degree (i.e., increasing k_p value). The original IC-based method generally shows the poorest performance and provides essentially no reverberation suppressions in the 0.5 m condition.

A.4.2.2 Noise-mask ratio

In Fig. A.6, ΔNMR (white bars) is shown, where smaller values correspond to less audible noise or better sound quality. For the different processing conditions, the original IC-based approach shows the best overall performance for both source-receiver distances. Considering the very small amount of dereverberation that is provided by this algorithm (see Sec. A.4.2.1 and Fig. A.6), this observation is not surprising since the algorithm only has a minimal effect on the signal. The performance of the proposed method for high degrees of processing (i.e., $k_p = 0.01$) is similar or slightly better than that obtained with the spectral subtraction approach. For decreasing degrees of processing (i.e., $k_p = 0.2$ and 0.3) the performance of the proposed method increases but, at the same time, the strength of dereverberation (as indicated by segSRR) also decreases (see grey bars in Fig. A.6). Considering both measures, segSRR and the NMR, the proposed method is superior for close sound sources (i.e., the 0.5 m condition with $k_p = 0.2$) and exhibits performance similar to the spectral subtraction method for the 5 m condition.

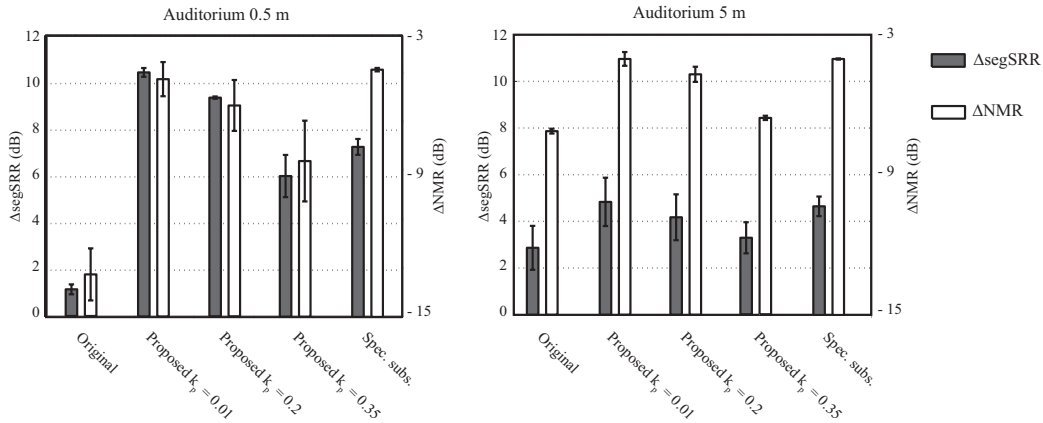


FIGURE A.6: ΔsegSRR (reverberation suppression) and ΔNMR (loss of quality) between the estimated clean signal and the equalized reverberant signal for different methods for the 0.5 m source-receiver distance (left panel) and 5 m source-receiver distance (right panel).

A.4.2.3 Subjective evaluation

The results from the subjective evaluation for each processing method are shown in Fig. A.7. For better comparison with the objective results, the measured data were inverted

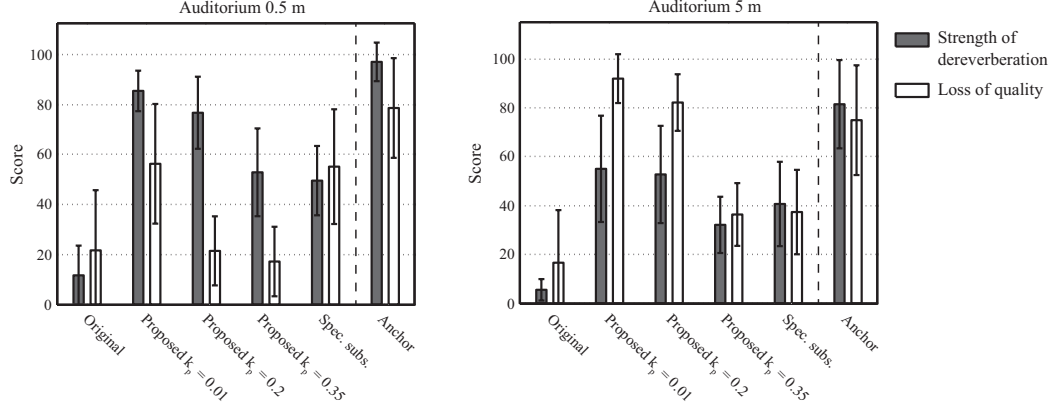


FIGURE A.7: The mean and standard deviation of subjective results judging “Strength of dereverberation” and “Overall loss of quality” for the 0.5 m source-receiver distance (left panel) and 5 m source-receiver distance (right panel).

