Near-End Listening Enhancement: Theory and Application

Von der Fakultät für Elektrotechnik und Informationstechnik
der Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines Doktors
der Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Diplom-Ingenieur

Bastian Sauert

aus Düsseldorf

Berichter: Universitätsprofessor Dr.-Ing. Peter Vary
Universitätsprofessor Dr.-Ing. Ulrich Heute

Tag der mündlichen Prüfung: 10. Oktober 2013

AACHENER BEITRÄGE ZU DIGITALEN NACHRICHTENSYSTEMEN

Herausgeber:

Prof. Dr.-Ing. Peter Vary
Institut für Nachrichtengeräte und Datenverarbeitung
Rheinisch-Westfälische Technische Hochschule Aachen
Muffeter Weg 3a
52074 Aachen
Tel.: 02 41-80 26 956
Fax: 02 41-80 22 186

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische
Daten sind im Internet über `http://dnb.ddb.de` abrufbar

1. Auflage Aachen:
Wissenschaftsverlag Mainz in Aachen
(Aachener Beiträge zu Digitalen Nachrichtensystemen, Band 36)
ISSN 1437-6768
ISBN 978-3-86130-729-7

Wissenschaftsverlag Mainz
Süsterfeldstr. 83, 52072 Aachen
Tel.: 02 41 / 2 39 48 oder 02 41 / 87 34 34
Fax: 02 41 / 87 55 77
www.verlag-mainz.de

Herstellung: Druckerei Mainz GmbH
Süsterfeldstr. 83, 52072 Aachen
Tel.: 02 41 / 87 34 34
Fax: 02 41 / 87 55 77
www.druckereimainz.de

Gedruckt auf chlorfrei gebleichtem Papier

D 82 (Diss. RWTH Aachen University, 2013)

To Alexandra

For your love, your encouragement, and your patience

Acknowledgments

This thesis was written during my time as research assistant and academic councillor at the Institute of Communication Systems and Data Processing (IND) at the RWTH Aachen University. I would like to take the opportunity to thank all those who contributed to the success of my work in so many ways.

First, I would like to express my sincere gratitude to my supervisor Prof. Dr.-Ing. Peter Vary for his continuous support, his numerous ideas and suggestions, as well as for creating such an inspiring and pleasant working environment. I am also indebted to Prof. Dr.-Ing. Ulrich Heute for being my co-supervisor and showing such a great interest in my work.

Furthermore, I want to thank all my current and former colleagues at the IND for providing a unique and very enjoyable work climate, for numerous fruitful discussions, very valuable proof-readings, and many interesting collaborations. Thank you, A. Heitzig, Andreas W., Annika, Aulis, Benedikt, Bernd, Birgit, Carsten, Christoph E., Christoph N., Florian, Frank, Gerald, Hauke, Heiner, Helge, H. Krott, K. Nguyen, Laurent, Magnus, Marc A., Marc W., Marco, Matthias P., Matthias R., Moritz, Peter J., Simone, Stefan, Sylvia, Thomas B., Thomas E., Thomas L., Thomas S., Thorsten, Tim, and Tobias. I am particularly grateful to Christiane Antweiler for her extraordinary support and encouragement. Very special thanks go to Max Mascheraux, who fabulously realized all the hardware designs I needed for my studies on loudspeaker protection.

I would also like to express my appreciation to all the students who made significant contributions to my work.

I owe thanks to the Institute for Communications Engineering (IENT) and the Institute of Technical Acoustics (ITA) for generously giving me access to their anechoic chambers and especially to Markus Müller-Trapet from ITA for his patient support during our measurements.

Thanks are extended to Siemens Mobile Phones and BenQ Mobile, Siemens Audiology Solutions, and Intel Mobile and Communications Group for providing funding and for the good and friendly collaboration over the years. Thank you, Bruno, Christophe, Fabrice, Henning, Martin, and Robert.

Finally, I wish to express my deepest thanks to my family and my friends. I cannot thank them enough for their encouragement and support.

Bastian Sauert

Abstract

Mobile telephony is often conducted in the presence of acoustical background noise such as traffic or babble noise. In this situation, the near-end listener perceives a mixture of clean far-end (downlink) speech and environmental noise from the near-end side, which goes along with an increased listening effort and possibly reduced speech intelligibility. As in many cases the noise signal cannot be influenced, the manipulation of the far-end signal is the only way to effectively improve speech intelligibility and to ease listening effort for the near-end listener by digital signal processing. We call this approach *near-end listening enhancement* (NELE).

In this thesis, innovative solutions for the problem of near-end listening enhancement are developed. These optimize the intelligibility of the far-end speech in local background noise with respect to the objective criterion *Speech Intelligibility Index* (SII). In contrast to state-of-the-art techniques, the developed methods tackle the problem for the first time from the application perspective considering also the requirements and restrictions of realistic scenarios such as in mobile phones. It is of particular importance that the processing adapts dynamically to the sound characteristics of the ambient noise. Hence, an effective intelligibility enhancement is provided in the presence of background noise, while in silence *no* audible modification is applied. The utilized noise tracking algorithm estimates the noise spectrum blindly from the microphone signal, the only access to the acoustical environment. Furthermore, a power limitation in critical bands ensures that the ear of the near-end listener is protected from damage and pain.

In mobile phones, the restrictions of the so-called micro-loudspeakers need to be considered and were thus experimentally evaluated and modeled in this thesis. Especially the maximum thermal load of the micro-loudspeaker constitutes a major limitation. This leads to an optimization of the SII with the constraint that the total audio power may only be increased up to a maximum power.

Besides the protection of the human ear, damage of the loudspeaker due to excessive excursions of the membrane or overheating must be prevented. Therefore, a *loudspeaker protection* scheme for mobile phones with a frequency dependent limitation has been developed. In contrast to the human ear protection, much shorter attack time constants are required. This leads to tight constraints on the filterbank design.

Although the presented algorithms for near-end listening enhancement are driven by real application constraints, this thesis also includes the derivation of theoretical bounds, instrumental measures, and auditory evaluations. As a result, significant improvements of speech intelligibility under adverse acoustical conditions are achieved. In the most difficult scenario where an increase of total audio power is not allowed, the word recognition rate improves with the proposed algorithms by up to 22 percentage points.

It is shown, that the developed new concepts can also be applied in different devices such as mobile phones, headphones, hands-free conference terminals, car multimedia systems, public address systems, and hearing aids.

Contents

Chapter 1

Introduction

In the beginning of telephony, the terminals were connected by wire and calls were mostly conducted indoor. At that time, the acoustical background noise could be controlled and was not a major problem. With the advent of cellular phones, people often make phone calls in challenging acoustical environments where a conversation is eventually perceptually impossible.

In these situations, strong acoustical background noise such as traffic or babble noise is often present at the near-end side. This has three major implications:

- The near-end user modifies her/his speaking style as a consequence of the exposure to the near-end noise, an effect known as Lombard reflex (Lombard 1911; Summers et al. 1988).
- The near-end noise is captured by the microphone together with the near-end speech. Several noise reduction techniques have been proposed, to reduce this noise signal before speech coding and transmission.
- The near-end user perceives a mixture of the clean or noise reduced far-end speech and the local acoustical background noise at the near-end side. Thus, the user experiences an increased listening effort and possibly a reduced speech intelligibility, which is addressed in this thesis.

The noisy environment can usually not be influenced easily, like car noise at a busy street or speech babble noise in a cafeteria. Although one ear of the near-end listener is “covered” to some extend by the mobile phone in handset mode, the noise signal is nevertheless perceived by both ears. As there is no possibility to intercept the near-end noise, the manipulation of the far-end signal is the only way to effectively improve speech intelligibility for the near-end listener by signal processing. We call this approach *near-end listening enhancement* (NELE).

A number of speech modification algorithms have been presented in literature to tackle the problem of NELE, which is also known as “speech intelligibility enhancement”, “speech reinforcement”, or simply “speech enhancement”. To date, most of the proposed algorithms are noise independent, i. e., the same processing with the same setup is performed regardless of the signal-to-noise ratio (SNR) or other noise characteristics. These noise independent methods include

- boosting of the consonant-vowel-ratio (Kretsinger & Young 1960; Thomas & Niederjohn 1970; Niederjohn & Grotelueschen 1976; Harris & Skowronski

2002; Yoo et al. 2007; Tantibundhit et al. 2007; Rasetshwane et al. 2009; Chanda & S. Park 2007),
- formant enhancement (Thomas & Ohley 1972; McLoughlin & Chance 1997; Hall & Flanagan 2010; Jokinen et al. 2012),
- manipulation of duration and prosody (Huang et al. 2010), and
- more advanced manipulations of the temporal structure (Rankovic 1991; H. Park et al. 2010; Zorilă et al. 2012).

They, however, result in a modified speech signal even in quiet environments.

Only recently, some techniques have been studied which utilize prior knowledge or estimates of the noise context. These approaches include

- formant enhancement (Brouckxon et al. 2008),
- modification of the local SNR (Choi et al. 2009[1]; Tang & Cooke 2011),
- spectral shaping and dynamic range compression (Erro et al. 2012), and
- optimization with respect to an objective criterion (Taal et al. 2012; Tang & Cooke 2012).

A different approach preserves the (partial) loudness of the speech signal despite the noise (J. W. Shin et al. 2009; H. S. Shin et al. 2010) and thereby requires increasing the signal energy.

In this thesis, innovative solutions for the problem of near-end listening enhancement are developed. These optimize the intelligibility of the far-end speech in local background noise with respect to the objective criterion *Speech Intelligibility Index* (SII). In contrast to state-of-the-art techniques, the developed methods tackle the problem for the first time from the application point of view considering also the requirements and restrictions of realistic scenarios such as in mobile phones. It is of particular importance that the processing adapts dynamically to the sound characteristics of the ambient noise. Hence, an effective intelligibility enhancement is provided in the presence of background noise, while in silence *no* audible modification is applied. The utilized noise tracking algorithm estimates the noise spectrum blindly from the microphone signal – the only access to the acoustical environment – and at the same time disregards the voice of the near-end user in double-talk situations. Furthermore, a power limitation in critical bands ensures that the ear of the near-end listener is protected from damage and pain.

Chapter 2 discusses the system model for NELE in mobile phones and the processing framework used in this thesis. Two different objective measures to judge speech intelligibility are presented, the SII and the *speech-based revised Speech Transmission Index* ($\mathrm{STI}_{\mathrm{sr}}$). Finally, a literature overview of NELE is given.

The optimization with respect to the SII is explained in Chapter 3. It results in an upper performance bound, which can only be reached with high-end loudspeakers. In mobile phones, however, the restrictions of the so-called micro-loudspeakers need to be considered. Especially the maximum thermal load of the micro-loudspeaker constitutes a major limitation. Thus, the total audio power may only be increased

[1] This contribution is actually an extension of (Sauert & Vary 2006b).

up to this maximum, leading to a *constrained* optimization of the SII, which is discussed in Chapter 4. In the extreme, the maximum allowed power is limited to the input power, in other words the total audio power of the speech signal may not be increased. This can be interpreted as a special case for sound reproduction systems without head-room in terms of output power. Besides the objective evaluations with instrumental measures, one of the developed algorithms proves its effectiveness in two large scale formal subjective listening tests with natural and synthetic speech.

Nowadays, modern mobile phones are required to provide a loud sound reproduction with good quality in use cases like hands-free telephony, portable radio receiver, music and video player, games console, ... This pushes the micro-loudspeakers to their limits. Therefore, their acoustical distortions, membrane excursion, and progress of temperature are studied by experiments in Chapter 5. A *loudspeaker protection* (LOPRO) scheme for mobile phones with frequency dependent limitation is developed, which prevents damage of the loudspeaker due to excessive excursions of the membrane or overheating. In contrast to the human ear protection, which is integrated in the proposed NELE framework, LOPRO requires much shorter time constants. This imposes tight constraints for instance on the filterbank design.

Although the development of new algorithms for near-end listening enhancement is chiefly driven by mobile phone application, these methods can also be applied in different devices such as headphones, hands-free conference terminals, car multimedia systems, public address systems, and hearing aids. These examples of application are finally discussed in Chapter 6 and point out the relevance of the presented new concepts.

Parts of this thesis have been presented in the following references published by the author: (Sauert et al. 2006; Sauert & Vary 2006a; Sauert & Vary 2006b; Schönle et al. 2006; Sauert et al. 2008; Sauert & Vary 2009; Sauert & Vary 2010a; Sauert & Vary 2010b; Schäfer et al. 2010; Sauert & Vary 2011; Sauert & Vary 2012a; Sauert & Vary 2012b; Cooke et al. 2013; Valentini-Botinhao et al. 2013). Throughout this thesis, these references are marked by underlining the year.

Chapter 2

Models, Methodology, and Literature Overview

In this chapter, the models and methodology of this thesis are introduced and described.

The system model is defined in Section 2.1 together with the assumptions made for system simulation. The framework for all near-end listening enhancement algorithms is presented in Section 2.2. Section 2.3 introduces the measures to evaluate speech intelligibility, followed by a description of the simulation environment in Section 2.4.

Finally, Section 2.5 gives a literature overview of near-end listening enhancement.

2.1 System Model

Figure 2.1 illustrates the application of handset telephony in the presence of acoustical background noise. The far-end (downlink) speech signal, which is assumed to be either clean or sufficiently noise reduced, is manipulated by *near-end listening enhancement* (NELE) to improve intelligibility for the near-end user utilizing an estimate of the near-end noise. Subsequently, *loudspeaker protection* (LOPRO) is applied to the enhanced far-end speech signal to prevent damage of the loudspeaker.

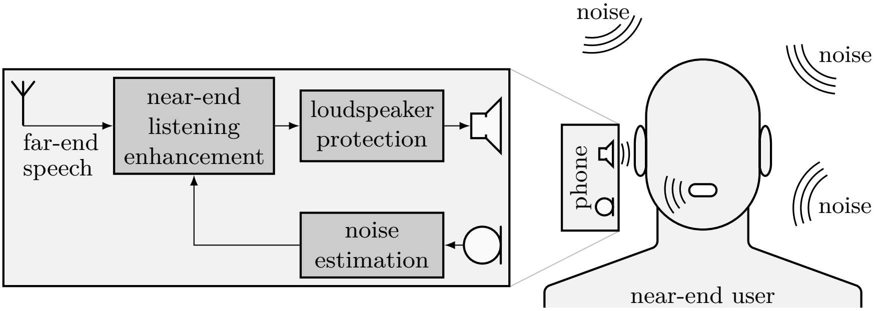


Figure 2.1: Handset telephony in noise.

In handset mode, the mobile phone is held at one ear, which is named “covered” in the following, whereas the other ear is “open”. As a consequence, the desired signal from the mobile phone is presented monaural at the covered ear, while the noise signal is perceived (differently) by both ears. In addition to the shadowing of the head, the noise signal is modified at the covered side by the frequency and direction of arrival (DOA) dependent acoustic characteristic of the covering phone.

Although speech and noise are differently presented to the ears, the system model and the derivation of the NELE concepts consider only the covered ear, which is motivated in the following:

Jeub et al. (2011) showed, that the signals at the ears of a human head in a diffuse noise field have a very low coherence for frequencies above 400 Hz. This “cut-off” frequency of the coherence is lower than in free-field condition due to the shadowing influence of the head. As the coherence basically is a measure of the correlation between the frequency components of two signals (Gardner 1992), the noise field at the human ears is approximately uncorrelated for frequencies above 400 Hz.

The binaural masking level difference (BMLD) indicates the amount by which the SNR of a signal must be increased or decreased to give the same detection score as the monaural reference condition, where both speech and noise are presented monaurally at only one ear. Wilbanks and Whitmore (1968) as well as Dolan and Robinson (1967) reported a BMLD of approximately 0 dB for monaurally presented speech in a binaural noise field with an interaural noise correlation of up to 30 %. This means, that a monaural speech signal is detected at the same level whether it is presented in monaural noise at the same ear or in a diffuse noise field with the same SNR on both ears.

As the narrow-band telephone speech is bandpass filtered with a lower cut-off frequency of about 300 Hz and the micro-loudspeakers of mobile phones have difficulties to properly present a signal at that frequencies, this justifies that the binaural presentation of the noise can be ignored during the derivation of the NELE algorithm.

In double-talk situations, where the far-end speaker and the near-end user speak at the same time, the microphone signal contains not only the ambient noise from the near-end side but also the interfering speech signal. In this case, it is crucial to apply a noise estimation algorithm which is capable of disregarding the near-end speech. Otherwise, a feedback loop between the near-end loudspeaker signal delivered to the near-end user and her/his speech signal would arise, which is at least distracting and annoying, if it does not block communication at all.

Section 2.1.1 describes the model of the acoustic transfer functions involved at the near-end side. Some of those transfer functions were evaluated in a measurement study, which is presented and discussed in Section 2.1.2. The complete system model with all relevant blocks and transfer functions is then described in Section 2.1.3, while Section 2.1.4 discusses the assumptions which yield the two system models used for simulations. Finally, Section 2.1.5 describes the calibrations and normalizations which are the link between the acoustic sound pressure and the unit-less entities of digital signal processing.

2.1.1 Model of Acoustic Transfer Functions

Figure 2.2 depicts the acoustic transfer functions involved in mobile telephony. After digital-analog conversion, the loudspeaker signal is amplified and played by the loudspeaker, summarized with the effective (electro-acoustic) transfer function $H_{\mathrm{ls}}(f)$ with f being the continuous frequency. The acoustic signal propagates from the loudspeaker to the near-end listener's covered ear, denoted with the acoustic transfer function $H_{\mathrm{ear}}(f)$.

Although the mobile phone covers the near-end listener's ear, the ear is also reached by the sound waves from the near-end noise source with direction of arrival (DOA) θ. This effect is called "acoustical leakage" (Krebber 1995) and is described by the acoustic transfer function $H_{\mathrm{leak},\theta}(f)$. In general, there can be multiple noise sources with different DOAs. For the presented system model this makes, however, no difference as will be shown later.

At the ear, the loudspeaker signal and the near-end noise signal are summed up. Potential non-linear effects due to saturation of the human ear are assumed to be considered by the objective intelligibility measures discussed in Section 2.3.

The near-end noise additionally reaches the microphone of the mobile phone via the acoustic transfer functions $H_{\mathrm{noise},\theta}(f)$. In a double-talk situation, the near-end speech signal, i. e., the voice of the near-end user, propagates from her/his mouth to the microphone, expressed by the effective magnitude response $H_{\mathrm{speech}}(f)$.

The loudspeaker signal is also fed back to the microphone as echo, attenuated according to the echo path $H_{\mathrm{echo}}(f)$. At the microphone, the sum of these three signals is recorded and lowpass filtered for the analog-digital conversion, which is modelled by the transfer function $H_{\mathrm{mic}}(f)$.

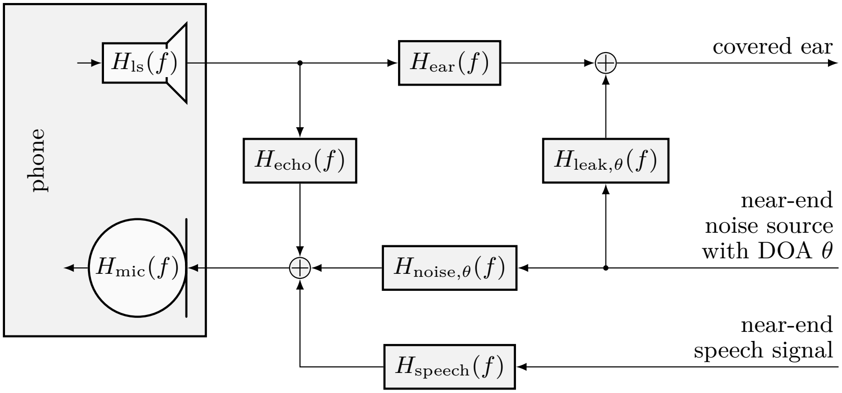


Figure 2.2: Model of acoustic transfer functions.

Discussion of Acoustic Transfer Functions

In the following, all relevant acoustic transfer functions but $H_{\mathrm{ls}}(f)$ are discussed. The frequency response $H_{\mathrm{ls}}(f)$ is investigated in detail in Chapter 5.

Generally, the transfer function $H_{\mathrm{ear}}(f)$ between loudspeaker and (covered) ear depends in handset mode on the mobile phone itself, the contact pressure between ear and mobile phone, and on how the mobile phone is held. In the situation of a noisy environment, however, listeners can be expected to find a position and especially a contact pressure which gives an optimal listening experience (Krebber 1995). Even though this position will slightly differ for each listener, it is reasonable to assume that the resulting frequency response is close to the response the manufacturer measured during the tuning of the mobile phone. It can therefore be compensated well within the feasible frequency range during loudspeaker equalization.

As the loudspeaker of the mobile phone is quasi "coupled" to the covered ear, the acoustical echo path $H_{\mathrm{echo}}(f)$ from the loudspeaker to the microphone is very weak in handset mode (3GPP TS 26.131 2011) and can be neglected.

The DOA dependent transfer functions $H_{\mathrm{leak},\theta}(f)$ and $H_{\mathrm{noise},\theta}(f) \cdot H_{\mathrm{mic}}(f)$ were investigated in (Schäfer 2005) and are presented in the next section. These transfer functions are important for NELE in mobile phones as the near-end noise field can only be recorded at the position of the microphone. However, the near-end listener perceives the enhanced far-end speech signal together with the near-end noise at the position of her/his ear. Therefore, the "virtual" magnitude response

$$H_{\mathrm{match},\theta}(f) = \left| \frac{H_{\mathrm{leak},\theta}(f)}{H_{\mathrm{noise},\theta}(f)} \right| \tag{2.1}$$

from the microphone of the mobile phone to the covered ear of the near-end listener is used to compensate for this mismatch such that the NELE algorithm can utilize the power spectral density (PSD) of the noise signal that is present at the listener's ear. Since the DOA can hardly be estimated using the microphone(s) of a mobile phone, $H_{\mathrm{match},\theta}(f)$ is approximated by an "average" magnitude response $\overline{H}_{\mathrm{match}}(f)$. Therefore, the dependency of $H_{\mathrm{match},\theta}(f)$ on θ and the accuracy of the approximation is also investigated.