(i.e., shown as $100 - \text{measured score}$). The attributes “amount of reverberation” and “overall quality” were consequently changed to “strength of dereverberation” and “loss of quality”. Considering the strength of dereverberation, indicated by the gray bars, the proposed approach exhibited the best performance for $k_p = 0.01$ at both distances. As the degree of processing decreases (i.e., for increasing values of k_p), the strength of dereverberation decreases. The improvement relative to the spectral subtraction approach is considerably higher for the 0.5 m distance (left panel) than for the 5 m distance (right panel). The original approach of Allen *et al.* (1977) produced the lowest strength of dereverberation for both source-receiver distances. The differences in scores between the original approach and the others were noticeably larger for the 0.5 m distance than for 5 m. This indicates that, for very close sound sources, the other methods are more efficient than the original IC approach.

The loss of quality of the signals processed with the proposed IC-based method were found to be substantially smaller for the 0.5 m condition than for the 5 m condition. This difference is not as large with the original approach as well as the spectral subtraction method, indicating that the proposed IC-based method is particularly successful for very close sound sources. As in the objective quality evaluation, increasing the degree of dereverberation processing (i.e., by decreasing k_p) results in a drop of the overall quality. However, this effect is not as prominent when decreasing k_p from 0.35 to 0.2 at the 0.5 m distance.

Considering both subjective measures, the proposed method with $k_p = 0.2$ clearly exhibits the best overall performance at the 0.5 m distance. Even when applying the

highest degree of processing (i.e., $k_p = 0.01$), the quality is similar to that obtained with spectral subtraction but the strength of dereverberation is substantially higher. For the 5 m distance, increasing the degree of processing has a negligible effect on the strength of dereverberation but is detrimental for the quality. However, for $k_p = 0.35$, the performance of the proposed method is comparable to that obtained with the spectral subtraction approach.

An analysis of variance (ANOVA) showed significance for the sample effect at source-receiver distances of 0.5 m [$F = 97.65$, $p < 0.001$] and 5 m [$F = 41.31$, $p < 0.001$]. No significant subject effect was found.

A.5 Discussion

According to the subjective results of the present study, the proposed method outperformed the two reference methods in all conditions. The original IC-based (reference) method proposed by Allen *et al.* (1977) did not provide any significant effect on the considered signals, but resulted in very low dereverberation scores but very high quality scores. The spectral-subtraction-based dereverberation method based on Lebart *et al.* (2001) generally provided a significant amount of dereverberation, but always reduced the overall quality. In particular, for the 0.5 m distance, the proposed method provided the strongest dereverberation effect as well as best quality for all processing degrees (k_p) that were considered. In the 5 m condition, the proposed method slightly outperformed the reference methods, both in terms of dereverberation and quality, but only for the lowest processing degree ($k_p = 0.35$).

The subjective evaluation method employed here is particularly sensitive to small differences between processing methods. However, the subjective data for the 0.5 m and 5 m conditions cannot directly be compared because they are presented with separate references. Due to the substantially different characteristics in the two conditions, a simultaneous presentation would result in scores at the end of the scale, which is known as compression bias (Zahorik *et al.*, 2005). For comparisons on an absolute scale, the objective measures applied here are more suitable.

When comparing the objective results between the 0.5 m and the 5 m conditions from Fig. A.6, the strength of dereverberation (i.e., segSRR) was generally higher in the nearer condition. In particular, the proposed method showed a better performance in the 0.5

m condition. In terms of quality loss (NMR difference), all algorithms performed better in the 0.5 m condition. There are two main reasons for the differences between the 0.5 m and 5 m conditions. First, at 0.5 m, where the DRR is substantially higher than at 5 m, the amount of required processing is lower, resulting in a signal of higher quality. Second, the high coherence arising from the direct sound and the early reflections is distinguishable from the diffuse sound-field with low coherence (Fig. A.7 panel (b)), i.e., a bimodal coherence distribution can be observed. Considering the narrow coherence distribution for the 5 m condition in Fig. A.5c, no high coherence values are present that clearly separate the direct and the diffuse field. Therefore, dereverberation using coherence information becomes less effective and produces more artifacts.

A good overall correspondence of the subjective and objective results was found (Sec. A.4.2). Considering the strength of dereverberation, the segSRR slightly underpredicted the effectiveness of the proposed approach when compared to the subjective results. A likely reason is that the subjects used cues for reverberation estimation that are not reflected in the objective measures. For instance, when using the original implementation of the segSRR without thresholding a very poor correlation with the subjective data was found. This is because the contribution from non-reverberant frames substantially alter the segSRR estimates. When the thresholding was introduced, the correspondence with the perceptual results increased dramatically. However, additional modifications or different methods need to be derived to achieve better correspondence between subjective and objective results. In the quality evaluation, the NMR seemed to overestimate the distortion and artifacts introduced by the proposed method at 0.5 m and to underestimate them at 5 m. Moreover, the subjects showed higher sensitivity to the distortions and artifacts produced by the proposed method than the NMR measure. In comparison, the anchor was rated with the lowest score (resulting in the highest values in Fig. A.7) in all conditions (except the proposed method at 5 m). This indicates that the subjects experienced the signal degradation introduced by the proposed method to be detrimental in these conditions. As pointed out by Tsilfidis and Mourjopoulos (2011), none of the quality measures (including the NMR measure) was developed to cope specifically with dereverberation and the artifacts introduced by such processing. Generally, none of the commonly applied objective measures are well correlated with subjective scores (Wen *et al.*, 2006).