2.1.2 Measurement of Direction of Arrival Dependent Acoustic Transfer Functions

This section describes the measurements of the DOA dependent acoustic transfer functions $H_{\mathrm{leak},\theta}(f)$ and $H_{\mathrm{noise},\theta}(f) \cdot H_{\mathrm{mic}}(f)$ of a *Siemens M65* mobile phone which were conducted by Magnus Schäfer in his student research project (Schäfer 2005).

Measurement Setup

The measurement setup consists of a *Genelec 1030A* high-end loudspeaker, which outputs the measurement signal, and an artificial head with ear coupler and mounted

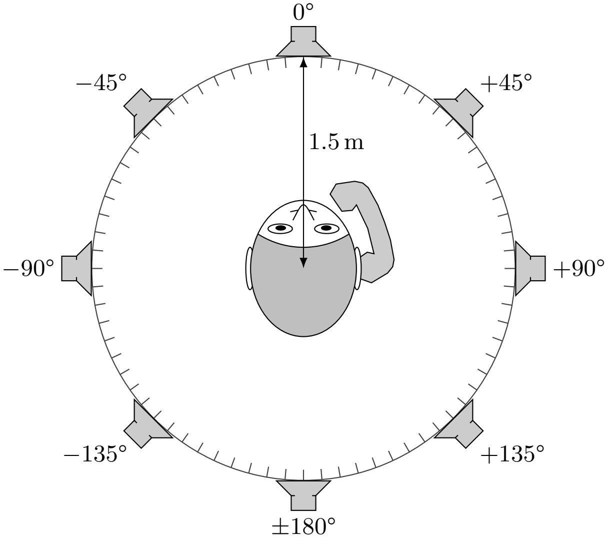


Figure 2.3: Model of the measurement setup, cf. (Schäfer 2005).

mobile phone, which records the response. The measurement took place in the anechoic chamber at the Institute for Communications Engineering at the RWTH Aachen University.

Loudspeaker and artificial head are placed in a distance of 1.5 m with loudspeaker and ears being at the same height. In each measurement step, the responses at the ears and the microphone of the mobile phone are recorded for a different DOA. A model of the measurement setup is depicted in Figure 2.3. Instead of rotating the loudspeaker around the artificial head to obtain the responses in the whole horizontal plane, the artificial head is turned by 5° after each step. Both have the same effect in an anechoic chamber, but the latter has a faster and more accurate handling.

The measurements are performed with *HEAD acoustics*' artificial head measurement system *HMS II.4* with ear simulator and the simplified pinna simulator according to (ITU-T P.57 2009, Type 3.4). The *Siemens M65* is connected to the *HMS II.4* using the handset positioner *HHP II* (see Figure 2.4) with a contact pressure of 8 N. According to Krebber (1995), this force is assumed to be typically chosen by humans for mobile communication in a noisy environment.

Measurement Signal

As the system under inspection does not change during each measurement step, it can be considered time invariant. Therefore, the real and imaginary part of a complex chirp signal were transmitted sequentially over the real acoustical system

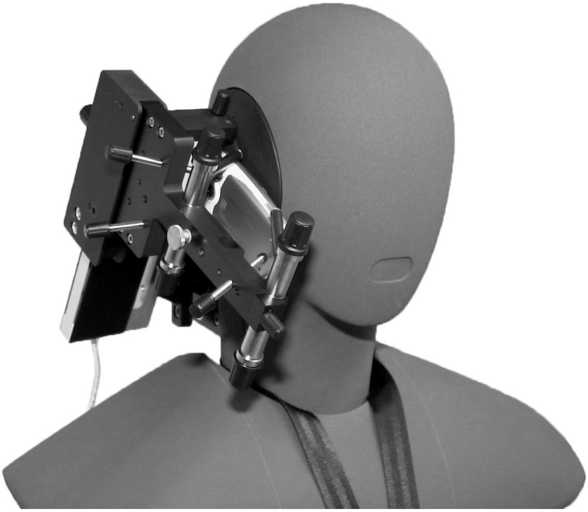


Figure 2.4: *HMS II.4* with *HHP II* and mounted mobile phone.

to perform a system identification. The complex chirp signal was chosen due to its constant amplitude spectrum and its high crest factor (Vary 1980).

In order to reduce the influence of the transient behaviour of the system, a modified complex chirp signal

$$x(k) = \exp\left\{\mathrm{j}\frac{\pi}{M} \cdot \left(\frac{M}{2} - k\right)^2\right\}, \quad k \in \{0, 1, \ldots, M-1\}, \tag{2.2}$$

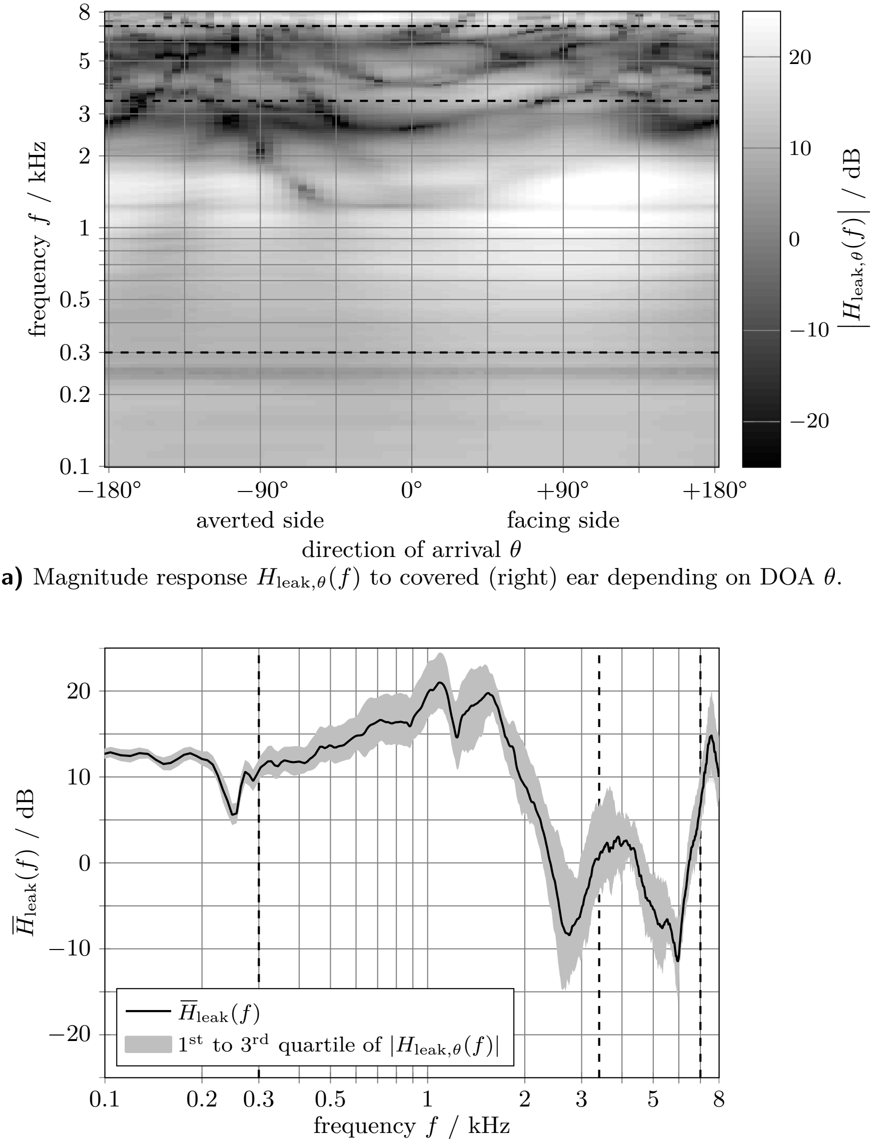
is used with a discrete Fourier transform (DFT) of size $M = 32000$ at a sampling rate of 32 kHz. Thus, the chirp signal starts and ends at 16 kHz, i. e., beyond the range of interest, and the relevant frequency range can be presented without steps or other discontinuities. The real and imaginary part of the complex chirp signal $x(k)$ are each played seven times to eliminate potential instationary noise sources and averaged afterwards, yielding the response signal $y(k)$.

The measured transfer function is finally obtained by normalization of the response signal $y(k)$ to the chirp signal $x(k)$ in frequency domain.

Results

Figure 2.5a shows the measured "acoustical leakage" $H_{\mathrm{leak},\theta}(f)$ from sound source to covered (right) ear as a function of the DOA θ. For frequencies below 1 kHz, $H_{\mathrm{leak},\theta}(f)$ is almost independent of θ. Above 2.5 kHz, it shows some zeros, which are mainly caused by reflections of the sound waves at the handset positioner.

To better illustrate the general behaviour of $H_{\mathrm{leak},\theta}(f)$ and its dependency on



(a) Magnitude response $H_{\text{leak},\theta}(f)$ to covered (right) ear depending on DOA θ.

(b) Average magnitude response $\overline{H}_{\text{leak}}(f)$ to covered (right) ear.

Figure 2.5: Magnitude responses to covered (right) ear with mounted *Siemens M65* mobile phone. The dashed lines denote the relevant cut-off frequencies for narrow-band and wide-band telephony.

the DOA besides the zeros, the average in dB over all θ of $|H_{\text{leak},\theta}(f)|$,

$$20 \log\left\{\overline{H}_{\text{leak}}(f)\right\} = \underset{\theta}{\text{mean}}\, 20 \log\left\{|H_{\text{leak},\theta}(f)|\right\}, \tag{2.3}$$

is plotted in Figure 2.5b together with the range between the first and the third quartile[1] of $|H_{\text{leak},\theta}(f)|$ w. r. t. θ. The quartiles mainly cut off poles and zeros and show that most $|H_{\text{leak},\theta}(f)|$ are within 4 dB around the average $\overline{H}_{\text{leak}}(f)$.

The acoustic transfer function $H_{\text{noise},\theta}(f) \cdot H_{\text{mic}}(f)$ from the sound source to the microphone of the mobile phone including the microphone characteristic is depicted in Figure 2.6a as a function of θ. For frequencies below about 300 Hz, it is rather independent of the DOA. Above, the magnitude response is larger in the frontal direction facing the sound source and smaller when the sound source is behind the head. Above 1.2 kHz, $H_{\text{noise},\theta}(f) \cdot H_{\text{mic}}(f)$ also exhibits zeros, especially for DOAs from behind, which are again expected to be caused by reflections at the handset positioner.

Figure 2.6b presents the average in dB over all θ of $|H_{\text{noise},\theta}(f) \cdot H_{\text{mic}}(f)|$,

$$20 \log\left\{\overline{H}_{\text{noise}}(f) \cdot |H_{\text{mic}}(f)|\right\} = \underset{\theta}{\text{mean}}\, 20 \log\left\{|H_{\text{noise},\theta}(f) \cdot H_{\text{mic}}(f)|\right\}, \tag{2.4}$$

and the range between the first and the third quartile of $|H_{\text{noise},\theta}(f) \cdot H_{\text{mic}}(f)|$ w. r. t. θ. It again shows that most $|H_{\text{noise},\theta}(f)|$ are within 4 dB around the average $\overline{H}_{\text{noise}}(f)$.

These measured acoustic transfer functions are used to calculate the "virtual" magnitude response

$$\frac{H_{\text{match},\theta}(f)}{|H_{\text{mic}}(f)|} = \left| \frac{H_{\text{leak},\theta}(f)}{H_{\text{noise},\theta}(f) \cdot H_{\text{mic}}(f)} \right| \tag{2.5}$$

from the output of the microphone of the mobile phone to the covered ear as described above. Its average magnitude response

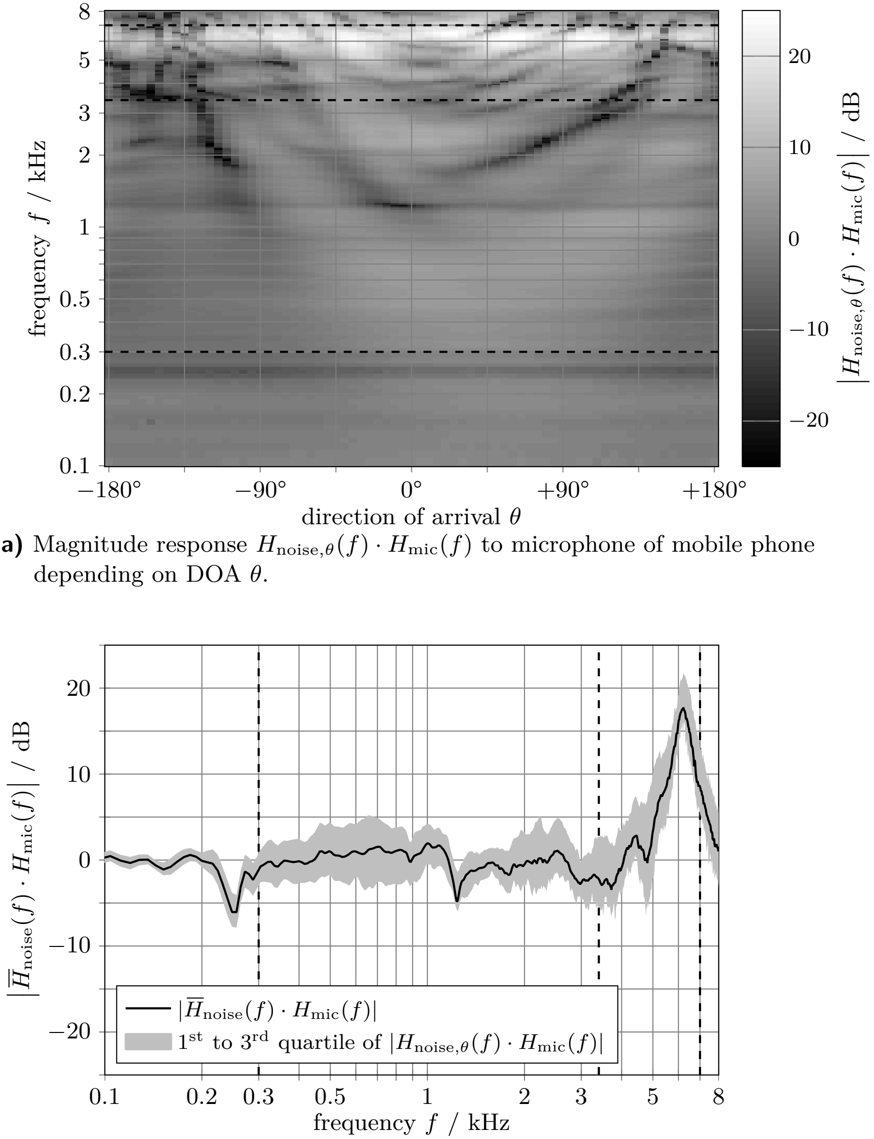
$$20 \log\left\{\frac{\overline{H}_{\text{match}}(f)}{|H_{\text{mic}}(f)|}\right\} = \underset{\theta}{\text{mean}}\, 20 \log\left\{\frac{H_{\text{match},\theta}(f)}{|H_{\text{mic}}(f)|}\right\} \tag{2.6}$$

is depicted in Figure 2.7. For frequencies below 1.2 kHz, the range between first and third quartile w. r. t. θ are within 2 dB around the average $\overline{H}_{\text{match}}(f) \cdot |H_{\text{mic}}(f)|^{-1}$. For most other frequencies below 2.1 kHz and above 3.4 kHz, it is within 5 dB around the average.

Discussion

As a result, all measured transfer functions can be approximated within a 4 to 5 dB range by their average over all DOAs.

[1] The first quartile splits for each f the lowest 25 % of $|H_{\text{leak},\theta}(f)|$ w. r. t. θ, whereas the third quartile splits the highest 25 %.



(a) Magnitude response $H_{\text{noise},\theta}(f) \cdot H_{\text{mic}}(f)$ to microphone of mobile phone depending on DOA θ.

(b) Average magnitude response $\overline{H}_{\text{noise}}(f) \cdot H_{\text{mic}}(f)$ to microphone.

Figure 2.6: Magnitude responses to microphone of mounted *Siemens M65* mobile phone. The dashed lines denote the relevant cut-off frequencies for narrow-band and wide-band telephony.

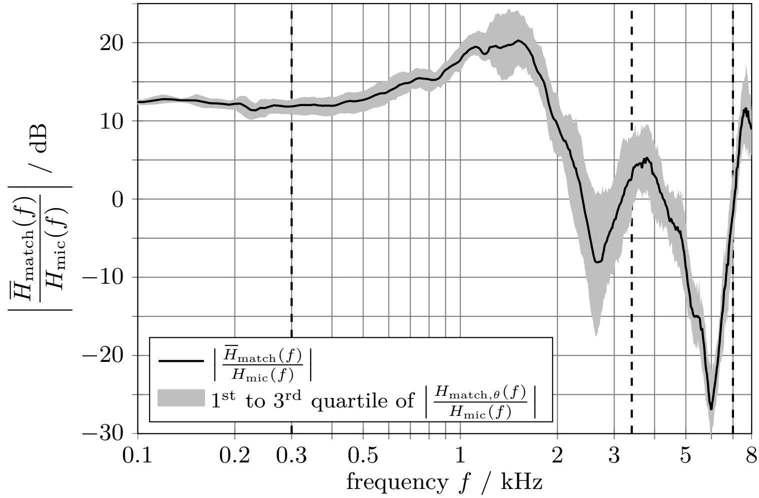


Figure 2.7: Average magnitude response $\overline{H}_{\text{match}}(f) \cdot H_{\text{mic}}^{-1}(f)$ from microphone of mounted *Siemens M65* mobile phone to covered ear. The dashed lines denote the relevant cut-off frequencies for narrow-band and wide-band telephony.

The "omnidirectional" average magnitude response $\overline{H}_{\text{match}}(f)$ is a valid first approximation for $H_{\text{match},\theta}(f)$, especially for frequencies below 1.2 kHz. This holds in particular as most real-world noise fields are diffuse. As a consequence, the noise characteristics at the covered ear of the near-end listener can be derived from the microphone signal. For frequencies above 1.2 kHz, the approximation of the DOA dependent $H_{\text{match},\theta}(f)$ with the average magnitude response $\overline{H}_{\text{match}}(f)$ is less precise and NELE algorithms should not rely on an exact estimate of the noise PSD.

2.1.3 Complete System Model

The complete model of signal flow is depicted in Figure 2.8. The acoustical part on the right-hand side corresponds to the model of acoustic transfer functions of Section 2.1.1 with the difference that here the near-end noise source with DOA θ is replaced by a diffuse noise source.

The digital signal processing part on the left-hand side is explained in the following: Using the clean far-end speech signal $s^{\text{in}}(k)$ with sample index k and the filtered near-end microphone signal $y(k)$ (see below) as input, the *near-end listening enhancement* (NELE) algorithm produces an enhanced speech signal $s^{\text{out}}(k)$. Loudspeaker equalization is then applied to flatten the overall transfer function to the ear such that the listener perceives the enhanced speech signal at

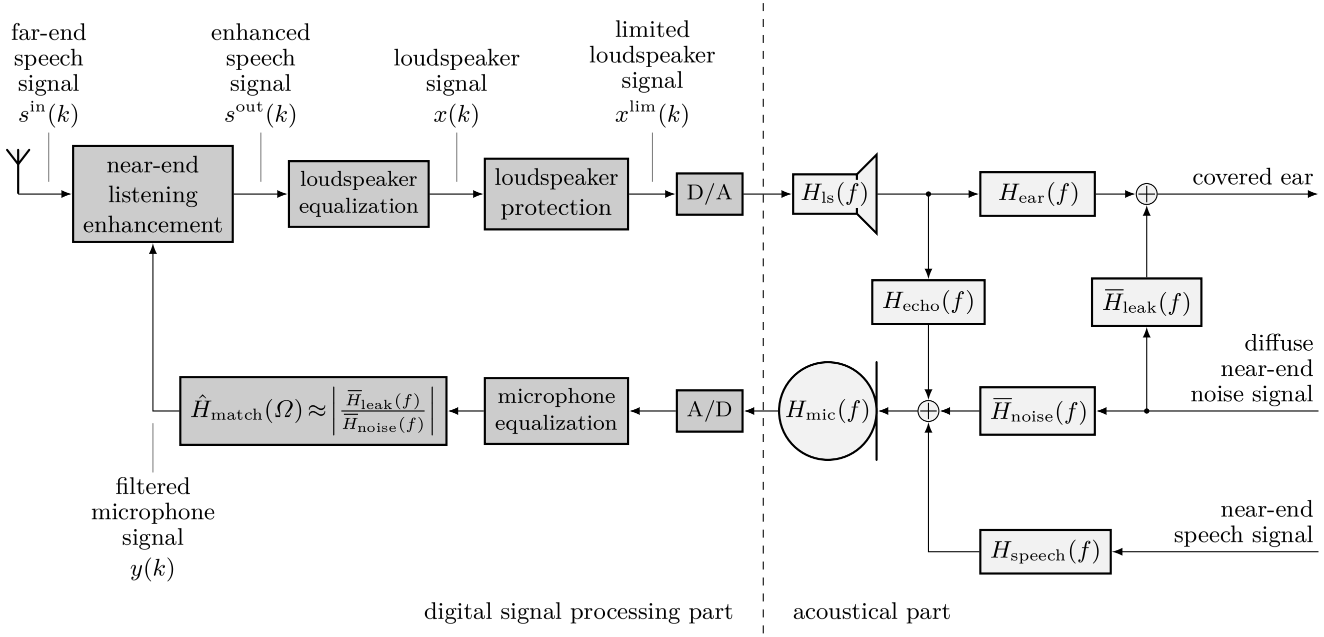

far-end speech signal $s^{\mathrm{in}}(k)$
near-end listening enhancement
enhanced speech signal $s^{\mathrm{out}}(k)$
loudspeaker equalization
loudspeaker signal $x(k)$
loudspeaker protection
limited loudspeaker signal $x^{\mathrm{lim}}(k)$
D/A
$H_{\mathrm{ls}}(f)$
$H_{\mathrm{ear}}(f)$
covered ear
$H_{\mathrm{echo}}(f)$
$\overline{H}_{\mathrm{leak}}(f)$
diffuse near-end noise signal
$\hat{H}_{\mathrm{match}}(\Omega) \approx \left|\frac{\overline{H}_{\mathrm{leak}}(f)}{\overline{H}_{\mathrm{noise}}(f)}\right|$
microphone equalization
A/D
$H_{\mathrm{mic}}(f)$
$\overline{H}_{\mathrm{noise}}(f)$
filtered microphone signal $y(k)$
near-end speech signal
$H_{\mathrm{speech}}(f)$
digital signal processing part
acoustical part

Figure 2.8: Complete system model with near-end listening enhancement, loudspeaker protection, and acoustic transfer functions.

the correct power with no or little coloration due to the loudspeaker. Afterwards and as a last step in digital domain, the *loudspeaker protection* (LOPRO) algorithm limits the loudspeaker signal $x(k)$ to prevent damage and failure of the loudspeaker. The limited loudspeaker signal $x^{\mathrm{lim}}(k)$ is – after digital-analog conversion – played by the loudspeaker.