From the results of the present study, it can be concluded that the effectiveness of the

proposed approach strongly depends on the coherence distribution in a given acoustical scenario and the applied coherence-to-gain mapping. The coherence estimation mainly depends on the window-length of the STFT analysis and the recursion constant α . The window-length represents the trade-off between time- and frequency-resolution. A frequency resolution consistent with literature was chosen here, but this could perhaps be optimized. The temporal resolution is reflected in the recursion constant α (Eq. A.5), which was here also chosen according to the relevant literature. Lowering the integration time (decreasing the recursion constant) increased the noisiness of the coherence estimates and resulted in higher limit for the lowest obtainable coherence values. This effectively reduces the processing range of the dereverberation algorithm and, thus, its effectiveness. If larger integration times were chosen, the spread of coherence would be lost, again reducing the effective processing range. An alternative approach, for instance, would be to change the recursion constant dynamically. As in dynamic-range compression (e.g., Kates, 2008), the concept of an attack time and release time could be adopted in order to improve the temporal resolution at signal onsets and decrease the resolution in case of signal decays.

The proposed coherence-to-gain mapping had a substantial effect on the performance both for dereverberation and quality (see Sec. A.3). For close source-receiver distances a large processing degree should be applied for best performance (e.g., $k_p = 0.01$). For larger distances the value of k_p should be increased. Hence, the k_p value should adapt based on source-receiver distance, which should be considered in future algorithm improvements. With reference to Fig. A.5, the average coherence across frequency seems to correlate well with source-receiver distance and thus, may be used as a measure for automatically adjusting the value of k_p . However, other source-receiver distance measures may be even more appropriate for controlling k_p (Vesa, 2009).

Roman and Woodruff (2011) investigated intelligibility with ideal binary masks (IBMs) applied to reverberant speech both in noise and concurrent speech. They found significant improvements in intelligibility especially when reverberation and noise were suppressed while early reflections were preserved. The IBMs however, require a priori information about the time-frequency representation of the reverberation and noise. For very low values of k_p and narrow distributions of IC the mapping steepens and it resembles a binary mask. In future studies, IC could be a measure for determining time-frequency frames in a binary mask framework.

Only the slope of the coherence-to-gain mapping was altered in the present study to minimize the number of free parameters. However, shifting the function may allow better tuning of the coherence-to-gain mapping and, thus, may further improve performance. This could be an effective addition to the processing proposed here. Furthermore, the shape of the mapping could be adapted based on the current coherence distribution. The sigmoidal parameters are currently updated at a rate of $t_{sig} = 3$ s. However, In some acoustic scenarios, the coherence distribution may change at a different rate. Hence, t_{sig} may need to be changed or controlled by a measure of the changes in the overall coherence statistics.

A.6 Summary and conclusion

An interaural-coherence based dereverberation method was proposed. The method applies a sigmoidal coherence-to-gain mapping function that is frequency dependent. The coherence-to-gain functions are controlled by an (online) estimate of the present interaural coherence statistics which allows an automatic adaptation to a given acoustic scenario. By varying the overall processing degree with the parameter k_p , a trade-off between the amount of dereverberation and sound quality can be adjusted. The objective measures segSRR and NMR were applied and compared to subjective scores associated with "amount of reverberation" and "overall quality", respectively. The objective and the subjective evaluation methods showed that, when a significant spread in coherence is provided by the binaural input signals, the proposed dereverberation method exhibits superior performance compared to existing methods both in terms of reverberation reduction and overall quality.

A.7 Appendix

A.7.1 Measuring binaural impulse responses

In order to evaluate the coherence as a function of source-receiver distance, binaural room impulse responses (BRIRs) were recorded in an auditorium using a B&K Head and Torso Simulator (HATS) in conjunction with a computer running MatLab for playback and recording. The auditorium had a reverberation time of $T_{60} = 1.9$ s at 2 kHz and a volume of 1150 m³. The corresponding reverberation distance is 1.4 m (see Kuttruff, 2000).

A DynAudio BM6P 2-Way loudspeaker was used as the sound source. This speaker-type was chosen to roughly approximate the directivity pattern of a human speaker while providing an appropriate signal-to-noise ratio. The BRIRs were measured using logarithmic upward sweeps (for details see Müller and Massarani, 2001). Anechoic speech samples with a male speaker (taken from Hansen and Munch, 1991) were convolved with the BRIRs to simulate reverberant signals.

A.7.2 Acknowledgments

The authors would like to thank Dr. A. Tsilfidis (University of Patras, Greece) for his contribution to the evaluation of the method. This work was supported by an International Macquarie University Research Excellence Scholarship (iMQRES) and Widex A/S.