In the microphone path, microphone equalization is applied to the microphone signal after analog-digital conversion with sampling rate f_{s} in order to equalize the microphone characteristic. The equalized microphone signal is filtered with $\hat{H}_{\mathrm{match}}(\mathit{\Omega})$, which is a digital filter estimate of $\overline{H}_{\mathrm{match}}(f) = \left|\frac{\overline{H}_{\mathrm{leak}}(f)}{\overline{H}_{\mathrm{noise}}(f)}\right|$ with $\mathit{\Omega} = \frac{2\pi f}{f_{\mathrm{s}}}$ denoting the normalized frequency in digital domain. As described above, this makes the filtered microphone signal $y(k)$ alike the noise signal at the ear and accounts for the mismatch of the noise field between the listener's ear and the microphone.

The filtered microphone signal $y(k)$ is finally used for noise estimation in the NELE algorithms as described in the beginning.

2.1.4 Simulation System Models

In Chapter 5, a model of the signal flow is used for the simulations, which is sketched in Figure 2.9a. Coming from the complete model of Figure 2.8, the following assumptions are made:

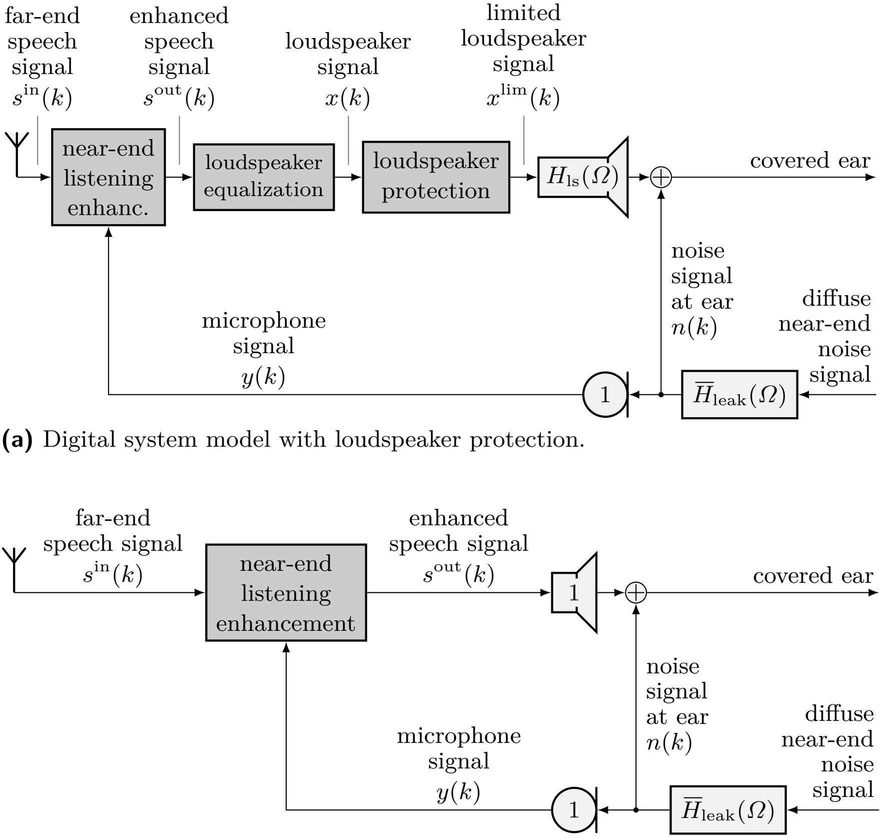
1. all signals are lowpass signals without components above the Nyquist frequency $\frac{f_{\mathrm{s}}}{2}$,
2. the echo path is zero, i. e., no echo occurs,
3. there is no interfering near-end speech signal in the simulation,
4. the transfer function $H_{\mathrm{ear}}(f)$ is implicitly included in $H_{\mathrm{ls}}(f)$ and compensated by the loudspeaker equalization,
5. microphone equalization works perfectly, and
6. the compensation of the noise field mismatch between microphone and ear works perfectly, i. e., $\hat{H}_{\mathrm{match}}(\mathit{\Omega}) = \left|\frac{\overline{H}_{\mathrm{leak}}(f)}{\overline{H}_{\mathrm{noise}}(f)}\right|$ with $\mathit{\Omega} = \frac{2\pi f}{f_{\mathrm{s}}}$ for $0 \leq f \leq \frac{f_{\mathrm{s}}}{2}$.

In the Chapters 3 and 4, a model of signal flow without LOPRO is used for the simulations, which is depicted in Figure 2.9b. It additionally includes the assumption, that

7. loudspeaker equalization works perfectly.

Discussion

The first assumption neglects high-frequency components above $\frac{f_{\mathrm{s}}}{2}$ of the analog near-end speech and noise signals. In the microphone path, both signals are lowpass filtered to $\frac{f_{\mathrm{s}}}{2}$ during analog-digital conversion anyway. In the acoustical part, the noise signal reaches the ear with full bandwidth, but, since the loudspeaker signal itself is a lowpass signal without components above $\frac{f_{\mathrm{s}}}{2}$, this assumption does not


far-end speech signal $s^{\mathrm{in}}(k)$
enhanced speech signal $s^{\mathrm{out}}(k)$
loudspeaker signal $x(k)$
limited loudspeaker signal $x^{\mathrm{lim}}(k)$
near-end listening enhanc.
loudspeaker equalization
loudspeaker protection
$H_{\mathrm{ls}}(\Omega)$
covered ear
noise signal at ear $n(k)$
diffuse near-end noise signal
microphone signal $y(k)$
1
$\overline{H}_{\mathrm{leak}}(\Omega)$
far-end speech signal $s^{\mathrm{in}}(k)$
near-end listening enhancement
enhanced speech signal $s^{\mathrm{out}}(k)$
1
covered ear
noise signal at ear $n(k)$
diffuse near-end noise signal
microphone signal $y(k)$
1
$\overline{H}_{\mathrm{leak}}(\Omega)$

(a) Digital system model with loudspeaker protection.

(b) Digital system model without loudspeaker protection.

Figure 2.9: System models used for simulation.

influence the intelligibility of the loudspeaker signal. It, however, allows to simulate the whole system in digital domain without the need for oversampling.

The second assumption is motivated in Section 2.1.1 with the quasi-"coupling" of the mobile phone to the covered ear.

A noisy environment has various effects on a dialog communication, which are not covered by the objective measures used for evaluation. Therefore, the model is restricted with the third assumption to the single-talk case, although the developed algorithms can easily cope with double-talk situations as described later.

In handset, headset, and hands-free mode, the frequency response of the loudspeaker is measured including the path to the ear (3GPP TS 26.132 2011) and compensated by the loudspeaker equalization (3GPP TS 26.131 2011). Accordingly, Assumption 4 holds for all applications with a mobile phone.

With the fifth assumption, the concatenation of microphone equalization and microphone transfer function $H_{\mathrm{mic}}(f)$ simplifies to a perfectly spectral flat characteristic of 1. While this is idealized for very low and very high frequencies, it is a reasonable assumption in the frequency range of interest.

With the sixth assumption, $\hat{H}_{\mathrm{match}}(\Omega)$ perfectly compensates $\left|\frac{\overline{H}_{\mathrm{noise}}(f)}{\overline{H}_{\mathrm{leak}}(f)}\right|$ independent of the diffuseness or the DOA of the noise. In the simulation, this allows to move $\overline{H}_{\mathrm{leak}}(\Omega)$ to the noise source. The filtering of the diffuse near-end noise signal thus yields the near-end noise signal $n(k)$ at the ear as well as the microphone signal $y(k)$. In a real application, especially with directional noise, this assumption results in an estimation error of the noise PSD at the listener's ear of 2 to 5 dB as shown in Section 2.1.2. Therefore, NELE algorithms should not rely on an exact estimate of the noise PSD.

Due to the seventh assumption (of the second model), the combination of loudspeaker equalization and loudspeaker transfer function $H_{\mathrm{ls}}(f)$ simplifies to a perfectly flat spectral characteristic of 1. In handset mode, this assumption is valid for frequencies between about 200 Hz and 5 kHz (see Figure 5.11), which is wider than the frequency range of narrow-band telephone speech. In hands-free mode, the assumption is also reasonable for frequencies above 500 Hz. Below about 300 Hz, it is idealized due to the distinct highpass characteristic of the speaker (see Figure 5.5), which is, however, of less importance as narrow-band telephone speech does not contain components in this frequency range.

2.1.5 Sound Pressure Calibration

In order to "map" between the acoustic sound pressure of the analog sound wave and the unit-less entities of digital signal processing, the microphone and loudspeaker path as well as the analysis filterbank must be calibrated and normalized, which is discussed in the following.

Microphone

The microphone converts the sound wave with sound pressure p_i to a proportional voltage (Vorländer 2008), which is amplified by the microphone amplifier. During analog-digital conversion, the amplified voltage is converted to a (quantized) digital signal and thereby inherently scaled. For a sufficiently large quantization word length, all three steps are assumed to be approximately linear. The (overall) proportionality factor between sound pressure and digital signal is denoted by g_{mic}. This factor with unit $^1/_{\mathrm{Pa}}$ depends on the microphone, the amplifier settings, and the analog-digital conversion. It must be calibrated during development of the mobile phone. As the mobile phone manufacturer must measure each product series anyway, g_{mic} is available at no additional cost.

Using g_{mic}, a reference power P_0 is defined as the (digital signal) power which originates from a sine wave with the reference sound pressure p_0 of 20 µPa:

$$P_0 = g_{\mathrm{mic}}^2 \cdot p_0^2 \, . \tag{2.7}$$

Filterbank

Instead of individual bandpass filters, a DFT analysis filterbank with window function $h(l)$, $l \in \{0, 1, \ldots, L-1\}$, of length L is used in this thesis to derive the subband signals. In order to acquire the correct subband power of a time-domain signal, it is important to compensate the "gain" of the analysis filterbank. Therefore, the real-valued normalization factor g_{fb} is calculated with the unit impulse sequence

$$\updelta(k) = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

as input of a DFT analysis filterbank without downsampling. The sum of the normalized energies of the filterbank output signals shall then be equal to the energy of the input:

$$\begin{aligned}
&\sum_{\mu=0}^{M-1} \sum_{k=-\infty}^{\infty} g_{\mathrm{fb}} \cdot \left| \sum_{l=0}^{L-1} \updelta(k-l) \cdot h(l) \cdot \exp\left\{-\mathrm{j}\tfrac{2\pi}{M}\mu l\right\} \right|^2 \\
&\quad = \sum_{\mu=0}^{M-1} \sum_{k=0}^{L-1} g_{\mathrm{fb}} \cdot \left| h(k) \cdot \exp\left\{-\mathrm{j}\tfrac{2\pi}{M}\mu k\right\} \right|^2 && (2.9) \\
&\quad = g_{\mathrm{fb}} \cdot \sum_{k=0}^{L-1} h^2(k) \cdot \sum_{\mu=0}^{M-1} \left| \exp\left\{-\mathrm{j}\tfrac{2\pi}{M}\mu k\right\} \right|^2 && (2.10) \\
&\quad = g_{\mathrm{fb}} \cdot \sum_{k=0}^{L-1} h^2(k) \cdot M && (2.11) \\
&\quad \overset{!}{=} \sum_{k=-\infty}^{\infty} \left| \updelta(k) \right|^2 = 1\,. && (2.12)
\end{aligned}$$

It finally follows that

$$g_{\mathrm{fb}} = \frac{1}{M \cdot \sum\limits_{l=0}^{L-1} h^2(l)}\,. \tag{2.13}$$

Loudspeaker

In the loudspeaker path, the digital-analog conversion converts the final loudspeaker signal to a voltage, which is amplified by the loudspeaker amplifier and converted to a sound wave by the loudspeaker. These three steps are again assumed to be approximately linear and the overall proportionality factor is compensated before digital-analog conversion.

2.2 Framework for Near-End Listening Enhancement

The framework inside the NELE block of Figures 2.8 and 2.9 is described in this section and depicted in Figure 2.10. In essence, a time-domain filtering is performed with filter coefficients calculated in the frequency-domain.

The far-end speech signal $s^{\text{in}}(k)$ is transformed in a (warped) analysis filterbank (see Section 2.2.1) to the sub-sampled DFT coefficients $\mathcal{S}_\mu^{\text{in}}(\kappa)$. The time index in the sub-sampled domain is given by

$$\kappa = \left\lfloor \frac{k}{R} \right\rfloor \cdot R \tag{2.14}$$

where $R \in \mathbb{N}$ is the downsampling rate.

Instead of calculating the spectral weights directly using the M complex-valued DFT coefficients $\mathcal{S}_\mu^{\text{in}}(\kappa)$ with DFT index $\mu \in \{0, 1, \ldots, M-1\}$, they are based on the corresponding $\frac{M}{2} + 1$ real-valued *subband* signals $s_i^{\text{in}}(k)$ with subband index $i \in \{0, 1, \ldots, \frac{M}{2}\}$ (for even M).

In theory, the subband signal $s_i^{\text{in}}(k)$ can be derived from a DFT analysis filterbank without sub-sampling by utilizing the complex conjugate symmetry $\mathcal{S}_\mu^{\text{in}}(k) = [\mathcal{S}_{M-\mu}^{\text{in}}(k)]^*$ of the DFT of the real-valued signal $s^{\text{in}}(k)$:

$$s_i^{\text{in}}(k) = \begin{cases} \mathcal{S}_i^{\text{in}}(k) & \text{if } i = 0 \text{ or } i = \frac{M}{2} \text{ for an even } M \\ \mathcal{S}_i^{\text{in}}(k) + \mathcal{S}_{M-i}^{\text{in}}(k) & \text{otherwise} \end{cases} \tag{2.15}$$

$$= \begin{cases} \mathcal{S}_i^{\text{in}}(k) & \text{if } i = 0 \text{ or } i = \frac{M}{2} \text{ for an even } M \\ \mathcal{S}_i^{\text{in}}(k) + \left[\mathcal{S}_i^{\text{in}}(k)\right]^* & \text{otherwise} \end{cases} \tag{2.16}$$

$$= g_{\text{sym},i} \cdot \operatorname{Re}\left\{\mathcal{S}_i^{\text{in}}(k)\right\} \tag{2.17}$$

with the DFT symmetry factor

$$g_{\text{sym},i} = \begin{cases} 1 & \text{if } i = 0 \text{ or } i = \frac{M}{2} \text{ for an even } M \\ 2 & \text{otherwise}\,. \end{cases} \tag{2.18}$$

In practice, only the short-term subband power estimates $\hat{P}_{s,i}^{\text{in}}(\kappa)$ of the real-valued subband signals $s_i^{\text{in}}(k)$ are needed to calculate the subband weights $W_i(\kappa)$. These estimates $\hat{P}_{s,i}^{\text{in}}(\kappa)$ can, however, be calculated with the *sub-sampled* DFT coefficients $\mathcal{S}_\mu^{\text{in}}(\kappa)$ as shown in Section 2.2.3.

The near-end microphone signal $y(k)$, which is a mixture of the near-end noise signal $n(k)$ and the interfering near-end speech signal, is analogously transformed to DFT coefficients $\mathcal{Y}_\mu(\kappa)$. Since the interfering near-end speech signal should not be considered during NELE, a noise tracking algorithm is used to estimate the short-term subband power estimates $\hat{P}_{n,i}(\kappa)$ of the near-end noise signal, which is discussed in Section 2.2.4.

The subband weights $W_i(\kappa)$ are calculated based on both subband power estimates $\hat{P}_{s,i}^{\text{in}}(\kappa)$ and $\hat{P}_{n,i}(\kappa)$. The choice of the subband weights resembles the

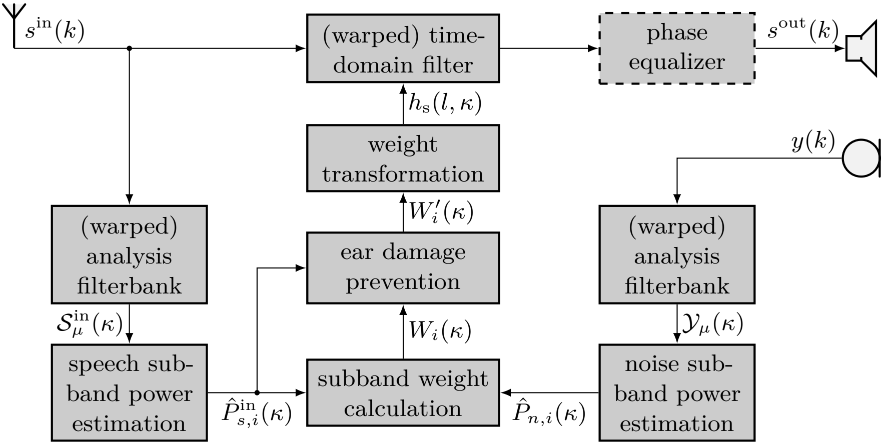

$s^{\mathrm{in}}(k)$
(warped) time-domain filter
phase equalizer
$s^{\mathrm{out}}(k)$
$h_{\mathrm{s}}(l,\kappa)$
weight transformation
$y(k)$
$W'_i(\kappa)$
(warped) analysis filterbank
ear damage prevention
(warped) analysis filterbank
$\mathcal{S}^{\mathrm{in}}_{\mu}(\kappa)$
$W_i(\kappa)$
$\mathcal{Y}_{\mu}(\kappa)$
speech sub-band power estimation
subband weight calculation
noise sub-band power estimation
$\hat{P}^{\mathrm{in}}_{s,i}(\kappa)$
$\hat{P}_{n,i}(\kappa)$

Figure 2.10: Framework for NELE with sample index k, sub-sampled time index κ, DFT index μ, and subband index i.

"core" of the NELE algorithm and is content of the Chapters 3 and 4. In order to prevent damage of the listener's ear, the subband weights $W_i(\kappa)$ are limited as described in Section 2.2.5. The limited subband weights $W'_i(\kappa)$ are transformed to the coefficients $h_{\mathrm{s}}(l,\kappa)$ of a (warped) time-domain filter, which is applied to the far-end speech signal $s^{\mathrm{in}}(k)$ (see Section 2.2.1). As a side effect of the frequency warping of the filter, the phase of the signal is altered. Since the human ear is quite insensitive towards phase modifications (Zwicker & Fastl 1999), these phase distortions are mostly tolerable for speech processing. However, strong, audible modifications may need to be "corrected" with a so-called phase equalizer. The final enhanced far-end speech signal $s^{\mathrm{out}}(k)$ is played back on the loudspeaker.

2.2.1 Filterbank Equalizer

In this thesis, the warped *filterbank equalizer* (FBE) (Vary 2006; Löllmann & Vary 2007) is utilized. Opposed to the DFT analysis-synthesis filterbank (AS FB), which is conventionally used for speech enhancement, this structure easily allows a processing with approximately Bark-scaled spectral resolution according to the human auditory system. Additionally, this concept separates filter calculation from signal modification, which avoids the need for a signal re-synthesis and features a (very) low signal delay.

It was shown in (Sauert et al. 2008) that the results achieved with the common AS FB with an appropriate, i. e., larger DFT size are comparable to the FBE but require a higher delay.

It should be noted that only an overview of the FBE is given here. Prototype filter design, efficient implementations using the fast Fourier transform (FFT), and many other aspects of this concept are treated in (Löllmann 2011; Löllmann & Vary 2007; Vary 2006) in detail.

Concept of Uniform FBE

The (clean) far-end speech signal $s^{\text{in}}(k)$ and the near-end microphone signal $y(k)$ with sample index k are split into M DFT coefficients $\mathcal{S}^{\text{in}}_{\mu}(\kappa)$ and $\mathcal{Y}_{\mu}(\kappa)$ by means of a DFT analysis filterbank with downsampling:

$$\mathcal{S}^{\text{in}}_{\mu}(\kappa) = \sum_{l=0}^{L-1} s^{\text{in}}(\kappa - l) \cdot h(l) \cdot \exp\left\{-\mathrm{j}\tfrac{2\pi}{M}\mu l\right\}, \tag{2.19}$$

$$\mathcal{Y}_{\mu}(\kappa) = \sum_{l=0}^{L-1} y(\kappa - l) \cdot h(l) \cdot \exp\left\{-\mathrm{j}\tfrac{2\pi}{M}\mu l\right\}. \tag{2.20}$$

The time index in the sub-sampled domain is given by $\kappa = \lfloor k/R \rfloor \cdot R$ with the downsampling rate $R \in \mathbb{N}$. The real-valued impulse response of the prototype filter of length L is denoted by $h(l)$ and chosen according to (Löllmann & Vary 2007) using a Hann window sequence. Note, that reasonable choices for L are multiples of the DFT size M, e. g., $L = M$.

The subband signals $\mathcal{S}^{\text{in}}_{\mu}(\kappa)$ and $\mathcal{Y}_{\mu}(\kappa)$ are used to estimate the subband powers and to calculate the spectral weights $W'_i(\kappa)$ as described later in Chapters 3 and 4. Filtering the far-end speech signal $s^{\text{in}}(k)$ with the time-varying coefficients $h_{\text{s}}(l,\kappa)$ yields the enhanced speech signal $s^{\text{out}}(k)$:

$$s^{\text{out}}(k) = \sum_{l=0}^{L-1} s^{\text{in}}(k - l) \cdot h_{\text{s}}(l,\kappa). \tag{2.21}$$

This *single* time-domain filter is obtained by a generalized DFT (GDFT) of the spectral weights $W'_{\mu}(\kappa)$[2] according to

$$h_{\text{s}}(l,\kappa) = h(l) \cdot \sum_{\mu=0}^{M-1} W'_{\mu}(\kappa) \cdot \exp\left\{-\mathrm{j}\tfrac{2\pi}{M}\mu(l - l_0)\right\} \tag{2.22}$$

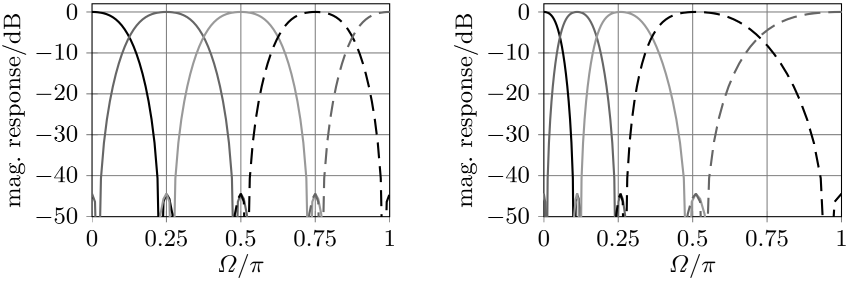
with filter tap $l \in \{0, 1, \ldots, L-1\}$ and filter delay $l_0 = \frac{L-1}{2}$.

Non-Uniform FBE

The FBE with non-uniform time-frequency resolution is designed by means of an allpass transformation. In the process, the delay elements of the discrete filters are replaced by (causal) allpass filters of first order

$$z^{-1} \rightarrow H_{\text{A}}(z) = \frac{z^{-1} - a}{1 - a\, z^{-1}} \tag{2.23}$$

[2] The "DFT weights" $W'_{\mu}(\kappa)$ with $0 \le \mu \le M-1$ consist of the "subband weights" $W'_i(\kappa)$ with $0 \le i \le \frac{M}{2}$ and their complex conjugate symmetric extension $[W'_{M-i}(\kappa)]^*$.



(a) Uniform filters.

(b) Warped filters.

Figure 2.11: Magnitude responses of uniform and warped ($a = 0.4$) DFT analysis filterbanks with $M = 8$ frequency bands.

with $-1 < a < 1$ being the allpass coefficient. Due to this allpass transformation, the uniform bandpass filters are converted into warped bandpass filters, which is illustrated in Figure 2.11. The allpass transformation accomplishes a variation of the bandwidths without changing certain filter properties such as stop-band attenuation. An allpass pole of $a = 0.4$ yields a good approximation of the Bark frequency scale at the considered sampling rate of $f_\mathrm{s} = 8\,\mathrm{kHz}$, cf. (Smith & Abel 1999).

Unfortunately, the allpass transformation changes not only the magnitude but also the phase response of the filters. This undesirable effect can be compensated by applying a phase equalizer to the output signal of the FBE (Löllmann 2011; Löllmann & Vary 2007). However, such phase equalization is not necessarily needed since the human auditory system is quite insensitive against phase modifications (Zwicker & Fastl 1999).

2.2.2 Handling of Not-Contributing Subbands

In mobile telephony, some subbands can, by system design, not contribute to the listening experience and thus to intelligibility:

- The first contributing subband $i_\mathrm{f} \geq 1$ is usually determined by the lower cut-off frequencies of the mobile phone's microphone and loudspeaker, which can be 150 Hz in handset mode (NXP 2010b) and 400 Hz in hands-free mode (Knowles 2011; NXP 2010a). Another restriction is given by the utilized speech codecs as, e. g., (3GPP TS 26.090 2009; 3GPP TS 26.190 2009) and transmission characteristics of the mobile phone, cf. (ITU-T P.310 2009; ITU-T P.311 2005).

- The last contributing subband i_l is bounded by the corresponding upper cut-off frequencies, which can be as low as 3.4 kHz for a narrow-band phone

(ITU-T P.310 2009) and ca. 7 kHz for a wide-band phone (ITU-T P.311 2005), as well as the Nyquist frequency of the digital processing system.

To sum up, the subband signals with index $i < i_\mathrm{f}$ and $i > i_\mathrm{l}$ have either not been transmitted over the telephone network or lie beyond the capabilities of the sound reproduction system. Either way, they can not be played back and can not be perceived by the near-end listener. Thus, all audio power spent in these subbands is wasted and, accordingly, the spectral weights in theses subbands are set to zero

$$W_i(\kappa) = 0 \qquad \forall \quad i < i_\mathrm{f} \vee i > i_\mathrm{l}\,. \tag{2.24}$$

All further spectral weight processing is only performed for the subbands with index $i_\mathrm{f} \leq i \leq i_\mathrm{l}$.

In this thesis, the contributing subbands cover the frequency range from 120 Hz to the Nyquist frequency for the given sampling rate f_s:

$$i_\mathrm{f} = \min\left\{\, i \;\middle|\; f_{\mathrm{h},i} > 120\,\mathrm{Hz} \,\right\}, \tag{2.25}$$

$$i_\mathrm{l} = \max\left\{\, i \;\middle|\; f_{\mathrm{l},i} < \tfrac{f_\mathrm{s}}{2} \,\right\}, \tag{2.26}$$

where $f_{\mathrm{h},i}$ and $f_{\mathrm{l},i}$ denote the upper and lower limiting frequency of the i-th subband, respectively. This way, the studies are independent of a specific speech codec or mobile phone, undesired effects at frequency subbands without speech content are avoided, and still an upper bound of the overall performance of the algorithms is compared.

2.2.3 Speech Subband Power Estimation

This section describes the calculation of the short-term subband power estimates $\hat{P}^\mathrm{in}_{s,i}(\kappa)$ of the far-end speech signal $s^\mathrm{in}(k)$ which is based on the moving average of the power of the subband signals in the past "speech segments" with voice activity using a look-back of (total) length τ_s in seconds (Sauert & Vary 2011).

For each speech signal segment of length R, i. e., each update interval κ, the voice activity according to the voice activity detector (VAD) of the G.729 codec (ITU-T G.729 2007) is determined. The time indices of the preceding $\frac{\tau_s \cdot f_\mathrm{s}}{R} \in \mathbb{N}$ segments with voice activity are collected in the set $\mathbb{K}_s(\kappa)$. The short-term speech subband power estimate $\hat{P}^\mathrm{in}_{s,i}(\kappa)$ is then calculated as the arithmetic mean over $\mathbb{K}_s(\kappa)$ of the squared magnitudes of $\mathcal{S}^\mathrm{in}_i(\kappa)$:

$$\hat{P}^\mathrm{in}_{s,i}(\kappa) = \underset{\zeta \in \mathbb{K}_s(\kappa)}{\mathrm{mean}}\; g_{\mathrm{sym},i} \cdot g_\mathrm{fb} \cdot \left|\mathcal{S}^\mathrm{in}_i(\zeta)\right|^2, \quad i_\mathrm{f} \leq i \leq i_\mathrm{l}\,, \tag{2.27}$$

where the symmetry factor $g_{\mathrm{sym},i}$ utilizes the complex conjugate symmetry of the DFT and the normalization factor g_fb achieves an analysis filterbank with approximately 0 dB gain.

The duration τ_s determines the memory of the speech subband power estimator. Too small values result in a high variance of the estimate and, thus, a fast and unpleasant fluctuation of the spectral weights. With a too large τ_s, the system can only slowly adapt to changes in intensity and spectral envelope of the far-end signal. In the following, the setting $\tau_s = 2\,\mathrm{s}$ is used (see Appendix A).

Note, that in a real implementation a more sophisticated approach might be necessary to cope with sudden changes in the far-end signal.

2.2.4 Noise Subband Power Estimation

To derive the short-term subband power estimates $\hat{P}_{n,i}(\kappa)$ of the near-end noise signal, two algorithms are examined (see also Sauert & Vary 2011):

1. the Minimum Statistics algorithm (Martin 2001, 2006) in the implementation of Brookes (2012) and
2. a minimum mean-square error (MMSE) based noise PSD tracking algorithm (Hendriks et al. 2010a) in an implementation provided by the authors (Hendriks et al. 2010b).

Quite remarkably, both algorithms perform out of the box equally well with the uniform as well as the non-uniform analysis filterbank of the FBE. In general, both algorithms are comparable in terms of average noise subband power estimate for most quasi-stationary noise signals. However, the MMSE based algorithm tends to track non-stationary noise as well as speech babble noise better and faster than the Minimum Statistics algorithm. Furthermore, it seems to cope better with interfering near-end speech.

Both algorithms are also compared to a simple moving average algorithm as of Section 2.2.3 with a memory of length τ_n in seconds, where the short-term noise subband power estimate is calculated as the arithmetic mean of the squared magnitudes of the subband signal $\mathcal{N}_i(\kappa)$ during the preceding $\frac{\tau_n \cdot f_\mathrm{s}}{R} \in \mathbb{N}$ update intervals:

$$\hat{P}_{n,i}(\kappa) = \underset{\zeta \in \mathbb{K}_n(\kappa)}{\mathrm{mean}}\; g_{\mathrm{sym},i} \cdot g_{\mathrm{fb}} \cdot \left|\mathcal{N}_i(\zeta)\right|^2 , \quad i_\mathrm{f} \le i \le i_\mathrm{l} , \tag{2.28}$$

with the set $\mathbb{K}_n(\kappa) = \left\{\kappa, \kappa - R, \kappa - 2R, \ldots, \kappa - \left(\frac{\tau_n \cdot f_\mathrm{s}}{R} - 1\right) \cdot R\right\}$. As this moving average algorithm interprets interfering near-end speech in double-talk situations as noise, it is, however, not suitable for most real-world applications.

2.2.5 Hearing Damage Prevention

The spectral weights are limited before filtering of the far-end speech signal in order to prevent damage of the listener's ear or pain.

In (Zwicker & Fastl 1999, Chapter 6), several listening experiments are described, which all show that the human ear integrates acoustic power over the critical bandwidths[3]. Therefore, it seems reasonable to limit the acoustic power also in critical bands in order to prevent hearing damage and pain.

[3] The critical band scale is also called Bark scale as proposed by Zwicker (1961).

As an approximation of this concept, each subband weight is restricted such that the resulting short-term subband power of the enhanced speech signal does not exceed a maximum subband power $P_s^{\max}$:

$$W_i'(\kappa) = \min\left\{W_i(\kappa),\, W_i^{\max}(\kappa)\right\} \quad \forall\, i_{\mathrm{f}} \le i \le i_{\mathrm{l}} \tag{2.29}$$

with the maximum subband weight

$$W_i^{\max}(\kappa) = \sqrt{\frac{P_s^{\max}}{\hat{P}_{s,i}^{\mathrm{in}}(\kappa)}}\,. \tag{2.30}$$

In accordance with (Zwicker & Fastl 1999, Figure 2.1), the value

$$10\log\left\{\frac{P_s^{\max}}{P_0}\right\} = 95\,\mathrm{dB_{SPL}} \tag{2.31}$$

is chosen.

2.3 Measures

This section describes the two objective measures which are used in this thesis to evaluate speech intelligibility: the Speech Intelligibility Index and a modified Speech Transmission Index.

2.3.1 Speech Intelligibility Index (SII)

The Speech Intelligibility Index (SII) (ANSI S3.5 1997) is a standardized objective measure which is correlated with the intelligibility of speech under a variety of adverse listening conditions.

The base of the SII calculation is the equivalent[4] (far-end) speech spectrum level E_i and the equivalent[4] (near-end) noise spectrum level N_i, where i denotes the subband index. The speech spectrum level can be calculated as

$$E_i = 10\log\left\{\frac{\dfrac{p_i^2}{f_{\Delta,i}}}{\dfrac{p_0^2}{1\,\mathrm{Hz}}}\right\}, \quad i_{\mathrm{f}} \le i \le i_{\mathrm{l}}\,, \tag{2.32}$$

where p_i^2 is the time-mean-square sound pressure of the speech in the i-th subband, measured through a bandpass filter with center $f_{\mathrm{c},i}$ and bandwidth $f_{\Delta,i}$ (ANSI S3.5 1997). The reference sound pressure of 20 µPa is denoted by p_0.

The SII is calculated in its finest spectral resolution according to the so-called critical band procedure, i. e., in the first 21 critical bands ranging from 100 Hz to 9.5 kHz.

[4] The equivalent spectrum level is defined as the spectrum level measured at the point corresponding to the center of the listener's head, with the listener absent, under the reference communication situation (ANSI S3.5 1997). In the following, the term "equivalent" is omitted from all spectrum levels for the sake of brevity.

Calculation of Spectrum Levels

In this thesis, the speech spectrum level E_i is computed for a given (digital) time-domain speech signal $s(k)$ as

$$E_i = 10 \log \left\{ \frac{\dfrac{\hat{P}_{s,i}}{f_{\Delta,i}}}{\dfrac{P_0}{1\,\mathrm{Hz}}} \right\}, \quad i_\mathrm{f} \leq i \leq i_\mathrm{l}\,, \tag{2.33}$$

where $\hat{P}_{s,i}$ denotes the estimate of the speech signal's subband power and P_0 is the digital reference power corresponding to the reference sound pressure of 20 µPa. The frequency bandwidth of the i-th critical band is given as

$$f_{\Delta,i} = f_{\mathrm{h},i} - f_{\mathrm{l},i} \tag{2.34}$$

with the upper and lower limiting frequency $f_{\mathrm{h},i}$ and $f_{\mathrm{l},i}$, respectively.

During *processing* with the NELE algorithm, the speech and noise subband power estimates are computed once per update interval using the estimators described in Sections 2.2.3 and 2.2.4.

To *evaluate* speech intelligibility after processing, the complete (processed) signal is partitioned in non-overlapping frames of 20 ms length. Those frames which contain voice activity according to the VAD of the G.729 codec (ITU-T G.729 2007) are selected and denoted with frame index m. These frames are multiplied with a Hann window and transformed to DFT coefficients $\mathcal{S}_\mu(m)$.

The speech subband power $\hat{P}_{s,i}$ is estimated as the sum of the squared magnitudes of all $\mathcal{S}_\mu(m)$ within the subband number i, averaged over all frames m with voice activity:

$$\hat{P}_{s,i} = \operatorname*{mean}_{m} \sum_{\mu \in \mathbb{M}_i} g_{\mathrm{sym},\mu} \cdot g_\mathrm{fb} \cdot \left|\mathcal{S}_\mu(m)\right|^2, \quad i_\mathrm{f} \leq i \leq i_\mathrm{l}\,, \tag{2.35}$$

with the set of the DFT indices in the i-th subband

$$\mathbb{M}_i = \left\{ \mu \in \mathbb{N}_0 \;\middle|\; f_{\mathrm{l},i} \leq \mu \cdot \frac{f_\mathrm{s}}{M} < f_{\mathrm{h},i} \wedge \mu \leq \frac{M}{2} \right\}. \tag{2.36}$$

As described above, the symmetry factor $g_{\mathrm{sym},i}$ utilizes the complex conjugate symmetry of the DFT and g_fb denotes the filterbank normalization factor. The DFT length $M = 512$ for sampling rate $f_\mathrm{s} = 8\,\mathrm{kHz}$ and $M = 1024$ for $f_\mathrm{s} = 16\,\mathrm{kHz}$ was found to be sufficient to avoid "quantization effects" due to the non-uniform subbands.

The noise spectrum level N_i is obtained analogously for the time-domain noise signal $n(k)$.

Calculation of the Speech Intelligibility Index

The following steps have to be performed for each of the 21 subbands to calculate the SII according to (ANSI S3.5 1997):

1. Determine the self-speech masking spectrum level

$$V_i(E_i) = E_i - 24\,\mathrm{dB}\,, \tag{2.37}$$

which considers the masking of higher frequency (far-end) speech components by lower frequency (far-end) speech components.

2. The masking spectrum level $Z_i(N_i)$ corresponds to the overall masking of the (far-end) speech, which includes within-band masking and out-of-band masking by the (near-end) noise as well as self-speech masking by the (far-end) speech. It therefore depends on the larger of the noise spectrum level N_i and the self-speech masking spectrum level $V_i(E_i)$.

   As, for the application of NELE, only those situations with significant background noise are of interest, it is assumed that N_i is always greater than $V_i(E_i)$. Thus, the overall masking of the (far-end) speech is only produced by the (near-end) noise, i. e., the noise spectrum level N_i, and is calculated as

$$Z_i(N_i) = 10\log\left\{10^{N_i/10} + \sum_{\zeta=1}^{i-1} 10^{\left[N_\zeta + 3.32 C_\zeta(N_\zeta)\log\left(f_{\mathrm{c},i}/f_{\mathrm{h},\zeta}\right)\right]/10}\right\} \tag{2.38}$$

   with the slope per octave of the spread of masking

$$C_i(N_i) = -80\,\mathrm{dB} + 0.6\left[N_i + 10\log(f_{\mathrm{h},i} - f_{\mathrm{l},i})\right]. \tag{2.39}$$

3. Determine the disturbance spectrum level $D_i(N_i)$ as the larger of the masking spectrum level $Z_i(N_i)$ and the internal noise spectrum level, which accounts for the threshold of hearing. With the same assumption of sufficiently high background noise as above, the disturbance spectrum level is

$$D_i(N_i) = Z_i(N_i)\,. \tag{2.40}$$

   Note, that the dependency of $D_i(N_i)$ on N_i is not written down in the following for the sake of brevity.

4. Determine the speech level distortion factor $L_i(E_i)$:

$$L_i(E_i) = \begin{cases} 1 & \text{if } E_i \leq U_i + 10\,\mathrm{dB} \\ 1 - \frac{E_i - U_i - 10\,\mathrm{dB}}{160\,\mathrm{dB}} & \text{if } U_i + 10\,\mathrm{dB} < E_i \leq U_i + 170\,\mathrm{dB} \\ 0 & \text{if } U_i + 170\,\mathrm{dB} < E_i\,, \end{cases} \tag{2.41}$$

   which considers the decrease in intelligibility caused by the distortion due to a high presentation level. The standard speech spectrum level at normal voice effort U_i is fixed and can be found in (ANSI S3.5 1997, Table 1). It has its maximum value of 34.75 dB in the second critical band with $f_{\mathrm{c},2} = 250\,\mathrm{Hz}$.

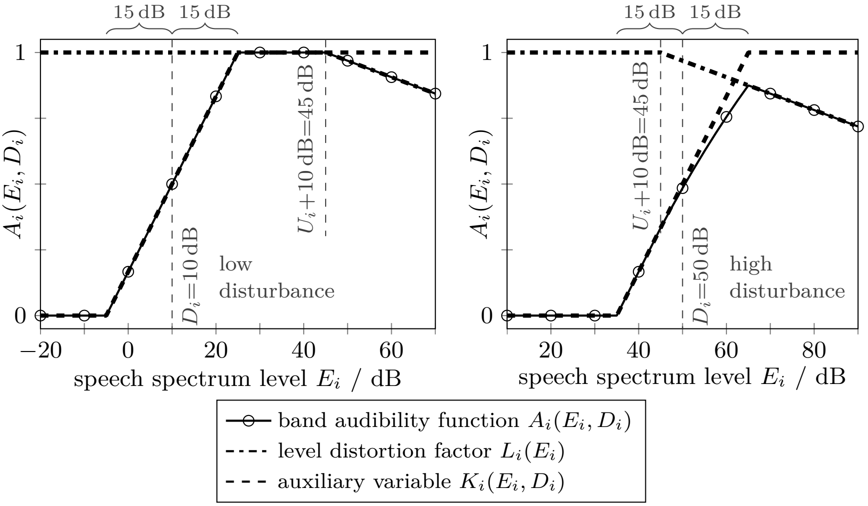


Figure 2.12: Exemplary plots of contributions to the band audibility function for low as well as high disturbance case.

5. Determine the band audibility function $A_i(E_i, D_i)$

$$A_i(E_i, D_i) = L_i(E_i) \cdot K_i(E_i, D_i) \tag{2.42}$$

using the auxiliary variable[5] $K_i(E_i, D_i)$

$$K_i(E_i, D_i) = \begin{cases} 0 & \text{if } E_i \leq D_i - 15\,\text{dB} \\ \frac{E_i - D_i + 15\,\text{dB}}{30\,\text{dB}} & \text{if } D_i - 15\,\text{dB} < E_i \leq D_i + 15\,\text{dB} \\ 1 & \text{if } D_i + 15\,\text{dB} < E_i\,. \end{cases} \tag{2.43}$$

The auxiliary variable $K_i(E_i, D_i)$ considers the loss of intelligibility due to the fact that the speech signal is masked by the noise. The band audibility function $A_i(E_i, D_i)$ specifies the effective proportion of the speech dynamic range within the subband that contributes to speech intelligibility.

Figure 2.12 shows the contributions to the band audibility function exemplarily for a low and a high disturbance.

Finally, the Speech Intelligibility Index $S(\underline{E}, \underline{D})$ is calculated as

$$S(\underline{E}, \underline{D}) = \sum_{i=1}^{21} I_i \cdot A_i(E_i, D_i)\,, \tag{2.44}$$

[5] In (ANSI S3.5 1997), K_i is called "temporary variable".

where $\underline{E}$ denotes the vector $(E_1, E_2, \ldots, E_{21})$ of all spectrum levels. The band importance function I_i (see ANSI S3.5 1997, Table 1) characterizes the relative significance of the subband to speech intelligibility. Since

$$\sum_{i=1}^{21} I_i = 1 \quad \text{and} \quad 0 \leq A_i(E_i, D_i) \leq 1\,, \tag{2.45}$$

the SII can take values from zero to one. Communication systems with an SII of $S(\underline{E}, \underline{D}) \geq 0.75$ are considered to be good, those with $S(\underline{E}, \underline{D}) \leq 0.45$ poor.

2.3.2 Speech Transmission Index (STI)

The Speech Transmission Index (STI) (Houtgast & Steeneken 1971; Houtgast et al. 2002; IEC 60268-16 2003; Steeneken & Houtgast 1980) is a well established intelligibility measure in room acoustics and for many types of transmission channels. In principle, it is based on the reduction of signal intensity modulation in frequency subbands.

For the traditional STI method, multiple bandpass filtered and intensity modulated probe stimuli are transmitted over the channel under consideration. The modulation depth of the stimulus and the response is compared for each frequency band and modulation frequency, mapped to a signal-to-noise ratio, and averaged to a single STI value between zero and one. Table 2.1 shows the quality rating of STI values according to (IEC 60268-16 2003). A summary of this method can be found, e. g., in (Goldsworthy & Greenberg 2004).