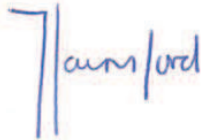
Appendix B

Ethics approvals



126 Greville Street
Chatswood NSW 2067
Australia
T (02) 9412 6872
F (02) 9411 8273
www.nal.gov.au

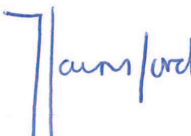


Australian Hearing Human Research Ethics Committee APPROVAL FOR RESEARCH INVOLVING HUMAN SUBJECTS	
APPROVAL NUMBER: AHHREC2012-13	
PROJECT NUMBER	PhD12.1
PROJECT TITLE	Release from masking through spatial separation in distance
CLASSIFICATION	Class1: Project with negligible risk
PRINCIPAL INVESTIGATORS	Adam Westermann, Jörg Buchholz, Virginia Best
DATE APPROVED/RATIFIED	17/5/2012
APPROVAL METHOD	Approved by the Research Director plus one other uninvolved senior NAL scientist as a Class 1 project with negligible risk.
<p>This approval is based on the information contained in the ethics application that was presented to the Research Director on 1/5/2012. A duplicate set of the documents is enclosed for your record.</p> <p>In compliance with the "National Statement on Ethical Conduct in Human Research" (2007), Annual reporting to the Committee on progress of the project is required. This will happen in March or September of each year and you will be reminded near the time.</p> <p>The Committee expects to be notified of any changes to the approved protocol or other issues that may have an impact on the ethics of the project either by means of the annual progress reports (checklists) or as an application for variation.</p> <p>All future correspondence relating to the ethical aspects of this project must quote the above Approval Number.</p>	
 Dr Tim Gainsford Operations & Finance Manager, NAL and AHHREC Secretary	



126 Greville Street
Chatswood NSW 2067
Australia
T (02) 9412 6872
F (02) 9411 8273
www.nal.gov.au



Australian Hearing Human Research Ethics Committee APPROVAL FOR RESEARCH INVOLVING HUMAN SUBJECTS	
APPROVAL NUMBER: AHHREC2013-1	
PROJECT NUMBER	PhD13.1
PROJECT TITLE	Informational masking in complex real-world environments
CLASSIFICATION	Class 1: project with negligible risk
PRINCIPAL INVESTIGATORS	Adam Westermann, Jörg Buchholz, Virginia Best
DATE APPROVED/RATIFIED	7 March 2013
APPROVAL METHOD	Approved by the Research Director plus one other uninvolved senior NAL scientist on 7 March 2013 as a Class 1 project with negligible risk.
<p>This approval is based on the information contained in the ethics application that was presented to the Research Director on 5 March 2013. A duplicate set of the documents is enclosed for your record.</p> <p>In compliance with the "National Statement on Ethical Conduct in Human Research" (2007), annual reporting to the Committee on progress of the project is required. This will happen in March or September of each year and you will be reminded near the time.</p> <p>The Committee expects to be notified of any changes to the approved protocol or other issues that may have an impact on the ethics of the project either by means of the annual progress reports (checklists) or as an application for variation.</p> <p>All future correspondence relating to the ethical aspects of this project must quote the above Approval Number.</p>	
 Dr Tim Gainsford Operations & Finance Manager, NAL and AHHREC Secretary	

Bibliography

- 3GPP TS26.073 (**2008**). “ANSI-C code for the Adaptive Multi Rate (AMR) speech codec”, Technical Report, 3rd Generation Partnership Project, Valbonne, France.
- Agus, T. R., Akeroyd, M. a., Gatehouse, S., and Warden, D. (**2009**). “Informational masking in young and elderly listeners for speech masked by simultaneous speech and noise.”, *Journal of the Acoustical Society of America* **126**, 1926–1940.
- Akeroyd, M. a., Gatehouse, S., and Blaschke, J. (**2007**). “The detection of differences in the cues to distance by elderly hearing-impaired listeners.”, *Journal of the Acoustical Society of America* **121**, 1077–1089.
- Akeroyd, M. A. and Guy, F. H. (**2011**). “The effect of hearing impairment on localization dominance for single-word stimuli”, *Journal of the Acoustical Society of America* **130**, 312–323.
- Allen, J. B., Berkley, D. A., and Blauert, J. (**1977**). “Multimicrophone signal-processing technique to remove room reverberation from speech signals”, *Journal of the Acoustical Society of America* **62**, 912–915.
- Allen, J. B. and Rabiner, L. R. (**1977**). “A unified approach to short-time fourier analysis and synthesis”, *Proceedings of the IEEE* **65**, 1558–1564.
- Arbogast, T. L., Mason, C. R., and Kidd, G. (**2002**). “The effect of spatial separation on informational and energetic masking of speech.”, *Journal of the Acoustical Society of America* **112**, 2086–2098.
- Bech, S. and Zacharov, N. (**2006**). *Perceptual Audio Evaluation: Theory, method and application*, 39–96 (Wiley and Sons Ltd., West Sussex, Great Britain).
- Bench, J., Kowal, A., and Bamford, J. (**1979**). “The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children”, *British Journal of Audiology* **13**, 108–112.