Two “extensions” of the traditional STI method have been proposed in literature: Firstly, the *revised Speech Transmission Index* (STI_r) is presented in (Steeneken & Houtgast 1991, 1999). It extends the STI with a redundancy correction, which accounts for the correlation of information between adjacent frequency bands. Furthermore, a separate assessment of male and female speech is introduced. The revised rule can also be found in (Houtgast et al. 2002; IEC 60268-16 2003).

Secondly, several variations were developed over time that use speech signals as stimuli instead of artificial probe stimuli. In (Goldsworthy & Greenberg 2004), various speech-based STI methods are analyzed and simple modifications are

STI value	quality
0.75 to 1	excellent
0.6 to 0.75	good
0.45 to 0.6	fair
0.3 to 0.45	poor
0 to 0.3	bad

Table 2.1: Mapping between STI values and quality rating according to (IEC 60268-16 2003).

proposed: “These modified STI methods are well correlated with the traditional STI for additive noise and reverberation and also exhibit qualitatively reasonable behavior for selected nonlinear operations. As a result, the modified STI methods are promising candidates to predict intelligibility of nonlinearly processed speech” (Goldsworthy & Greenberg 2004). Among these methods, the so-called envelope regression method is preferred, as it produces comparable results for conventional acoustical degradations as well as the considered nonlinear operations with less computational complexity.

Calculation of Speech-Based Revised Speech Transmission Index ($\mathrm{STI}_{\mathrm{sr}}$)

In this thesis, these two “extensions” are combined to a speech-based method with the higher prediction accuracy of the redundancy correction between adjacent frequency bands. This combination, which will be called *speech-based revised Speech Transmission Index* ($\mathrm{STI}_{\mathrm{sr}}$) in the following, consists of the envelope regression method, presented in (Goldsworthy & Greenberg 2004), with the redundancy correction of the improved $\mathrm{STI}_{\mathrm{r}}$.

The (delay compensated) probe and response signals are split into seven octave bands with center frequencies ranging from 125 Hz to 8 kHz using eighth-order Butterworth bandpass filters. The intensity envelopes $\breve{x}_i(k)$ and $\breve{y}_i(k)$ of the probe resp. response bandpass signals are calculated for each frequency band i by squaring the bandpass-filtered signals, lowpass filtering with an eighth-order Butterworth filter with 50 Hz cutoff frequency, and downsampling to 200 Hz.

For each frequency band, the modulation metric M_i is calculated as

$$M_i = \breve{\beta}_i \frac{\lambda_{\breve{x}\breve{y},i}}{\lambda_{\breve{x},i}} \tag{2.46}$$

with the normalization term $\breve{\beta}_i$, the covariance $\lambda_{\breve{x}\breve{y},i}$ between $\breve{x}_i(k)$ and $\breve{y}_i(k)$, and the variance $\lambda_{\breve{x},i}$ of $\breve{x}_i(k)$. The normalization term $\breve{\beta}_i$ considers the powers of the probe and response signals and is calculated as

$$\breve{\beta}_i = \frac{\mu_{\breve{x},i}}{\mu_{\breve{x},i} + \mu_{\breve{z},i}}\,, \tag{2.47}$$

where $\mu_{\breve{x},i}$ denotes the mean of the probe intensity envelope $\breve{x}_i(k)$ and $\mu_{\breve{z},i}$ is the mean of the noise envelope $\breve{z}_i(k) = |\breve{y}_i(k) - \breve{x}_i(k)|$. As a practical extension to (Goldsworthy & Greenberg 2004), $\breve{\beta}_i$ is set to zero in the unlikely case $\mu_{\breve{x},i} \leq 0$, which prevents a division by zero as well as negative normalization terms.

The (co)variances are calculated as an unbiased estimate, i. e.,

$$\lambda_{\breve{x}\breve{y},i} = \mathrm{E}\Big\{ \big(\breve{x}_i(k) - \mu_{\breve{x},i}\big)\big(\breve{y}_i(k) - \mu_{\breve{y},i}\big) \Big\} \tag{2.48}$$

and

$$\lambda_{\breve{x},i} = \mathrm{E}\Big\{ \big(\breve{x}_i(k) - \mu_{\breve{x},i}\big)^2 \Big\} \tag{2.49}$$

with $\mu_{\breve{y},i}$ denoting the mean of the response intensity envelope $\breve{y}_i(k)$.

In analogy to the $\mathrm{STI_r}$, a corrected modulation metric M_i' is calculated for each frequency band to consider the auditory spread of masking and the hearing threshold:

$$M_i' = M_i \cdot \frac{I_{\breve{y},i}}{I_{\breve{y},i} + I_{\mathrm{am},i} + I_{\mathrm{rs},i}}\,, \tag{2.50}$$

where

$$I_{\breve{y},i} = 10 \log \left\{ \frac{\operatorname*{mean}_k \left\{ \breve{y}_i(k) \right\}}{P_0} \right\} \tag{2.51}$$

represents the signal intensity, i. e., the power of the response bandpass signals, with the reference power P_0. $I_{\mathrm{am},i}$ denotes the intensity level of auditory masking and $I_{\mathrm{rs},i}$ accounts for the absolute hearing threshold. The values and calculation rules of $I_{\mathrm{am},i}$ and $I_{\mathrm{rs},i}$ can be reviewed in (IEC 60268-16 2003) and (Houtgast et al. 2002).

The apparent signal-to-noise ratio $aSNR_i$ and the transmission index TI_i for each frequency band are defined as

$$aSNR_i = 10 \log \left\{ \frac{M_i'}{1 - M_i'} \right\} \tag{2.52}$$

and

$$TI_i = \begin{cases} 0 & \text{if } aSNR_i \leq -15\,\mathrm{dB} \\ \frac{aSNR_i + 15\,\mathrm{dB}}{30\,\mathrm{dB}} & \text{if } -15\,\mathrm{dB} < aSNR_i \leq 15\,\mathrm{dB} \\ 1 & \text{if } 15\,\mathrm{dB} < aSNR_i\,. \end{cases} \tag{2.53}$$

Finally, the $\mathrm{STI_{sr}}$ is obtained by a weighted sum of the transmission indices for all seven octave bands and the corresponding redundancy correction:

$$STI_{\mathrm{sr}} = \sum_{i=1}^{7} \alpha_i\, TI_i - \sum_{i=1}^{6} \beta_i \sqrt{TI_i \cdot TI_{i+1}}\,, \tag{2.54}$$

with the octave-weighting factors α_i and the so-called redundancy correction factors β_i. The weighting and redundancy factors are different for male and female speech and can be found in (IEC 60268-16 2003) and (Houtgast et al. 2002).

The $\mathrm{STI_{sr}}$ can have values between zero and one, just as the traditional STI, and the same quality ratings of Table 2.1 are applied in this thesis.

2.3.3 SII Gain and STI Gain

To facilitate the comparison of the performance of the NELE algorithms, the concept of *SII gain* is introduced as the difference in dB between the input SNR which the *unprocessed* speech needs to yield an SII of 0.75 and the input SNR which the *processed* speech needs with the same noise. In other words the SII gain is the amount in decibels by which the input SNR may be decreased due to the processing while still retaining a “good” communication system. A positive SII gain implies thus an improvement, while a negative gain implies degradation.

The *STI gain* is defined analogously as the amount in decibels by which the input SNR may be decreased with processing compared to unprocessed speech while still retaining a “good” $\mathrm{STI}_{\mathrm{sr}}$ rating of 0.6.

In the simulation result diagrams, SII gain and STI gain are often indicated with arrows, e. g., in Figure 3.2.

2.4 Simulation Environment for Near-End Listening Enhancement

For this thesis, simulations are performed to obtain objective rating of the NELE algorithms. This section describes the utilized system models, input signals, configurations, and parameter settings. Individual exceptions are stated at the simulation results.

- The digital system models as described in Figure 2.9 are used.

- Each speech file of the TIMIT database (Garofolo et al. 1990) is taken as clean far-end speech input $s^{\mathrm{in}}(k)$, in total 6300 files and about 5.4 hours. Prior to processing, each speech file is scaled to match an overall active speech level (ITU-T P.56 1993) corresponding to a sound pressure level of $62.35\,\mathrm{dB}_{\mathrm{SPL}}$ as specified in (ANSI S3.5 1997) for normal voice effort.

- Noises from the NOISEX-92 database (SPIB 1995; Varga & Steeneken 1993), especially speech babble (`babble`), white noise (`white`), and car interior noise (`volvo`), are used as disturbing near-end noise field. The near-end noise $n(k)$ at the near-end listener’s ear is obtained by filtering these noises with the average “acoustical leakage” $\overline{H}_{\mathrm{leak}}(\Omega)$ measured for a *Siemens M65* mobile phone (see Figure 2.5b).

  The desired input SNRs ranging from $-40\,\mathrm{dB}$ to $40\,\mathrm{dB}$ in $2\,\mathrm{dB}$ steps are achieved by adjusting the overall unweighted power of the (original) noise file in relation to a sound pressure level of $62.35\,\mathrm{dB}_{\mathrm{SPL}}$ and scaling the filter coefficients of $\overline{H}_{\mathrm{leak}}(\Omega)$ to an energy of one.

  Figure 2.13 compares the spectrum level averaged over all speech files of the TIMIT database and the average disturbance spectrum levels of the considered noises from NOISEX-92 database.

- In order to avoid evaluating initial transient effects, all components, especially the speech and noise subband power estimators, should have reached their “steady state” before examination. Since the speech files of the TIMIT database are on average only 3 seconds long, each speech file is replicated and the two copies are concatenated. After processing, only the second of the concatenated copies is used for assessment.

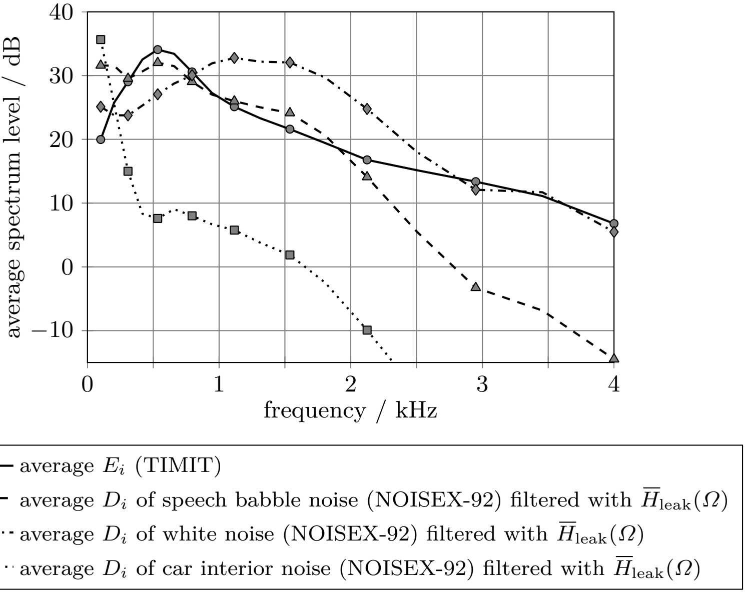


Figure 2.13: Comparison of average spectrum levels of TIMIT database and some noises of NOISEX-92 database filtered with $\overline{H}_{\text{leak}}(\Omega)$.

- As required by most objective measures, the algorithmic delay of the processed speech signal is compensated for before evaluation. The utilized cross-correlation method produced a consistent delay in all simulations.

- Each processed file is evaluated in terms of the average over all files of

  – the Speech Intelligibility Index (SII) using the so-called critical band procedure (ANSI S3.5 1997) (cf. Section 2.3.1) and
  – the speech-based revised Speech Transmission Index (STI_{sr}) as described in Section 2.3.2.

- As the objective measures SII and STI_{sr} do not cover the various effects of dialog communication, the simulations only consider single-talk (of the far-end speaker), although most proposed NELE algorithms are capable of dealing with double-talk.

- Processing is performed at sampling rate $f_{\text{s}} = 8\,\text{kHz}$.

- All evaluated NELE algorithms that utilize the framework sketched in Section 2.2 use the following parameters:

  – DFT size $M = 34$ for $f_{\text{s}} = 8\,\text{kHz}$ and $M = 42$ for $f_{\text{s}} = 16\,\text{kHz}$,
  – prototype filter length $L = M$,

- allpass coefficient $a = 0.4$ for $f_\mathrm{s} = 8\,\mathrm{kHz}$ and $a = 0.58$ for $f_\mathrm{s} = 16\,\mathrm{kHz}$,
- update interval $R \mathrel{\widehat{=}} 10\,\mathrm{ms}$,
- memory of speech subband power estimator $\tau_s = 2\,\mathrm{s}$,
- MMSE based noise PSD tracking algorithm as noise subband power estimator, and
- finite impulse response (FIR) phase equalizer of degree 50 for $f_\mathrm{s} = 8\,\mathrm{kHz}$ and of degree 96 for $f_\mathrm{s} = 16\,\mathrm{kHz}$, cf. (Löllmann & Vary 2007).

The choice of these parameters is discussed in Appendix A.

2.5 Near-End Listening Enhancement in Literature

A number of speech modification algorithms have been proposed to enhance the intelligibility of a speech signal perceived in a noisy environment. While most contributions deal with the modification of the wave form of a (natural or synthetic) speech signal – which is also the focus of this thesis – some other publications adapt the speech production stages of text-to-speech (TTS) systems (e. g., Langner & Black 2005; Raitio et al. 2011; Valentini-Botinhao et al. 2012).

To date, most of the known algorithms are noise-independent, i. e., they do not take into account different SNRs or other noise characteristics, like the spectral distribution. Section 2.5.1 provides an overview of these conventional noise-independent speech modification algorithms, which include, e. g., boosting of weak speech events, formant enhancement, and more advanced manipulations of the temporal structure.

Only recently, techniques have been studied which actually utilize prior knowledge or estimates of the background noise context and adapt their processing and weighting depending on it. These approaches, which are presented in Section 2.5.2, include formant enhancement, modification of the local SNR, optimization w. r. t. an objective criterion, and recovery of the partial loudness.

2.5.1 Noise-Independent Methods

Boosting of Consonant-Vowel-Ratio

Around the mid of the last century, first methods have been developed to enhance the intelligibility of speech perceived in noise. They were intended to emphasize those perceptual speech features which contribute most to intelligibility. Thus, weak speech events (generally consonants) are enhanced relative to the speech events with greater amplitude (generally vowels). Most of these methods use some kind of amplitude equalization, sometimes in combination with highpass filtering to emphasize the second formant frequencies relative to the first formant.

Early methods involved *peak clipping* to equalize amplitudes, which, however, introduces harmonic and intermodulation distortions due to its non-linear nature. Later, compression with *fast limiting* was studied and found to be beneficial over peak clipping (Kretsinger & Young 1960).

In (Thomas & Niederjohn 1968, 1970), highpass filtering followed by *infinite amplitude clipping* is proposed, which maps all positive values to the maximum positive amplitude and all negative values to the maximum negative amplitude and thus results in a “binary” time-domain signal. This method is shown to increase speech intelligibility in bandpass filtered white noise by up to 50 percentage points at an SNR of 0 dB. The same setup was used in (Thomas & Ohley 1972), to assess the intelligibility of highpass filtered speech in white noise without any clipping. As a result of these works, it has been shown that clipping enhances the intelligibility of highpass filtered speech in white noise for SNRs above -2 dB and reduces it below. Apparently, the severe distortions introduced by clipping cancel the positive effect of increasing the power of consonants on speech intelligibility. Therefore, a *rapid amplitude compression* is proposed in (Niederjohn & Grotelueschen 1976, 1978). This method gains a similar intelligibility score at 0 dB SNR as in (Thomas & Niederjohn 1970), but has a much better performance at lower SNRs.

More recently, an energy redistribution from voiced regions to unvoiced regions was presented in (Harris & Skowronski 2002; Skowronski & Harris 2006). Unvoiced segments of the speech signal are identified using a simple spectral flatness measure and boosted by 7.4 dB with a smooth interpolation of the gain factor at the boundaries of the segments. Finally, the energy of the output speech signal is normalized to the energy of the input signal. In a two-choice, forced-decision experiment in additive white Gaussian noise, their algorithm yielded a 3 percentage points higher mean score compared to the unmodified speech signal.

Yoo et al. (2004, 2007) proposed a decomposition into quasi-steady-state and transient components. In the former group, which represents “the steady portions of vowels and hubs of consonants”, the formants are enhanced, whereas the latter group includes the “transitions between vowels and consonants and within vowels” and is amplified by a factor of 12. The resulting speech yielded on average a 10.5 percentage points better word recognition rate over the original speech in speech-weighted noise at -10 dB SNR.

A more complex scheme is proposed by Tantibundhit et al. (2007) to decompose speech into tonal, transient, and residual components using a hidden Markov chain based on a modified discrete cosine transform and a wavelet-based hidden Markov tree. As before, the transient components are amplified by a factor of 12 and recombined with the original speech, followed by an energy normalization.

Coming from the same research group as Yoo, Rasetshwane et al. (2009) applied a wavelet packet-based method to extract an estimate of the transient speech components, which are then amplified and mixed to the original speech. Modified rhyme tests show that the proposed method is essentially as intelligible as (Yoo et al. 2007) and at least as good as (Tantibundhit et al. 2007).

Chanda and S. Park (2007) proposed a low-complexity system for intelligibility improvement. The speech signal is filtered with a tunable highpass shelving filter, which boosts the phonetic power of most consonants. The cut-off frequency of the filter is dynamically adjusted to make the input level approximately equal to the output level.

Formant Enhancement

Highpass filtering to emphasize the second formant relative to the first formant also has a long tradition starting with (Thomas 1968) and (Thomas & Ohley 1972).

McLoughlin and Chance (1997) proposed a somewhat different formant enhancement method using line spectral pairs. Each formant is shifted upwards in frequency to improve the "formant-to-noise" ratio. In addition, the formant bandwidth is widened to flatten the spectral tilt. An informal listening test with noise-independent settings indicates that intelligibility is improved by the formant shift alone by 10 percentage points and by formant bandwidth adjustment alone by 14 percentage points.

More recently, Hall and Flanagan (2010) compared differentiation, i. e., a sample-wise first-order backward difference, with formant equalization. Both filter coefficient sets are constant and boost high frequencies. In a diagnostic rhyme test, the probability of a correct response is increased with both methods by approximately the same amount of 16 percentage points. Formant equalization was, however, preferred by the listeners over differentiation.

In (Jokinen et al. 2012), two further post-filters which attenuate the first formant and enhance the second are compared with the formant equalization of (Hall & Flanagan 2010). The first post-filter adaptively tracks the formant locations, while the second one uses fixed locations. A speech recognition threshold (SRT) test indicates that all three filters improve intelligibility over unprocessed speech by about 5.6 dB.

Optimization with Respect to Objective Criterion

Rankovic (1991) used the *Articulation Index* (AI) (ANSI S3.5 1969) to fit the hearing aids of subjects with sensorineural hearing loss. The method called *AIMax* tries to "'position' the 30 dB dynamic range of short-term speech levels entirely above the pure-tone thresholds without allowing speech to surpass discomfort levels" based on the long-term average speech spectrum.

Enhancement of Pitch and Temporal Envelope

H. Park et al. (2010) presented a noise independent method which strengthens the pitch structure and the temporal modulation in seven subbands. The primary aim here was improving the perceptual quality of a speech signal.

Manipulation of Duration and Prosody

Lombard speech commonly refers to a naturally modified speaking style spoken by humans under noise exposure (Lombard 1911; Summers et al. 1988).

In (Huang et al. 2010), the authors try to mimic the Lombard effect by manipulating, e. g., phoneme duration, fundamental frequency, formant frequencies, and spectral envelope using the speech manipulation system STRAIGHT.

Spectral Shaping and Dynamic Range Compression

Very recently, Zorilă et al. (2012) described a system with spectral shaping and dynamic range compression (DRC). The spectral shaping consists of three parts, an adaptive formant enhancement, an adaptive pre-emphasis filter, and a fixed highpass filter to combat the attenuation of high frequencies during signal reproduction.

2.5.2 Noise-Dependent Methods

Formant Enhancement

A method to enhance the first three formants is proposed in (Brouckxon et al. 2008). For each formant, the signal-to-masking ratio between the current speech sound pressure level (SPL) and the hearing threshold based on the instantaneous background noise is calculated. Amplification factors are derived to ensure a certain predefined goal signal-to-masking ratio, which are afterwards smoothed over time. A small subjective evaluation indicates an about 4 dB lower SRT for the processed utterances.

Modification of Local Signal-to-Noise Ratio

In (Sauert & Vary 2006a,b), the SNR recovery algorithm (SNRrecov (A2)), cf. Section 3.2.3, is presented, which amplifies the speech signal in a time and frequency dependent way to reestablish a certain local target SNR. This concept is refined with a non-uniform filterbank in (Sauert et al. 2008).

In (Sauert et al. 2006), two strategies for speech power reallocation are compared which satisfy a strict output audio power constraint: the reallocation to *noisier* frequencies and the reallocation to *less noisy* frequencies, the latter resulting in the method of maximal power transfer (MaxTransfer (A6)), cf. Section 4.2.1.

So far, the far-end signal was considered to be clean, i. e., recorded in a noise-free environment or sufficiently cleaned before transmission with a state-of-the art noise reduction algorithm. In contrast, Choi et al. (2009) assume a *noisy* far-end speech signal and extend SNRrecov (A2) (Sauert & Vary 2006a,b) to this case. A speech-absence probability for the far-end signal is included into the weighting rule (3.27), which aims at amplifying only the far-end speech but not the far-end noise.

Different strategies to reallocate energy over time and/or frequency are presented in (Tang & Cooke 2010, 2011) and evaluated with listening tests and objective measures. As a result, the *SelectBoost* approach had the most notable intelligibility improvement of 10 to 38 percentage points depending on noise characteristic and SNR. It boosts speech energy in those time/frequency regions above 1.8 kHz with a local SNR of less than 5 dB.