- Bernstein, J. G. W. and Grant, K. W. (2009). “Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners.”, *Journal of the Acoustical Society of America* **125**, 3358–3372.
- Best, V., Kalluri, S., McLachlan, S., Valentine, S., Edwards, B., and Carlile, S. (2010). “A comparison of CIC and BTE hearing aids for three-dimensional localization of speech.”, *International Journal of Audiology* **49**, 723–732.
- Best, V., Keidser, G., Buchholz, J. M., and Freeston, K. (2013a). “Psychometric effects of adding realism to a speech-in-noise test”, in *Proceedings of the Acoustical Society of America meeting*, volume 19, 050067.
- Best, V., Marrone, N., Mason, C. R., and Kidd, G. (2012). “The influence of non-spatial factors on measures of spatial release from masking.”, *Journal of the Acoustical Society of America* **131**, 3103–3110.
- Best, V., Thompson, E. R., Mason, C. R., and Kidd, G. (2013b). “An Energetic Limit on Spatial Release from Masking.”, *Journal of the Association for Research in Otolaryngology* 603–610.
- Blauert, J. (1996). *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*, 271–287, 334–337 (The MIT Press, Cambridge, MA, USA).
- Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, USA).
- Bolia, R. S., Nelson, W. T., Ericson, M. a., and Simpson, B. D. (2000). “A speech corpus for multitalker communications research”, *Journal of the Acoustical Society of America* **107**, 1065–1066.
- Bradley, J. S., Sato, H., and Picard, M. (2003). “On the importance of early reflections for speech in rooms”, *Journal of the Acoustical Society of America* **113**, 3233–3244.
- Brandenburg, K. (1987). “Evaluation of quality for audio encoding at low bit rates”, in *Proceedings of the Audio Engineering Society Convention* (London, Great Britain).
- Bregman, A. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*, A Bradford book (Bradford Books).

- Bronkhorst, A. (2000). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions", *Acta Acustica united with Acustica* **86**, 117–128.
- Bronkhorst, A. and Plomp, R. (1990). "A clinical test for the assessment of binaural speech perception in noise", *International Journal of Audiology* **29**, 275–285.
- Brungart, D. S. (2012). "Better-ear glimpsing efficiency with symmetrically-placed interfering talkers", *Journal of the Acoustical Society of America* **132**, 2545–2556.
- Brungart, D. S. and Rabinowitz, W. M. (1999). "Auditory localization of nearby sources. Head-related transfer functions.", *Journal of the Acoustical Society of America* **106**, 1465–1479.
- Brungart, D. S. and Simpson, B. D. (2002). "The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal.", *Journal of the Acoustical Society of America* **112**, 664–676.
- 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", *Journal of the Acoustical Society of America* **110**, 2527–2538.
- Buchholz, J. M. (2007). "Characterizing the monaural and binaural processes underlying reflection masking", *Hearing Research* **232**, 52–66.
- Byrne, D., Dillon, H., and Tran, K. (1994). "An international comparison of long-term average speech spectra", *Journal of the Acoustical Society of America* **96**, 2108–2120.
- Cameron, S. and Dillon, H. (2007). "Development of the Listening in Spatialized Noise-Sentences Test (LISN-S).", *Ear and Hearing* **28**, 196–211.
- Carhart, R., Tillman, T. W., and Greetis, E. S. (1969). "Perceptual masking in multiple sound backgrounds.", *The Journal of the Acoustical Society of America* **45**, 694–703.
- Cherry, E. (1953). "Some experiments on the recognition of speech, with one and with two ears", *Journal of the Acoustical Society of America* **25**, 975–979.
- Christiansen, T. U. and Henriksen, P. J. (2011). "Objective evaluation of consonant-vowel pairs produced by native speakers of danish", in *Proceedings of Forum Acusticum 2011*.

- Culling, J. F. (**2013**). “Energetic and Informational Masking in a Simulated Restaurant Environment”, in *Basic aspects of Hearing*, edited by B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel, volume 787 of *Advances in Experimental Medicine and Biology*, 511–518 (Springer New York, New York, NY).
- Daneman, M. and Carpenter, P. (**1980**). “Individual differences in working memory and reading”, *Journal of Verbal Learning and Verbal Behavior* **466**, 450–466.
- Darwin, C. (**2008**). “Spatial hearing and perceiving sources”, in *Auditory Perception of Sound Sources*, edited by W. Yost, A. Popper, and R. Fay, volume 29 of *Springer Handbook of Auditory Research*, 215–232 (Springer US).
- Dau, T., Püschel, D., and Kohlrausch, a. (**1996**). “A quantitative model of the ”effective” signal processing in the auditory system. I. Model structure.”, *Journal of the Acoustical Society of America* **99**, 3615–3622.
- Dillon, H. (**2001**). *Hearing aids* (Thieme).
- Durlach, N. (**1963**). “Equalization and Cancellation Theory of Binaural Masking Level Differences”, *Journal of the Acoustical Society of America* **426**, 416–426.
- Durlach, N. I., Braida, L. D., and Ito, Y. (**1986**). “Towards a model for discrimination of broadband signals.”, *The Journal of the Acoustical Society of America* **80**, 63–72.
- Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). “Note on informational masking (L)”, *Journal of the Acoustical Society of America* **113**, 2984.
- Favrot, S. and Buchholz, J. (**2010**). “LoRA: A loudspeaker-based room auralization system”, *Acta Acustica united with Acustica* **96**, 364–375.
- Festen, J. M. and Plomp, R. (**1990**). “Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing.”, *Journal of the Acoustical Society of America* **88**, 1725–1736.
- Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). “Spatial release from informational masking in speech recognition”, *Journal of the Acoustical Society of America* **109**, 2112–2122.

- Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (2004). “Effect of number of masking talkers and auditory priming on informational masking in speech recognition”, *Journal of the Acoustical Society of America* **115**, 2246–2256.
- Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (1999). “The role of perceived spatial separation in the unmasking of speech.”, *Journal of the Acoustical Society of America* **106**, 3578–3588.
- Furuya, K. and Kataoka, A. (2007). “Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction.”, *IEEE Transactions on Audio Speech and Language Processing* **15**, 1579–1591.
- Gillespie, B. W., Malvar, H. S., and Florncio, D. A. F. (2001). “Speech dereverberation via maximum-kurtosis subband adaptive filtering”, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 2001*, 3701–3704.
- Glasberg, B. R. and Moore, B. C. (1990). “Derivation of auditory filter shapes from notched-noise data.”, *Hearing Research* **47**, 103–38.
- 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.”, *Journal of the Acoustical Society of America* **134**, 2937–2945.
- Glyde, H., Cameron, S., Dillon, H., Hickson, L., and Seeto, M. (2012). “The effects of hearing impairment and aging on spatial processing.”, *Ear and hearing* **34**, 15–28.
- Goetze, S., Albertin, E., Kallinger, M., Mertins, A., and Kammeyer, K.-D. (2010). “Quality assessment for listening-room compensation algorithms”, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2450–2453 (Dallas, TX, USA).
- Golden, C. J. and Freshwater, S. M. (2002). *The Stroop Color and Word Test* (Sterling, Wood dale, IL).
- Goverts, S. T., Houtgast, T., and van Beek, H. H. (2001). “The precedence effect for lateralization for the mild sensory neural hearing impaired”, *Hearing Research* **163**, 82–92.

- Greenberg, J. E., Peterson, P. M., and Zurek, P. M. (1993). “Intelligibility-weighted measures of speech-to-interference ratio and speech system performance.”, *Journal of the Acoustical Society of America* **94**, 3009–10.
- Habets, E. A. P. (2010). “Speech dereverberation using statistical reverberation models”, in *Speech Dereverberation*, edited by P. A. Naylor and N. D. Gaubitch, *Signals and Communication Technology*, 57–93 (Springer London).
- Hansen, J. H. L. and Pellom, B. L. (1998). “An Effective Quality Evaluation Protocol For Speech Enhancement Algorithms”, *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia 2819–2822.
- Hansen, V. and Munch, G. (1991). “Making recordings for simulation tests in the archimedes project”, *Journal of the Audio Engineering Society* **39**, 768–774.
- Hatziantoniou, P. D. and Mourjopoulos, J. N. (2000). “Generalized fractional-octave smoothing of audio and acoustic responses”, *Journal of the Audio Engineering Society* **48**, 259–280.
- Helfer, K. and Freyman, R. (2005). “The role of visual speech cues in reducing energetic and informational masking”, *Journal of the Acoustical Society of America* **117**, 842–849.
- Helfer, K. S., Chevalier, J., and Freyman, R. L. (2010). “Aging, spatial cues, and single- versus dual-task performance in competing speech perception.”, *Journal of the Acoustical Society of America* **128**, 3625–3633.
- Helfer, K. S. and Freyman, R. L. (2008). “Aging and speech-on-speech masking.”, *Ear and Hearing* **29**, 87–98.
- IEC 60268-16 (2011). “Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index”, *International Electrotechnical Commission* .
- Ihlefeld, A. and Shinn-Cunningham, B. (2008). “Spatial release from energetic and informational masking in a selective speech identification task.”, *Journal of the Acoustical Society of America* **123**, 4369–4379.

- ISO 389-7 (1996). "Acoustics - Reference zero for the calibration of audiometric equipment - Part 7: Reference threshold of hearing under free-field and diffuse-field listening conditions", International Organization of Standardization .
- Jacobsen, F. and Roisin, T. (2000). "The coherence of reverberant sound fields", *Journal of the Acoustical Society of America* **108**, 204–210.
- Jeub, M., Schäfer, M., and Vary, P. (2009). "A binaural room impulse response database for the evaluation of dereverberation algorithms", in *Proceedings of the 16th international conference on Digital Signal Processing*, 550–554 (IEEE Press).
- Jeub, M. and Vary, P. (2010). "Model-based dereverberation preserving binaural cues", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* **18**, 1732–1745.
- Jørgensen, S. and Dau, T. (2011). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing.", *Journal of the Acoustical Society of America* **130**, 1475–87.
- Jørgensen, S. and Dau, T. (2013). "The role of high-frequency envelope fluctuations for speech masking release", in *Proceedings of the Acoustical Society of America meeting*, volume 19, 060126.
- Kates, J. M. (2008). *Digital Hearing Aids*, 221–262 (Plural Publishing, San Diego, CA, USA).
- Keidser, G., Dillon, H., Mejia, J., and Nguyen, C.-V. (2013). "An algorithm that administers adaptive speech-in-noise testing to a specified reliability at selectable points on the psychometric function.", *International Journal of Audiology* **52**, 795–800.
- Kidd, G. and Mason, C. (2005). "The role of reverberation in release from masking due to spatial separation of sources for speech identification", *Acta Acustica united with Acustica* **91**, 526–536.
- Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (2007). "Informational Masking", in *Auditory Perception of Sound Sources*, edited by R. Yost, William A. and Popper, Arthur N. and Fay, 143–189 (Springer US).