Spectral Shaping and Dynamic Range Compression

Erro et al. (2012) proposed a two-step parametric approach based on full-band harmonic modelling. At first, the spectral slope is increased to mimic the effect of a higher vocal effort. Then, in a DRC step, the energy of the signal is redistributed over time to amplify meaningful low-energy parts of the signal.

Optimization with Respect to Objective Criterion

In (Sauert & Vary 2009), the bounded SII-based optimization (OptSIIbound (A1)), cf. Section 3.2.1, is proposed which improves speech intelligibility by optimizing the spectral weights w. r. t. the SII. This optimization concept is extended to a total audio power constraint in (Sauert & Vary 2010a,b), resulting in the recursive closed-form power-constrained SII-based optimization (OptSIIrecur (A4)), cf. Section 4.1.2, which is later refined in (Sauert & Vary 2011, 2012b).

The approach presented in (Tang & Cooke 2012) applies stationary spectral weights which are optimized offline for different noise types at a range of SNRs using a genetic algorithm and the glimpse proportion (Cooke 2006) as optimization criterion. Online, i. e., during application, only noise descriptors, like noise type and SNR, must be estimated to choose the correct pre-optimized parameter settings.

Recently, Taal et al. (2012) presented an optimization algorithm based on a spectro-temporal perceptual distortion measure. As the underlying auditory model considers the temporal envelope, the proposed method is sensitive to transient regions and amplifies them compared to vowels.

Recovery of Partial Loudness

In the presence of noise, the signal of interest is partially masked and thus perceived with a reduced loudness. This effect can be described, e. g., with the loudness perception model of Moore and Glasberg (Moore et al. 1997). The model describes the loudness per ERB[6] of a signal in silence, named specific loudness, as well as the partial specific loudness of a signal perceived in noise.

J. W. Shin et al. (2009, 2007) propose an algorithm which aims at an unchanged listening experience, i. e., an unchanged loudness, despite the noise. The input speech signal is amplified such that the partial specific loudness of the amplified speech in noise becomes the same as the specific loudness of the noise-free signal. This method requires to increase the power of the speech signal and improves intelligibility just as a side effect.

In mobile communication, one ear usually is open and perceives just the background noise signal, whereas the other ear is covered by the mobile phone and thus hears a mixture of the far-end speech signal and a filtered (attenuated) background noise signal (see Section 2.1). This situation is considered in (H. S. Shin et al. 2010), where the above idea is applied to the binaural loudness model of Moore and Glasberg (2007). This model adds a non-linear correction term, which compensates for the different (partial) specific loudnesses on the two ears. A preference test shows that the binaural based system is preferred over the monaural based system.

[6] The equivalent rectangular bandwidth (ERB) scale is closely related to the critical bands.

Chapter 3

Near-End Listening Enhancement *without* Total Power Constraint

In this chapter, near-end listening enhancement (NELE) algorithms are discussed which improve speech intelligibility without constraining the total audio power. The only applied limitation is the prevention of listener's hearing damage, which is described in Section 2.2.5.

The Speech Intelligibility Index (SII) is chosen as optimization criterion for these algorithms due to its proven ability to predict speech intelligibility and its calculation rules, which are suitable for algorithm design.

While the standard SII, as outlined in Section 2.3.1, uses long-term power averages, e. g., over a whole utterance, the time-varying subband weights of the algorithms developed in this thesis are based on the short-term subband powers using the original SII calculation rules as criterion. This approach is supported by Rhebergen and Versfeld, who propose “an extension to the SII model [. . .] with the aim to predict the speech intelligibility in both stationary and fluctuating noise. The basic principle [. . .] is that both speech and noise signal are partitioned into small time frames. Within each time frame the conventional SII is determined, yielding the speech information available to the listener at that time frame. Next, the SII values of these time frames are averaged [. . .]” (Rhebergen & Versfeld 2005).

The basic idea of this class of algorithms is to first determine an optimum speech spectrum level $E_i^{\mathrm{opt}}(\kappa)$ for each subband i which maximizes the SII $S(\underline{E}, \underline{D}(\kappa))$, cf. (2.44), under consideration of the current disturbance spectrum level $D_i(\kappa)$:

$$\underline{E}^{\mathrm{opt}}(\kappa) = \arg\max_{\underline{E}} \left\{ S\big(\underline{E}, \underline{D}(\kappa)\big) \right\} \tag{3.1}$$

optionally subject to some major or minor constraints, which are discussed later. Due to the transfer characteristics of the mobile phone's microphone and loudspeaker as well as the utilized speech codec, not all subbands are available for enhancement in a mobile phone application as described in Section 2.2.2. Therefore, the optimization is only performed for the $i_{\mathrm{l}} - i_{\mathrm{f}} + 1$ subbands from the first contributing subband i_{f} to the last contributing subband i_{l}. Accordingly, the vector notation $\underline{E}$ denotes the vector $(E_{i_{\mathrm{f}}}, E_{i_{\mathrm{f}}+1}, \ldots, E_{i_{\mathrm{l}}})$ of the spectrum levels in all contributing subbands[1].

[1] Using the default parameters given in Section 2.4, there are 17 contributing subbands at sampling rate $f_{\mathrm{s}} = 8\,\mathrm{kHz}$ and 21 contributing subbands at $f_{\mathrm{s}} = 16\,\mathrm{kHz}$.

As second step, the subband weights are calculated which are necessary to achieve the optimum speech spectrum level $E_i^{\mathrm{opt}}(\kappa)$ with the input speech signal $s^{\mathrm{in}}(k)$ at the ear of the listener.

For this application, the speech spectrum level $E_i^{\mathrm{in}}(\kappa)$ of the input speech signal and the disturbance spectrum level $D_i(\kappa)$ are calculated as described in Section 2.3.1 based on the short-term subband power estimates $\hat{P}_{s,i}^{\mathrm{in}}(\kappa)$ and $\hat{P}_{n,i}(\kappa)$. With the (to be determined) subband weights $W_i(\kappa)$, the short-term subband power estimate $\hat{P}_{s,i}^{\mathrm{out}}(\kappa)$ of the (enhanced) subband output speech signal $s_i^{\mathrm{out}}(\kappa) = W_i(\kappa) \cdot s_i^{\mathrm{in}}(\kappa)$ can be expressed as

$$\hat{P}_{s,i}^{\mathrm{out}}(\kappa) = W_i^2(\kappa) \cdot \hat{P}_{s,i}^{\mathrm{in}}(\kappa)\,, \quad i_{\mathrm{f}} \leq i \leq i_{\mathrm{l}}\,, \tag{3.2}$$

which leads to the speech spectrum level of the output speech signal

$$E_i^{\mathrm{out}}(\kappa) = 10\log\left\{\frac{\hat{P}_{s,i}^{\mathrm{out}}(\kappa)}{f_{\Delta,i}}\right\} = 20\log\big\{W_i(\kappa)\big\} + E_i^{\mathrm{in}}(\kappa) \tag{3.3}$$

and results in the subband weights

$$W_i(\kappa) = 10^{\left[E_i^{\mathrm{out}}(\kappa) - E_i^{\mathrm{in}}(\kappa)\right]/20}\,, \quad i_{\mathrm{f}} \leq i \leq i_{\mathrm{l}}\,. \tag{3.4}$$

3.1 Analysis of SII Calculation Rules

In this section the calculation rules of the SII, which have been presented in Section 2.3.1, are briefly analyzed w. r. t. their dependency on the speech spectrum level E_i. This gives the foundation for the following SII-based optimizations.

The band audibility function $A_i(E_i, D_i)$

$$A_i(E_i, D_i) = L_i(E_i) \cdot K_i(E_i, D_i) \qquad \text{[2.42, p. 29]}$$

as a function of E_i is determined by two factors with diametrically opposed impact:

- The auxiliary variable

$$K_i(E_i, D_i) = \begin{cases} 0 & \text{if } E_i \leq D_i - 15\,\mathrm{dB} \\ \frac{E_i - D_i + 15\,\mathrm{dB}}{30\,\mathrm{dB}} & \text{if } D_i - 15\,\mathrm{dB} < E_i \leq D_i + 15\,\mathrm{dB} \\ 1 & \text{if } D_i + 15\,\mathrm{dB} < E_i\,, \end{cases} \qquad \text{[2.43, p. 29]}$$

  which accounts for the masking of the speech signal by the noise, *increases* monotonically with increasing speech spectrum level E_i.

- The level distortion factor

$$L_i(E_i) = \begin{cases} 1 & \text{if } E_i \leq U_i + 10\,\mathrm{dB} \\ 1 - \frac{E_i - U_i - 10\,\mathrm{dB}}{160\,\mathrm{dB}} & \text{if } U_i + 10\,\mathrm{dB} < E_i \leq U_i + 170\,\mathrm{dB} \\ 0 & \text{if } U_i + 170\,\mathrm{dB} < E_i\,, \end{cases} \qquad \text{[2.41, p. 28]}$$

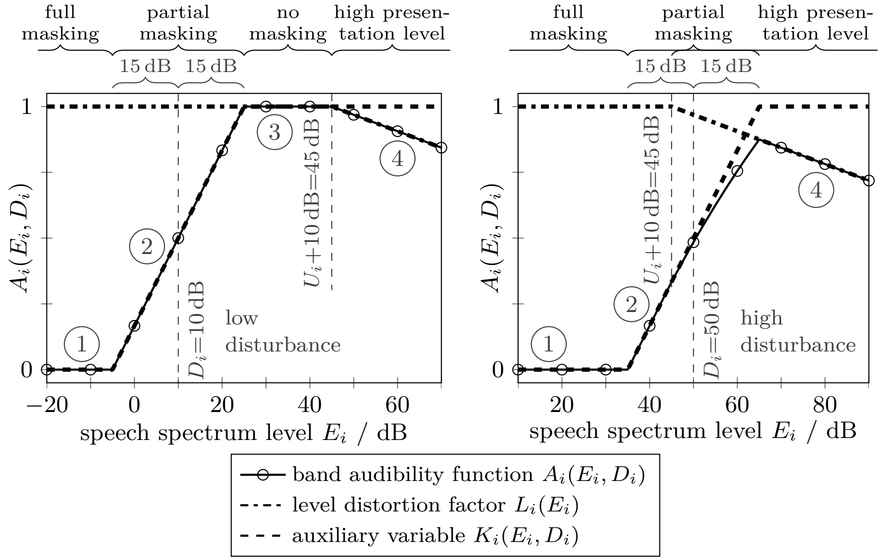


Figure 3.1: Exemplary plots of contributions to the band audibility function for low as well as high disturbance case.
The four segments are marked with circles.

which considers the distortion due to a high presentation level, *decreases* monotonically with increasing speech spectrum level E_i. The standard speech spectrum level at normal voice effort U_i is fixed and can be found in (ANSI S3.5 1997, Table 1). It has its maximum value of 34.75 dB in the second critical band with $f_{c,2} = 250\,\text{Hz}$.

As a consequence, two cases of practical interest exist for $A_i(E_i, D_i)$ depending on the disturbance spectrum level D_i, which are exemplarily depicted in the left resp. right diagram of Figure 3.1:

- *Low disturbance case:* The Segment ② with increasing $K_i(E_i, D_i)$ ends before the start of the Segment ④ with decreasing $L_i(E_i)$.

- *High disturbance case:* The Segments ② and ④ with increasing $K_i(E_i, D_i)$ and decreasing $L_i(E_i)$ overlap.

- The third case where the segment with increasing $K_i(E_i, D_i)$ starts after $L_i(E_i)$ has vanished is not of practical interest since it occurs only for $D_i > U_i + 185\,\text{dB}$ and results in a band audibility function identical to zero.

3.1.1 Low Disturbance Case: $D_i + 15\,\mathrm{dB} \leq U_i + 10\,\mathrm{dB}$

As sketched in the left diagram of Figure 3.1, the band audibility function $A_i(E_i, D_i)$ exhibits in this case four segments of practical interest:

① For very low E_i, the speech is assumed to be fully masked by the noise and the i-th frequency band has thus no contribution to intelligibility.

② For $D_i - 15\,\mathrm{dB} < E_i \leq D_i + 15\,\mathrm{dB}$, the speech signal is only partially masked and the band audibility function increases.

③ For $D_i + 15\,\mathrm{dB} < E_i \leq U_i + 10\,\mathrm{dB}$ on, masking is assumed to be irrelevant and the frequency band has full contribution to intelligibility.

④ For speech spectrum levels $E_i \geq U_i + 10\,\mathrm{dB}$, intelligibility is reduced again due to the high presentation level.

The resulting band audibility function $A_i(E_i, D_i)$ is continuous and piecewise linear:

$$A_i(E_i, D_i) = \begin{cases} 0 & \text{if } E_i \leq D_i - 15\,\mathrm{dB} \\ \frac{E_i - D_i + 15\,\mathrm{dB}}{30\,\mathrm{dB}} & \text{if } D_i - 15\,\mathrm{dB} < E_i \leq D_i + 15\,\mathrm{dB} \\ 1 & \text{if } D_i + 15\,\mathrm{dB} < E_i \leq U_i + 10\,\mathrm{dB} \\ 1 - \frac{E_i - U_i - 10\,\mathrm{dB}}{160\,\mathrm{dB}} & \text{if } U_i + 10\,\mathrm{dB} < E_i \leq U_i + 170\,\mathrm{dB} \\ 0 & \text{if } U_i + 170\,\mathrm{dB} < E_i\,. \end{cases} \tag{3.5}$$

It can be easily seen, that the maximum value

$$\max_{E_i} \big\{ A_i(E_i, D_i) \big\} = 1 \tag{3.6}$$

is reached throughout the third segment where masking is irrelevant:

$$D_i + 15\,\mathrm{dB} \leq E_i \leq U_i + 10\,\mathrm{dB}\,. \tag{3.7}$$

3.1.2 High Disturbance Case: $D_i + 15\,\mathrm{dB} > U_i + 10\,\mathrm{dB}$

As shown in the last section, the boundaries between the Segments ①, ②, and ③ “move” with a changing disturbance spectrum level $D_i(\kappa)$, whereas the boundary to Segment ④ is independent of $D_i(\kappa)$. Thus, for a high disturbance spectrum level, the Segments ② and ④ overlap and Segment ③ vanishes as depicted in the right diagram of Figure 3.1. The resulting band audibility function $A_i(E_i, D_i)$ is a continuous, downward opened parabola in the overlapping segment and piecewise linear elsewhere:

$$A_i(E_i, D_i) = \begin{cases} 0 & \text{if } E_i \le D_i - 15\,\text{dB} \\ \frac{E_i - D_i + 15\,\text{dB}}{30\,\text{dB}} & \text{if } D_i - 15\,\text{dB} < E_i \le \xi_{\text{b},i} \\ \left(\frac{E_i - D_i + 15\,\text{dB}}{30\,\text{dB}}\right) \cdot \left(1 - \frac{E_i - U_i - 10\,\text{dB}}{160\,\text{dB}}\right) & \text{if } \xi_{\text{b},i} < E_i \le \xi_{\text{e},i} \\ 1 - \frac{E_i - U_i - 10\,\text{dB}}{160\,\text{dB}} & \text{if } \xi_{\text{e},i} < E_i \le U_i + 170\,\text{dB} \\ 0 & \text{if } U_i + 170\,\text{dB} < E_i \end{cases} \tag{3.8}$$

with the beginning of the quadratic segment $\xi_{\text{b},i} = \max\{U_i + 10\,\text{dB},\, D_i - 15\,\text{dB}\}$ and its end $\xi_{\text{e},i} = \min\{D_i + 15\,\text{dB},\, U_i + 170\,\text{dB}\}$.

In the segment $\xi_{\text{b},i} < E_i \le \xi_{\text{e},i}$, the gradient

$$\frac{\mathrm{d}A_i(E_i, D_i)}{\mathrm{d}E_i} = \frac{1}{30\,\text{dB}} \cdot \left(1 - \frac{E_i - U_i - 10\,\text{dB}}{160\,\text{dB}}\right) - \frac{E_i - D_i + 15\,\text{dB}}{30\,\text{dB}} \cdot \frac{1}{160\,\text{dB}} \tag{3.9}$$

is always positive if $D_i < U_i + 125\,\text{dB}$. Thus

$$\frac{\mathrm{d}A_i(E_i, D_i)}{\mathrm{d}E_i} \begin{cases} = 0 & \text{if } E_i \le D_i - 15\,\text{dB} \\ > 0 & \text{if } D_i - 15\,\text{dB} < E_i \le \xi_{\text{b},i} \\ > 0 & \text{if } \xi_{\text{b},i} < E_i \le \xi_{\text{e},i} \\ < 0 & \text{if } \xi_{\text{e},i} < E_i \le U_i + 170\,\text{dB} \\ = 0 & \text{if } U_i + 170\,\text{dB} < E_i\,. \end{cases} \tag{3.10}$$

It follows, that the maximum value

$$\max_{E_i}\bigl\{A_i(E_i, D_i)\bigr\} = 1 - \frac{D_i - U_i + 5\,\text{dB}}{160\,\text{dB}} \tag{3.11}$$

is reached for

$$E_i = \xi_{\text{e},i} = D_i + 15\,\text{dB}\,. \tag{3.12}$$

3.1.3 Summary

The maximum value of the band audibility function $A_i(E_i, D_i)$ is for all disturbance spectrum levels D_i given by

$$\max_{E_i}\bigl\{A_i(E_i, D_i)\bigr\} = \min\left\{1,\, 1 - \frac{D_i - U_i + 5\,\text{dB}}{160\,\text{dB}}\right\}. \tag{3.13}$$

For the practically relevant case $D_i < U_i + 125\,\text{dB}$, this maximum value is reached for

$$D_i + 15\,\text{dB} \le E_i \le \min\bigl\{D_i + 15\,\text{dB},\, U_i + 10\,\text{dB}\bigr\}\,. \tag{3.14}$$

3.1.4 Theoretical Performance Bound (TheoPerfBound)

Following the above summary, the maximum reachable SII $\tilde{S}(\underline{D})$ given the disturbance spectrum level D_i is

$$\tilde{S}(\underline{D}) = \max_{\underline{E}}\Big\{S\big(\underline{E}, \underline{D}(\kappa)\big)\Big\} = \sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} I_i \cdot \min\Big\{1,\ 1 - \frac{D_i - U_i + 5\,\mathrm{dB}}{160\,\mathrm{dB}}\Big\}, \quad (3.15)$$

which gives a theoretical performance bound (TheoPerfBound) for all algorithms.

3.2 Optimization with Respect to SII

In this section, the SII should be maximized without limitation of the *total* audio power. With the prevention of listener's hearing damage, introduced in Section 2.2.5, this leads to the bounded maximization problem

$$\underline{E}^\mathrm{opt}(\kappa) = \arg\max_{\underline{E}}\Big\{S\big(\underline{E}, \underline{D}(\kappa)\big)\Big\} \qquad [3.1, \text{p. } 41]$$

with the SII $S(\underline{E}, \underline{D}(\kappa))$ according to (2.44), subject to the bounds

$$E_i \overset{!}{\leq} E_i^\mathrm{max} \quad \forall\, i_\mathrm{f} \leq i \leq i_\mathrm{l}\,. \quad (3.16)$$

The maximum allowed speech spectrum level

$$E_i^\mathrm{max} = E_i^\mathrm{in} + 20\log\big\{W_i^\mathrm{max}(\kappa)\big\} = 10\log\bigg\{\frac{P_s^\mathrm{max}}{f_{\Delta,i}}\bigg\}, \quad (3.17)$$

is determined with the maximum subband power P_s^max to prevent hearing damage.

Since the speech spectrum levels for the different frequency bands do not depend on each other (as opposed to the case considered in Chapter 4) and since the band audibility function of one frequency band only depends on the speech spectrum level of exactly that frequency band, the $(i_\mathrm{l} - i_\mathrm{f} + 1)$-dimensional bounded optimization problem (3.1) can be expressed by $i_\mathrm{l} - i_\mathrm{f} + 1$ different *one*-dimensional bounded optimization problems. In the range $E_i \leq \min\{D_i(\kappa) + 15\,\mathrm{dB},\ U_i + 10\,\mathrm{dB}\}$, where the band audibility function $A_i(E_i, D_i)$ increases monotonically (see Figure 3.1), this one-dimensional bounded optimization problem can furthermore be seen as an unconstrained optimization problem with subsequent bounding of the solution.

3.2.1 Bounded SII-Based Optimization (OptSIIbound (A1))

According to (3.14), the speech spectrum level

$$E_i(\kappa) = \max\Big\{D_i(\kappa) + 15\,\mathrm{dB}, \min\{E_i^\mathrm{in}(\kappa), U_i + 10\,\mathrm{dB}\}\Big\} \quad (3.18)$$

leads to the maximum SII. The second term of the maximum function prevents an attenuation of the speech signal in quiet environments.

Taking the minimum in the second term of (3.18) considers loud input speech signals, in which case it can be necessary to reduce speech signal power in some subbands in order to maximize the SII. This, however, tends to excessively attenuate high frequency components of fricatives and plosives at a sampling rate of 16 kHz. Therefore, the minimum in (3.18) is omitted and replaced by $E_i^{\text{in}}(\kappa)$ in the following.

The bound (3.16) is finally satisfied by the optimum speech spectrum level

$$E_i^{\text{opt}}(\kappa) = \min\Big\{\max\{D_i(\kappa) + 15\,\text{dB}, E_i^{\text{in}}(\kappa)\}, E_i^{\max}\Big\}\,, \tag{3.19}$$

which leads to the subband weights

$$W_i(\kappa) = 10^{\left[\min\left\{\max\left\{D_i(\kappa)+15\,\text{dB}, E_i^{\text{in}}(\kappa)\right\}, E_i^{\max}\right\} - E_i^{\text{in}}(\kappa)\right]/20} \tag{3.20}$$

$$= \min\left\{\max\left\{10^{\left[D_i(\kappa)+15\,\text{dB}-E_i^{\text{in}}(\kappa)\right]/20},\, 1\right\},\, W_i^{\max}(\kappa)\right\}. \tag{3.21}$$

This method is called *bounded SII-based optimization* (OptSIIbound (A1))[2] and was first published in (Sauert & Vary 2009).

3.2.2 Analysis

In this section, the OptSIIbound (A1) scheme presented above is studied concerning the characteristics of the resulting subband weights for different SNRs.

At very low SNRs, i. e., for $E_i^{\max} < \max\{D_i(\kappa) + 15\,\text{dB}, E_i^{\text{in}}(\kappa)\}$, the solution is determined by the maximum allowed subband power which prevents the listener's hearing damage:

$$E_i^{\text{opt}}(\kappa) = E_i^{\max}\,. \tag{3.22}$$

For low to mid SNRs, i. e., for $E_i^{\text{in}}(\kappa) \leq D_i(\kappa) + 15\,\text{dB} \leq E_i^{\max}(\kappa)$, the optimum solution lies at the upper bound of Segment ② (cf. Figure 3.1). In this case, (3.19) simplifies to

$$E_i^{\text{opt}}(\kappa) = D_i(\kappa) + 15\,\text{dB}\,. \tag{3.23}$$

In this case, the overall spectral shape of the output speech roughly approaches that of the noise. However, the temporal and spectral fine-structure of the speech signal is still preserved, since only few subband weights are applied in the comparably wide critical subbands and the subband weights change much more slowly than the phonemes of the speech signal.

At high SNR, i. e., for $D_i(\kappa) + 15\,\text{dB} < E_i^{\text{in}}(\kappa)$, the optimum solution lies in Segment ③ with

$$E_i^{\text{opt}}(\kappa) = E_i^{\text{in}}(\kappa) \tag{3.24}$$

and thus 0 dB subband weights, i. e., no modification is applied in quiet environments.

[2] For the sake of clarity, all presented algorithms are numbered and the number is stated together with the acronym in the following.

3.2.3 SNR Recovery Algorithm (SNRrecov (A2))

A reduction of complexity can be achieved compared to OptSIIbound (A1), if the masking of speech components by the noise is neglected in the calculation of the disturbance spectrum level $D_i(\kappa)$, i. e.,

$$\underline{D}(\kappa) = \underline{N}(\kappa)\,. \tag{3.25}$$

The time-varying weight factors of (3.21) reduce to

$$W_i'(\kappa) = \min\left\{ \max\left\{ 10^{\left[N_i(\kappa)+15\,\mathrm{dB}-E_i^{\mathrm{in}}(\kappa)\right]/20},\, 1\right\},\, W_i^{\max}(\kappa)\right\} \tag{3.26}$$

$$= \min\left\{ \max\left\{ 10^{15\,\mathrm{dB}/20} \cdot \sqrt{\frac{\hat{P}_{n,i}(\kappa)}{\hat{P}_{s,i}^{\mathrm{in}}(\kappa)}},\, 1\right\},\, W_i^{\max}(\kappa)\right\}. \tag{3.27}$$

This algorithm was found heuristically in (Sauert & Vary 2006a,b) using a DFT AS FB and was named *SNR recovery algorithm* (SNRrecov (A2)). In (Sauert et al. 2008), it was later adapted to the filterbank equalizer (FBE) described in Section 2.2.1.

3.3 Simulation Results

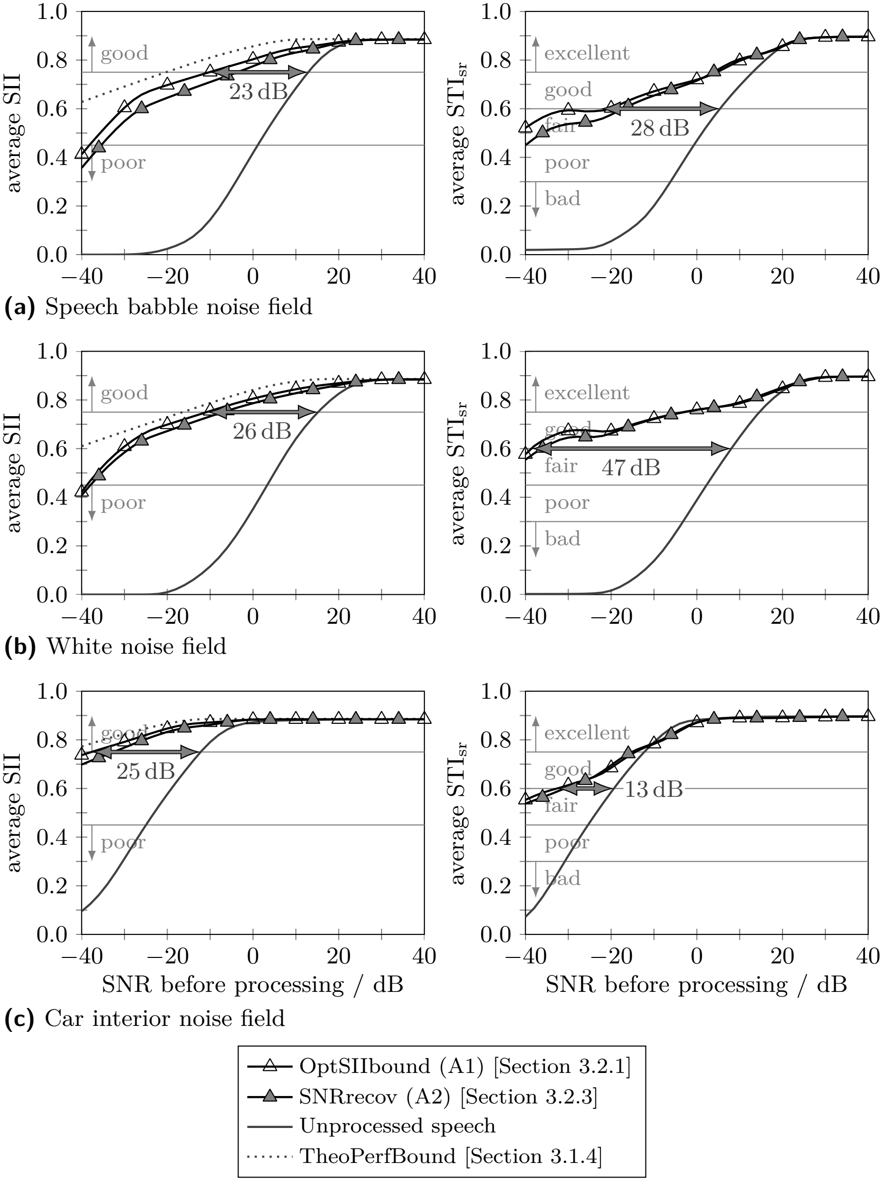
Figure 3.2 shows a comparison of the performance of the two algorithms presented in this chapter: OptSIIbound (A1) and SNRrecov (A2). Simulations are performed at 8 kHz sampling rate using the FBE framework with 17 non-uniform contributing subbands. A detailed description of the simulation parameters is given in Section 2.4.

The OptSIIbound (A1) algorithm yields an SII gain of 23 dB to 26 dB, i. e., it retains a “good” communication system at a 23 dB to 26 dB lower input SNR compared to unprocessed speech (see Section 2.3.3). Due to the shape of the $\mathrm{STI}_{\mathrm{sr}}$ curve, the STI gain is more heterogeneous between 28 dB and 47 dB for speech babble and white noise, respectively.

Since the car interior noise accumulates almost all its energy at very low frequencies (see Figure 2.13), most frequency bands remain almost undisturbed even at 0 dB average SNR. Therefore, the degradation generally starts at lower SNRs than for speech babble or white noise. Furthermore, OptSIIbound (A1) has a slightly worse $\mathrm{STI}_{\mathrm{sr}}$ than unprocessed speech at a mid SNR range of -10 dB to $+2$ dB. This is discussed in Section 4.2.3.

The SII performance of OptSIIbound (A1) lies below TheoPerfBound, which can be explained by the fact that the algorithms optimize the SII for each frame separately based on the current (smoothed) spectrum levels, whereas the average SII is calculated from the mean spectrum level over each entire sound file. At SNRs below 20 dB, the hearing damage prevention, i. e., bound (3.16), becomes active and enlarges the gap to TheoPerfBound.

SNRrecov (A2) neglects the masking effect and hence leads to an equal or smaller amplification in each frame. On average the SII performance is slightly



(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 3.2: Comparison of OptSIIbound (A1) and SNRrecov (A2). See Section 2.4 for simulation parameters. The arrows indicate the SII and STI gain of OptSIIbound (A1).

worse but still comparable. Apparently, the slope of masking has only a small influence on the performance of the algorithm.

Note, that due to the sampling rate $f_\mathrm{s} = 8\,\mathrm{kHz}$, the speech signal misses some frequency bands which are considered by the measures SII and STI_sr. Consequently, all algorithms as well as the theoretical bound can not reach a measure of one even at an infinitely high SNR. Please refer to Section 4.4 for examples with 16 kHz sampling rate.

3.3.1 Comparison with Frequency Independent Version

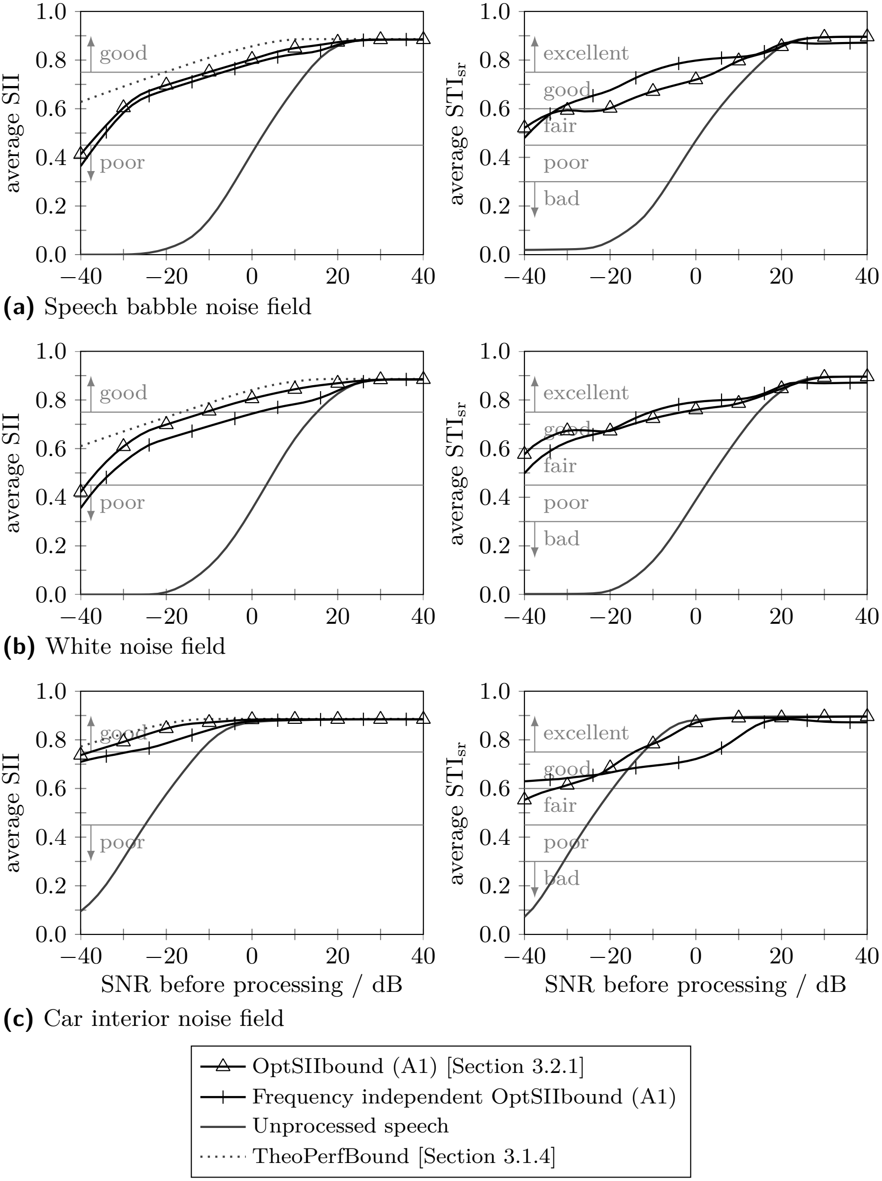
In order to evaluate the benefit of frequency dependent processing, a frequency independent version, i. e., a pure gain manipulation which uses the same power as OptSIIbound (A1) is evaluated. The far-end speech signal $s^\mathrm{in}(k)$ is partitioned in overlapping frames with 0.5 s length and 10 ms overlap, which are multiplied with a Hann window. These frames are scaled to have the same energy as the corresponding frames of the output of OptSIIbound (A1) and joined with an overlap-add technique.

The length of the frames and thus the fluctuation of the weighting factors is an important factor for the STI_sr rating. Therefore, a frame length of 0.5 s is chosen such that the loss in STI_sr at very high SNRs is acceptable.

In comparison to the frequency dependent approach OptSIIbound (A1), the average SII is consistently lower after frequency independent amplification as can be seen in Figure 3.3. This is in accordance with informal listening tests and justifies the frequency dependent approach chosen in this thesis.

In contrast, the frequency independent version shows better STI_sr rating for speech babble and white noise at medium SNRs than the frequency dependent version. This is, however, just an effect of the slow fluctuations due to the long frame length.

Since a speech signal and the speech babble noise have a similar average spectrum, the resulting weights of OptSIIbound (A1) are almost frequency independent. Consequently, the average SII of the frequency dependent and independent OptSIIbound (A1) are about the same for speech babble noise.



(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 3.3: Comparison of frequency dependent and independent versions of OptSIIbound (A1). See Section 2.4 for simulation parameters.

Chapter 4

Near-End Listening Enhancement *with* Total Power Constraint

In practical applications the overall loudspeaker signal power is constrained, i. e., the total short-term audio power of the contributing subbands of the optimized output signal must be less or equal than a maximum, possibly time-varying total short-term audio power $\mathfrak{P}^{\max}(\kappa)$:

$$\sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} \hat{P}_{s,i}^{\mathrm{out}}(\kappa) \overset{!}{\leq} \mathfrak{P}^{\max}(\kappa)\,. \tag{4.1}$$

Therefore, this chapter treats the $(i_\mathrm{l}-i_\mathrm{f}+1)$-dimensional non-linear maximization problem

$$\underline{E}^{\mathrm{opt}}(\kappa) = \arg\max_{\underline{E}}\left\{S\big(\underline{E},\underline{D}(\kappa)\big)\right\} \tag*{[3.1, p. 41]}$$

with the SII $S(\underline{E},\underline{D}(\kappa))$, subject to the new inequality constraint of the total audio power

$$\sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} f_{\Delta,i}\cdot 10^{E_i/10} \overset{!}{\leq} \mathfrak{P}^{\max}(\kappa) \tag{4.2}$$

and the bounds to prevent listener's hearing damage

$$E_i \overset{!}{\leq} E_i^{\max} \quad \forall\, i_\mathrm{f} \leq i \leq i_\mathrm{l}\,. \tag*{[3.16, p. 46]}$$

The vector notation $\underline{E}$ again denotes the vector $(E_{i_\mathrm{f}}, E_{i_\mathrm{f}+1}, \ldots, E_{i_\mathrm{l}})$ of the spectrum levels in all contributing subbands.

In contrast to the optimization problem considered by OptSIIbound (A1), the speech spectrum levels of the different subbands *do* depend on each other due to the inequality constraint. Therefore, this $(i_\mathrm{l}-i_\mathrm{f}+1)$-dimensional inequality constrained non-linear maximization problem can *not* be reduced to one-dimensional optimization problems as in Section 3.2 but must be solved in full size.

Concerning the constraint of the total audio output power, two variants are considered in this thesis:

Constraint 1: The loudspeaker signal power is constrained to the power of the original (input) signal, i. e., no additional audio power may be spent. In this case, the maximum allowed total audio power is time-variant and equal to the total short-term audio power of the contributing subbands of the input signal

$$\mathfrak{P}^{\max}(\kappa) = \sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} f_{\Delta,i} \cdot 10^{E_i^{\mathrm{in}}(\kappa)/10}. \tag{4.3}$$

This constraint is basically an extreme case for sound reproduction systems without head-room in terms of total output audio power. But it also proves useful for comparison with other NELE algorithms.

Constraint 2: One major limitation of the small loudspeakers used in mobile phones is the thermal load during continuous playback. In this realistic case, the maximum allowed total audio power $\mathfrak{P}^{\max}$ is constant and a parameter of the sound reproduction system, which could be derived during design of the device.

The *15x11x3.5 speaker* built by *NXP Semiconductors* can be considered a "typical" loudspeaker used for hands-free telephony. As indicated in its specification (NXP 2010a), the thermal limit is reached for this speaker at 500 mW (RMS), which is equivalent to a characteristic sensitivity of about

$$10\log\left\{\frac{\mathfrak{P}^{\max}}{P_0}\right\} = 90\,\mathrm{dB_{SPL}} \tag{4.4}$$

in 10 cm distance.

In Section 4.1, general solutions to the power-constrained optimization problem are developed and analyzed. These solutions are evaluated without increase of total audio power (Constraint 1) in Section 4.2, which also includes algorithmic modifications for problematic noise types. The second constraint, i. e., the increase of total audio power up to a constant maximum audio power given by the thermal limit of the loudspeaker is considered in Section 4.3. Section 4.4 discusses the influence of a higher sampling rate on the performance of the algorithms and, in Section 4.5, the proposed methods are compared with algorithms from the literature. Section 4.6 completes the chapter with results from three listening tests.

4.1 Power-Constrained Optimization with Respect to SII

It is shown in Section 3.1.3 that the (unconstrained) maximum SII is reached for

$$D_i(\kappa) + 15\,\mathrm{dB} \leq E_i \leq \min\left\{D_i(\kappa) + 15\,\mathrm{dB},\, U_i + 10\,\mathrm{dB}\right\}. \qquad \text{[cf. 3.14, p. 45]}$$

The term *admissible range* is introduced for the derivations of optimization schemes in this section as the speech spectrum level range below $D_i(\kappa) + 15\,\mathrm{dB}$ (see Figure 4.1). It has the upper border

$$E_i^{\mathrm{adm}}(\kappa) = \min\bigl\{D_i(\kappa) + 15\,\mathrm{dB},\, E_i^{\mathrm{max}}(\kappa)\bigr\}\,, \tag{4.5}$$

which also considers $E_i^{\mathrm{max}}(\kappa)$ to prevent listener's hearing damage.

This leads to two cases depending on the speech spectrum level E_i:

Case 1: $\underline{E} = \underline{E}^{\mathrm{adm}}(\kappa)$ *does not* fulfill the power constraint (4.2), i. e.,

$$\sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} f_{\Delta,i} \cdot 10^{E_i^{\mathrm{adm}}(\kappa)/10} > \mathfrak{P}^{\mathrm{max}}(\kappa)\,. \tag{4.6}$$

In this case, all power must be used to maximize the SII. The solution $\underline{E} = \underline{E}^{\mathrm{opt}}(\kappa)$ always fulfills equality in (4.2) as well as

$$E_i^{\mathrm{opt}} \leq E_i^{\mathrm{adm}}(\kappa) \quad \forall\, i_\mathrm{f} \leq i \leq i_\mathrm{l}\,. \tag{4.7}$$

Consequently, the above $(i_\mathrm{l} - i_\mathrm{f} + 1)$-dimensional *inequality* constrained maximization problem turns into the $(i_\mathrm{l} - i_\mathrm{f} + 1)$-dimensional *equality* constrained maximization problem

$$\underline{E}^{\mathrm{opt}}(\kappa) = \arg\max_{\underline{E}}\Bigl\{S\bigl(\underline{E}, \underline{D}(\kappa)\bigr)\Bigr\} \qquad [3.1,\ p.\,41]$$

with the SII $S(\underline{E}, \underline{D}(\kappa))$ according to (2.44), subject to the *equality* constraint

$$\sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} f_{\Delta,i} \cdot 10^{E_i/10} \overset{!}{=} \mathfrak{P}^{\mathrm{max}}(\kappa) \tag{4.8}$$

and the bounds for the speech spectrum level

$$E_i \overset{!}{\leq} E_i^{\mathrm{max}} \quad \forall\, i_\mathrm{f} \leq i \leq i_\mathrm{l}\,. \qquad [3.16,\ p.\,46]$$

In Section 4.1.1 an approach is presented which performs a numerical optimization of a concave approximation of the band audibility function. Computationally less complex is a recursive closed-form optimization of a linear approximation of the band audibility function which is derived in Section 4.1.2.

Case 2: $\underline{E} = \underline{E}^{\mathrm{adm}}(\kappa)$ *does* fulfill the power constraint (4.2), i. e.,

$$\sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} f_{\Delta,i} \cdot 10^{E_i^{\mathrm{adm}}(\kappa)/10} \leq \mathfrak{P}^{\mathrm{max}}(\kappa)\,. \tag{4.9}$$

Here, the maximum SII can be reached and the further audio power may be used to reduce the change of tone color. This topic is addressed in Section 4.1.3.

4.1.1 Numerical Optimization (OptSIInum (A3))

The above $(i_\mathrm{l} - i_\mathrm{f} + 1)$-dimensional *equality* constrained maximization problem can be transformed into a $(i_\mathrm{l} - i_\mathrm{f})$-dimensional *bound* constrained maximization problem by expressing the speech spectrum level in one frequency band as a function of the speech spectrum levels of the other frequency bands using (4.8). Without loss of generality, E_{i_f} is expressed as

$$E_{i_\mathrm{f}}(\underline{E}_{\setminus i_\mathrm{f}}) = 10 \log \left\{ \frac{1}{f_{\Delta, i_\mathrm{f}}} \cdot \left(\mathfrak{P}^{\max}(\kappa) - \sum_{i=i_\mathrm{f}+1}^{i_\mathrm{l}} f_{\Delta, i} \cdot 10^{E_i/10} \right) \right\} \tag{4.10}$$

with the sliced vector

$$\underline{E}_{\setminus i_\mathrm{f}} = (E_{i_\mathrm{f}+1}, E_{i_\mathrm{f}+2}, \ldots, E_{i_\mathrm{l}}), \tag{4.11}$$

resulting in the maximization problem

$$\underline{E}^{\mathrm{opt}}_{\setminus i_\mathrm{f}}(\kappa) = \arg\max_{\underline{E}_{\setminus i_\mathrm{f}}} \left\{ I_{i_\mathrm{f}} \cdot A_{i_\mathrm{f}}\big(E_{i_\mathrm{f}}(\underline{E}_{\setminus i_\mathrm{f}}), D_{i_\mathrm{f}}\big) + \sum_{i=i_\mathrm{f}+1}^{i_\mathrm{l}} I_i \cdot A_i(E_i, D_i) \right\} \tag{4.12}$$

subject to the *bound* constraint

$$-50\,\mathrm{dB} \overset{!}{\leq} E_i \overset{!}{\leq} E_i^{\mathrm{adm}}(\kappa) \quad \forall\, i_\mathrm{f} \leq i \leq i_\mathrm{l} \tag{4.13}$$

with

$$E_i^{\mathrm{adm}}(\kappa) = \min\big\{ D_i(\kappa) + 15\,\mathrm{dB},\ E_i^{\max}(\kappa) \big\}\,. \qquad [4.5,\ \mathrm{p.}\,55]$$

Note, that E_i is lower bounded only to stabilize the numerical optimization. The lower bound is chosen in concordance with (ANSI S3.5 1997).