- Kidd, 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.", *Journal of the Acoustical Society of America* **104**, 422–431.
- Knudsen, E. I. (2007). "Fundamental components of attention.", *Annual Review of Neuroscience* **30**, 57–78.
- Kock, W. E. (1950). "Binaural Localization and Masking", *Journal of the Acoustical Society of America* **22**, 801–804.
- Kokkinakis, K. and Loizou, P. C. (2011). "Evaluation of objective measures for quality assessment of reverberant speech", in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2420–2423 (Prague, Czech Republic).
- Kollmeier, B., Peissig, J., and Hohmann, V. (1993). "Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain", *Scandinavian Audiology, Supplement* **38**, 28–38.
- Kuttruff, H. (2000). *Room acoustics, 4th Edition*, 136–248 (Taylor & Francis).
- Lavandier, M. and Culling, J. F. (2007). "Speech segregation in rooms: effects of reverberation on both target and interferer.", *Journal of the Acoustical Society of America* **122**, 1713–1723.
- Lebart, K., Boucher, J. M., and Denbigh, P. N. (2001). "A new method based on spectral subtraction for speech dereverberation", *Acta Acustica united with Acustica* **87**, 359–366.
- Litovsky, R. Y., Colburn, S. H., Yost, W. A., and Guzman, S. J. (1999). "The precedence effect", *Journal of the Acoustical Society of America* **106**, 1633–1654.
- Ljung, R. and Kjellberg, A. (2010). "Long reverberation time decreases recall of spoken information", *Building Acoustics* **16**, 301–311.
- Lorho, G. (2010). "Perceived quality evaluation - an application to sound reproduction over headphones", Ph.D. thesis, Aalto University, School of Science and Technology, Department of Signal Processing and Acoustics, Finland.
- Martin, R. (2001). "Small microphone arrays with postfilters for noise and acoustic echo reduction", in *Microphone Arrays*, edited by M. Brandstein and D. Ward (Springer, Cambridge, MA, USA and London, Great Britain).

- McCowan, I. A. and Boulard, H. (2003). “Microphone Array Post-filter based on Noise Field Coherence”, *IEEE Transactions on Audio Speech and Language Processing* **11**, 709–716.
- Moncur, J. and Dirks, D. (1967). “Binaural and monaural speech intelligibility in reverberation”, *Journal of Speech and Hearing Research* **10**, 186–195.
- Moore, B. C. J. (2012). *An Introduction to the Psychology of Hearing*, 67–202 (Emerald, Bingley, United Kingdom).
- Mourjopoulos, J. (1992). “Digital equalization of room acoustics”, in *Proceedings of the Audio Engineering Society Convention 1992* (Vienna, Austria).
- Müller, S. and Massarani, P. (2001). “Transfer-function measurement with sweeps”, *Journal of the Acoustical Society of America* **49**, 443–471.
- Muller, S. and Massarani, P. (2001). “Transfer-Function Measurement with Sweeps *”, *Journal of the Audio Engineering Society* **49**, 443–471.
- Nábělek, A. and Robinson, P. (1982). “Monaural and binaural speech perception in reverberation for listeners of various ages”, *The Journal of the Acoustical Society* ... **86**, 1259–65.
- Naylor, P. A. and Gaubitch, N. D. (2010). *Speech Dereverberation*, 57–387 (Springer, London, Great Britain).
- Neely, S. T. and Allen, J. B. (1979). “Invertibility of a room impulse response”, *Journal of the Acoustical Society of America* **66**, 165–169.
- Oxenham, a. J. and Moore, B. C. (1994). “Modeling the additivity of nonsimultaneous masking.”, *Hearing research* **80**, 105–18.
- Plomp, R. (1976). “Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)”, *Acta Acustica united with Acustica* **34**, 200–211.
- Qin, M. K. and Oxenham, A. J. (2003). “Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers”, *Journal of the Acoustical Society of America* **114**, 446–454.

- RBS.1534-2001:, I. (2003). “Method for the subjective assessment of intermediate quality levels of coding systems”, Technical Report, International Telecommunications Union, Geneva.
- Rindel, J. (2000). “The use of computer modeling in room acoustics”, *Journal of Vibroengineering* **3**, 219–224.
- Roman, N. and Woodruff, J. (2011). “Intelligibility of reverberant noisy speech with ideal binary masking.”, *Journal of the Acoustical Society of America* **130**, 2153–2161.
- Ruggles, D., Bharadwaj, H., and Shinn-Cunningham, B. G. (2011). “Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication.”, *Proceedings of the National Academy of Sciences of the United States of America* **108**, 15516–15521.
- Salomons, A. (1995). “Coloration and binaural decoloration of sound due to reflections”, Ph.D. thesis, Technical University Delft, Department of Signal Processing and Acoustics, Netherlands.
- Scharrer, R. (2010). “Binaurale akustische umgebungserkennung”, in *Proceedings of the German Annual Conference on Acoustics (DAGA)*, 625–626 (Berlin, Germany).
- Seeber, B. U., Kerber, S., and Hafter, E. R. (2010). “A system to simulate and reproduce audio-visual environments for spatial hearing research.”, *Hearing research* **260**, 1–10.
- Shinn-Cunningham, B. G. (2008). “Object-based auditory and visual attention.”, *Trends in Cognitive Sciences* **12**, 182–186.
- Shinn-Cunningham, B. G., Schickler, J., Kopco, N., and Litovsky, R. (2001). “Spatial unmasking of nearby speech sources in a simulated anechoic environment.”, *Journal of the Acoustical Society of America* **110**, 1118–1129.
- Simmer, K., Fischer, S., and Wasiljeff, A. (1994). “Suppression of coherent and incoherent noise using a microphone array”, *Annals of Telecommunications* **49**, 439–446.
- Smeds, K., Wolters, F., and Rung, M. (2014). “Estimation of signal-noise ratios in realistic sound scenarios”, *Journal of the American Academy of Audiology* *Accepted*.
- Stone, M. a., Fullgrabe, C., Mackinnon, R. C., and Moore, B. C. J. (2011). “The importance for speech intelligibility of random fluctuations in “steady” background noise”, *The Journal of the Acoustical Society of America* **130**, 2874.

- Strelcyk, O., Pentony, S., Kalluri, S., and Edwards, B. (2014). “Effects of interferer facing orientation on speech perception by normal-hearing and hearing-impaired listeners.”, *Journal of the Acoustical Society of America* **135**, 1419–1432.
- Tsilfidis, A. and Mourjopoulos, J. (2009). “Signal-dependent constraints for perceptually motivated suppression of late reverberation”, *Signal Processing* **90**, 959–965.
- Tsilfidis, A. and Mourjopoulos, J. (2011). “Blind single-channel suppression of late reverberation based on perceptual reverberation modeling”, *Journal of the Acoustical Society of America* **129**, 1439–1451.
- Tsilfidis, A., Mourjopoulos, J., and Tsoukalas, D. (2008). “Blind estimation and suppression of late reverberation utilising auditory masking”, in *Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, 208–211 (Trento, Italy).
- Vesa, S. (2009). “Binaural sound source distance learning in rooms”, *Audio, Speech, and Language Processing, IEEE Transactions on* **17**, 1498–1507.
- Watson, C. S. (2005). “Some Comments on Informational Masking”, *Acta Acustica united with Acustica* **91**, 502–512.
- Watson, C. S., Kelly, W. J., and Wroton, H. W. (1976). “Factors in the discrimination of tonal patterns. II. Selective attention and learning under various levels of stimulus uncertainty.”, *The Journal of the Acoustical Society of America* **60**, 1176–1186.
- Wen, J., Gaubitch, N. D., Habets, E., Myatt, T., and Naylor, P. A. (2006). “Evaluation of speech dereverberation algorithms using the mardy database”, in *Proceedings of the International Workshop on Acoustic Echo and Noise Control, (IWAENC)* (Paris, France).
- Westermann, A. and Buchholz, J. (2014a). “The effect of spatial separation in distance on the intelligibility of speech in rooms.”, *Journal of the Acoustical Society of America* *Submitted*.
- Westermann, A. and Buchholz, J. (2014b). “The influence of informational masking in reverberant, multi-talker environments.”, *Journal of the Acoustical Society of America* *Submitted*.

- Westermann, A., Buchholz, J. M., and Dau, T. (2013). “Binaural dereverberation based on interaural coherence histograms.”, *Journal of the Acoustical Society of America* **133**, 2767–2777.
- Wilson, B. S. and Dorman, M. F. (2008). “Cochlear implants: a remarkable past and a brilliant future.”, *Hearing research* **242**, 3–21.
- Wood, N. and Cowan, N. (1995). “The cocktail party phenomenon revisited: how frequent are attention shifts to one’s name in an irrelevant auditory channel?”, *Journal of Experimental Psychology: Learning, Memory and Cognition* **21**, 255–260.
- Wu, M. and Wang, D. (2006). “A two-stage algorithm for one-microphone reverberant speech enhancement”, *IEEE Transactions on Audio Speech and Language Processing* **14**, 774–784.
- Yegnanarayana, B., Avendano, C., Hermansky, H., and Murthy, P. S. (1999). “Speech enhancement using linear prediction residual”, *Speech Communication* **28**, 25–42.
- Zahorik, P. (2002a). “Direct-to-reverberant energy ratio sensitivity.”, *Journal of the Acoustical Society of America* **112**, 2110–2117.
- Zahorik, P. (2002b). “Direct-to-reverberant energy ratio sensitivity.”, *Journal of the Acoustical Society of America* **112**, 2110–2117.
- Zahorik, P. (2005). “Auditory distance perception in humans: A summary of past and present research”, *Acta Acustica united with ...* **91**, 409–420.
- Zahorik, P., Brungart, D. S., and Bronkhorst, A. W. (2005). “Auditory distance perception in humans: A summary of past and present research”, *Acta Acustica united with Acustica* **91**, 409–420.
- Zurek, P. M. (1979). “Measurements of binaural echo suppression”, *Journal of the Acoustical Society of America* **66**, 1750–1757.