To ensure convergence of this numerical optimization scheme to a global maximum, the band audibility function is approximated by a function $\hat{A}_i(E_i, D_i)$, which is depicted in Figure 4.1. For this purpose, the limitations of the auxiliary variable $K_i(E_i, D_i)$ in (2.43) and the limitation to zero of the level distortion factor $L_i(E_i)$ in (2.41) are omitted, which results in

$$\hat{A}_i(E_i, D_i) = \frac{E_i - D_i(\kappa) + 15\,\mathrm{dB}}{30\,\mathrm{dB}} \cdot \min\left\{ 1,\ 1 - \frac{E_i - U_i - 10\,\mathrm{dB}}{160\,\mathrm{dB}} \right\} \tag{4.14}$$

and leads to a strictly concave optimization function.

The solution $\underline{E}^{\mathrm{opt}}_{\setminus i_\mathrm{f}}(\kappa)$ of this optimization problem can be found with the trust-region-reflective algorithm[1]. The optimum speech spectrum level of the preceding update interval $\underline{E}^{\mathrm{opt}}_{\setminus i_\mathrm{f}}(\kappa - 1)$ can be used as initial estimate for the solution in order to reduce the number of iterations.

In the following, this method, which was presented in a similar manner in (Sauert & Vary 2010a), is called *numerical power-constrained SII-based optimization* (OptSIInum (A3)).

[1] The implementation in the *Matlab* function `fmincon` was used for the simulations.

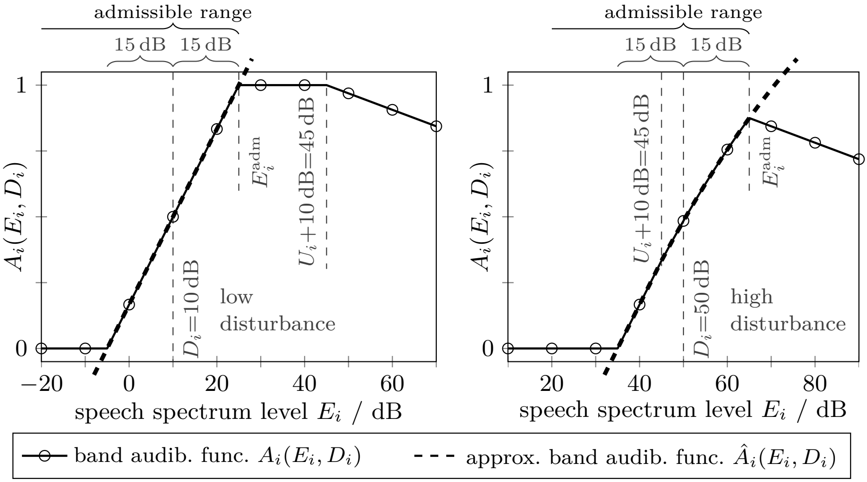


Figure 4.1: Exemplary plot of a *concave* approximation of the band audibility function for low as well as high disturbance case (cf. Figure 3.1).

4.1.2 Recursive Closed-Form Optimization (OptSIIrecur (A4))

For the closed-form optimization presented in this section, the band audibility function is approximated by the linear function

$$\hat{A}_i(E_i, D_i) = \frac{E_i - D_i(\kappa) + 15\,\mathrm{dB}}{30\,\mathrm{dB}} \cdot \min\left\{1,\ 1 - \frac{D_i(\kappa) + 15\,\mathrm{dB} - U_i - 10\,\mathrm{dB}}{160\,\mathrm{dB}}\right\}, \tag{4.15}$$

which is exemplarily plotted in Figure 4.2.

In the most relevant range $D_i(\kappa) - 15\,\mathrm{dB} \leq E_i \leq D_i(\kappa) + 15\,\mathrm{dB}$, the approximation $\hat{A}_i(E_i, D_i)$ is identical to $A_i(E_i, D_i)$ in the low disturbance case and slightly underestimates $A_i(E_i, D_i)$ in the high disturbance case.

The $(i_\mathrm{l} - i_\mathrm{f} + 1)$-dimensional equality constrained maximization problem (3.1) with constraint (4.8) can be solved using Lagrange multipliers. Following the definition of the SII (2.44), the Lagrange function is stated as

$$\hat{S}(\underline{E}, \underline{D}, \lambda) = \sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} I_i \cdot \hat{A}_i(E_i, D_i) + \lambda \cdot \left(\mathfrak{P}^{\max}(\kappa) - \sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} f_{\Delta,i} \cdot 10^{E_i/10}\right) \tag{4.16}$$

$$= \sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} \mathit{\Gamma}_i \cdot \frac{E_i - D_i(\kappa) + 15\,\mathrm{dB}}{30\,\mathrm{dB}} + \lambda \cdot \left(\mathfrak{P}^{\max}(\kappa) - \sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} f_{\Delta,i} \cdot 10^{E_i/10}\right) \tag{4.17}$$

with the Lagrange multiplier λ. The term

$$\mathit{\Gamma}_i = I_i \cdot \min\left\{1,\ 1 - \frac{D_i(\kappa) + 15\,\mathrm{dB} - U_i - 10\,\mathrm{dB}}{160\,\mathrm{dB}}\right\} \tag{4.18}$$

contains all factors of the summand $I_i \cdot \hat{A}_i(E_i, D_i)$ which are independent of E_i.

Differentiating $\hat{S}(\underline{E}, \underline{D}, \lambda)$ with respect to E_i and λ leads to a system of $i_\mathrm{l} - i_\mathrm{f} + 2$ equations

$$\frac{\mathrm{d}\hat{S}(\underline{E}, \underline{D}, \lambda)}{\mathrm{d}E_{i_\mathrm{f}}} = \frac{\mathit{\Gamma}_{i_\mathrm{f}}}{30\,\mathrm{dB}} - \lambda \cdot \frac{\ln(10)}{10} \cdot f_{\Delta,i_\mathrm{f}} \cdot 10^{E_{i_\mathrm{f}}/10} \overset{!}{=} 0 \tag{4.19}$$

$$\frac{\mathrm{d}\hat{S}(\underline{E}, \underline{D}, \lambda)}{\mathrm{d}E_{i_\mathrm{f}+1}} = \frac{\mathit{\Gamma}_{i_\mathrm{f}+1}}{30\,\mathrm{dB}} - \lambda \cdot \frac{\ln(10)}{10} \cdot f_{\Delta,i_\mathrm{f}+1} \cdot 10^{E_{i_\mathrm{f}+1}/10} \overset{!}{=} 0 \tag{4.19$'$}$$

$$\vdots \qquad\qquad\qquad \vdots$$

$$\frac{\mathrm{d}\hat{S}(\underline{E}, \underline{D}, \lambda)}{\mathrm{d}E_{i_\mathrm{l}}} = \frac{\mathit{\Gamma}_{i_\mathrm{l}}}{30\,\mathrm{dB}} - \lambda \cdot \frac{\ln(10)}{10} \cdot f_{\Delta,i_\mathrm{l}} \cdot 10^{E_{i_\mathrm{l}}/10} \overset{!}{=} 0 \tag{4.19$''$}$$

$$\frac{\mathrm{d}\hat{S}(\underline{E}, \underline{D}, \lambda)}{\mathrm{d}\lambda} = \mathfrak{P}^{\max}(\kappa) - \sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} f_{\Delta,i} \cdot 10^{E_i/10} \overset{!}{=} 0 \tag{4.20}$$

with the closed-form solution

$$E_i^{(1)} = 10 \log\left\{\frac{\mathit{\Gamma}_i}{\sum\limits_{i_\mathrm{f} \le \zeta \le i_\mathrm{l}} \mathit{\Gamma}_\zeta} \cdot \frac{\mathfrak{P}^{\max}(\kappa)}{f_{\Delta,i}}\right\}. \tag{4.21}$$

This solution might, however, fall outside the admissible range and violate the bound (4.7). Therefore, further recursion steps $\upsilon = 2, 3, \dots$ might be necessary (usually in less than 20 % of the coefficient updates). In this case, the solution $E_i^{(\upsilon-1)}$ of the preceding step is limited to $E_i^{\mathrm{adm}}(\kappa)$ and the closed-form solution (4.21) is adapted for the not limited subbands:

$$E_i^{(\upsilon)} = \begin{cases} E_i^{\mathrm{adm}}(\kappa) & \text{if } E_i^{(\upsilon-1)} \ge E_i^{\mathrm{adm}}(\kappa) \\ 10 \log\left\{\dfrac{\mathit{\Gamma}_i}{\sum\limits_{\substack{i_\mathrm{f} \le \zeta \le i_\mathrm{l} \,\wedge \\ E_\zeta^{(\upsilon-1)} < E_\zeta^{\mathrm{adm}}(\kappa)}} \mathit{\Gamma}_\zeta} \cdot \dfrac{\mathfrak{P}^{\max,(\upsilon)}(\kappa)}{f_{\Delta,i}}\right\} & \text{if } E_i^{(\upsilon-1)} < E_i^{\mathrm{adm}}(\kappa) \end{cases} \tag{4.22}$$

for $\upsilon > 1$, where

$$\mathfrak{P}^{\max,(\upsilon)}(\kappa) = \mathfrak{P}^{\max}(\kappa) - \sum_{\substack{i_\mathrm{f} \le \zeta \le i_\mathrm{l} \,\wedge \\ E_\zeta^{(\upsilon-1)} \ge E_\zeta^{\mathrm{adm}}(\kappa)}} f_{\Delta,\zeta} \cdot 10^{E_\zeta^{\mathrm{adm}}(\kappa)/10} \tag{4.23}$$

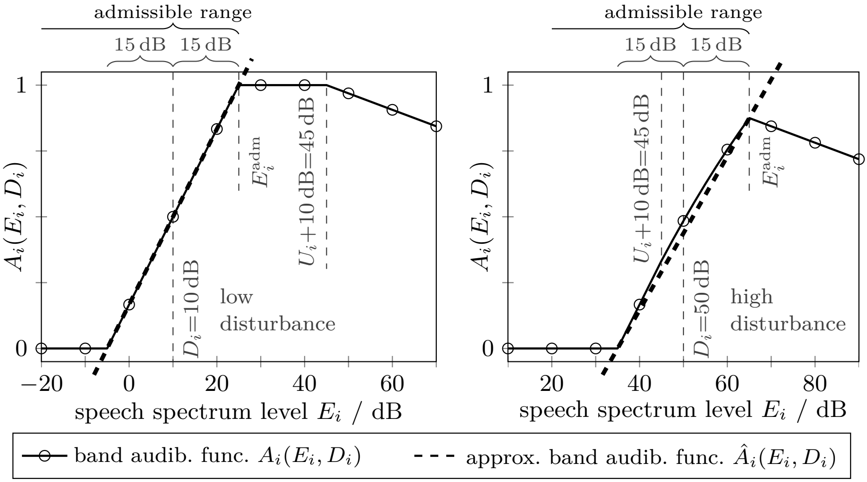


Figure 4.2: Exemplary plot of a *linear* approximation of band audibility function for low as well as high disturbance case (cf. Figure 3.1).

is the remaining power budget for the not limited subbands. Equation (4.22) is repeated recursively until all subbands fulfill $E_i^{(\upsilon)} \leq E_i^{\text{adm}}(\kappa)$, leading after $\varUpsilon \leq i_\text{l} - i_\text{f} + 1$ recursion steps to the final solution

$$\underline{E}^{\text{opt}}(\kappa) = \underline{E}^{(\varUpsilon)} \, . \tag{4.24}$$

In the vast majority of update intervals (usually more than 95 %), only $\varUpsilon \leq 2$ recursion steps are necessary to find the final solution.

Note, that in an actual implementation of this method, which is called *recursive closed-form power-constrained SII-based optimization* (OptSIIrecur (A4)), the calculations can be performed in the linear domain instead of decibel, which renders most instances of logarithm and exponentiation unnecessary and reduces the complexity of the weighting rule even further. OptSIIrecur (A4) was first published in (Sauert & Vary 2010b).

4.1.3 Reduction of Change of Tone Color

If $\underline{E} = \underline{E}^{\text{adm}}(\kappa)$ fulfills constraint (4.2), the remaining power budget cannot further increase the SII as discussed above. However, it can be used to reduce the attenuation and the change of tone color of the speech signal.

As for OptSIIbound (A1) in Section 3.2.1, a reduction of speech subband power should be prevented if the power constraint permits. But if the power constraint enforces at least some attenuation, the remaining power budget should be used to make the subbands weights as uniform as possible.

This leads to the following optimum speech spectrum level, cf. (3.19) of OptSIIbound (A1):

$$E_i^{\mathrm{opt}}(\kappa) = \min\Big\{\max\Big\{D_i(\kappa) + 15\,\mathrm{dB},\ E_i^{\mathrm{in}}(\kappa) - \varepsilon\Big\},\ E_i^{\max}(\kappa)\Big\} \tag{4.25}$$

with minimal $\varepsilon \geq 0\,\mathrm{dB}$ such that $\underline{E} = \underline{E}^{\mathrm{opt}}(\kappa)$ fulfills constraint (4.2). The minimal ε is found using a waterfilling technique.

If $\mathfrak{P}^{\max}(\kappa)$ is sufficiently high, constraint (4.2) is fulfilled with $\varepsilon = 0\,\mathrm{dB}$ and the optimization scheme turns into OptSIIbound (A1). Otherwise,

- subbands with a higher disturbance spectrum level $D_i(\kappa)$ are "lower bounded" to $D_i(\kappa) + 15\,\mathrm{dB}$,
- the speech spectrum levels of the remaining subbands (with a higher SNR) are chosen to a unified attenuation ε below the input speech spectrum level, with ε being as small as $\mathfrak{P}^{\max}(\kappa)$ permits.

Please note, that this quick overview disregards some details of (4.25).

4.1.4 Analysis

In this section, the OptSIIrecur (A4) scheme of Section 4.1.2 and the reduction of change of tone color of Section 4.1.3 are studied concerning the resulting subband weights for different SNRs. In order to characterize the general behaviour of the algorithm, it is again assumed that the maximum allowed speech spectrum level $E_i^{\max}$ is large enough such that $E_i^{\max} > D_i + 15\,\mathrm{dB}$ holds in all frequency bands and $E_i^{\max}$ can thus be ignored.

As derived in the following, with increasing SNR, the subband weights first have a bandpass characteristic, then the spectral shape of the output speech roughly follows that of the noise, and, finally, no modification is applied. These segments and the transitions between them are plotted in Figure 4.3 exemplarily for white noise and a constant allowed output audio power.

Low SNR: Bandpass Characteristic

For low SNR[2] with $D_i(\kappa) + 15\,\mathrm{dB} \geq E_i^{(1)}$ in all subbands, recursion stops after $\Upsilon = 1$ step with the first closed-form solution as optimum solution:

$$E_i^{\mathrm{opt}}(\kappa) = E_i^{(1)} = 10\log\left\{\frac{\Gamma_i}{\sum\limits_{i_{\mathrm{f}}\leq\zeta\leq i_{\mathrm{l}}}\Gamma_\zeta}\cdot\frac{\mathfrak{P}^{\max}(\kappa)}{f_{\Delta,i}}\right\}. \qquad \text{[cf. 4.21, p. 58]}$$

For speech babble noise, where $D_i(\kappa) - U_i$ and thus the second factor of

$$\Gamma_i = I_i \cdot \min\left\{1,\ 1 - \frac{D_i(\kappa) + 15\,\mathrm{dB} - U_i - 10\,\mathrm{dB}}{160\,\mathrm{dB}}\right\} \qquad \text{[4.18, p. 58]}$$

[2] The maximum total audio power $\mathfrak{P}^{\max}(\kappa)$ determines which SNR is "low"; a higher allowed power requires a lower SNR.

is approximately constant over all subbands, (4.21) simplifies to

$$E_i^{\mathrm{opt}}(\kappa) \approx 10 \log \left\{ \frac{I_i}{\sum\limits_{i_\mathrm{f} \leq \zeta \leq i_\mathrm{l}} I_\zeta} \cdot \frac{\mathfrak{P}^{\max}(\kappa)}{f_{\Delta,i}} \right\}. \tag{4.26}$$

For all other broadband noise types, (4.26) is not exact but still a reasonable approximation. Interestingly, in the low SNR situations considered here, the optimum speech spectrum level is thus approximately independent of the spectral characteristics of the noise.

It further follows that the optimum subband power $P_{s,i}^{\mathrm{opt}}(\kappa)$ of the output speech is approximately distributed according to the band importance function in that subband:

$$P_{s,i}^{\mathrm{opt}}(\kappa) \approx \frac{I_i}{\sum\limits_{i_\mathrm{f} \leq \zeta \leq i_\mathrm{l}} I_\zeta} \cdot \mathfrak{P}^{\max}(\kappa)\,. \tag{4.27}$$

The optimum subband weights accordingly result in

$$W_i(\kappa) \approx \sqrt{\frac{I_i}{\sum\limits_{i_\mathrm{f} \leq \zeta \leq i_\mathrm{l}} I_\zeta} \cdot \frac{\mathfrak{P}^{\max}(\kappa)}{\hat{P}_{s,i}^{\mathrm{in}}(\kappa)}}\,. \tag{4.28}$$

In (ANSI S3.5 1997, Table 1), the band importance function I_i is defined to be constant for all frequency bands between 400 Hz and 4.4 kHz with declining values for lower as well as higher frequencies. Accordingly, the optimum subband powers tend to be equally distributed over all subbands between 400 Hz and 4.4 kHz. Since the input speech usually has a spectral lowpass tilt, the spectral weighting shows (at higher sampling rates) a bandpass character with its maximum at about 5.8 kHz, cf. Figure 4.3b. At a sampling rate of 8 kHz, however, this bandpass acts like a highpass.

Medium SNR: Noise-Like Spectral Shape

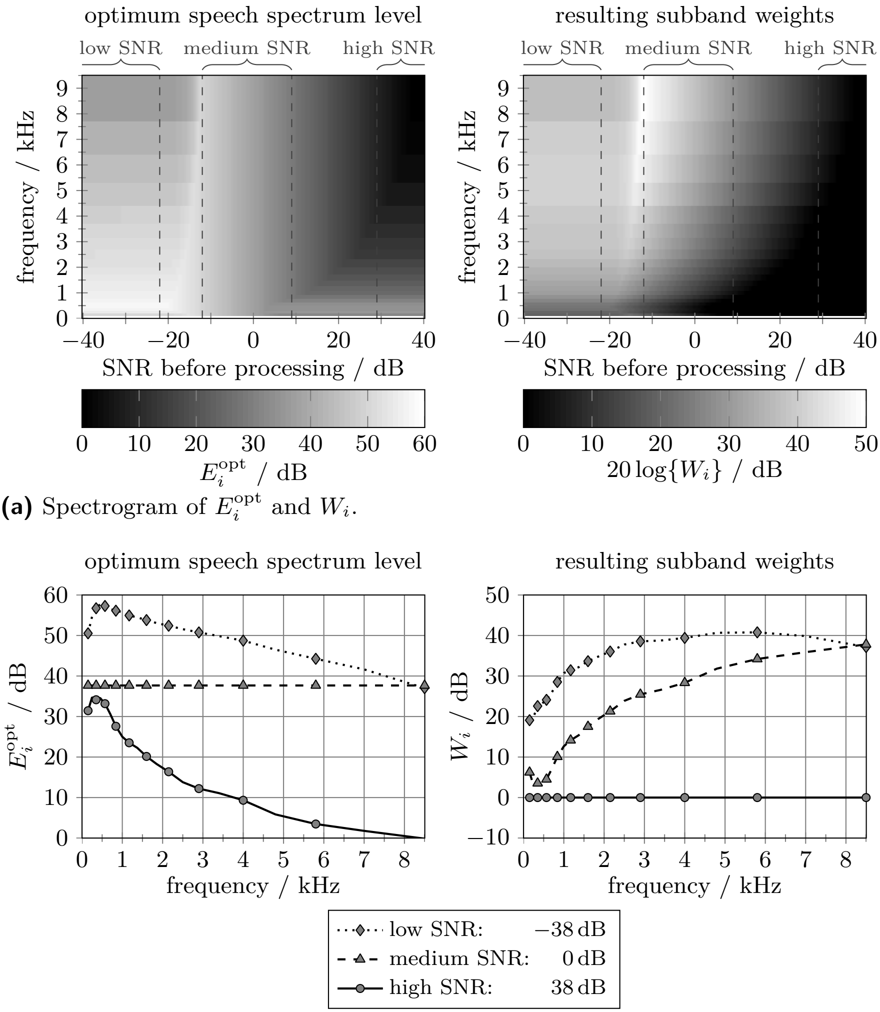
With increasing SNR[3], the disturbance spectrum level $D_i(\kappa)$ and thus $E_i^{\mathrm{adm}}(\kappa)$ decreases. Accordingly, in more and more subbands the first solution (4.21) will fall outside the admissible range and will, during the next recursion step, be limited to $E_i^{\mathrm{adm}}(\kappa) = D_i(\kappa) + 15\,\mathrm{dB}$.

Finally, the optimum solution converges towards

$$E_i^{\mathrm{opt}}(\kappa) = D_i(\kappa) + 15\,\mathrm{dB} \quad \forall\, i_\mathrm{f} \leq i \leq i_\mathrm{l} \qquad \text{[cf. 3.23, p. 47]}$$

and the spectral shape of the output speech roughly follows that of the noise. In this state of medium SNR, the situation is on the one hand bad enough, that

[3] Again, the maximum total audio power $\mathfrak{P}^{\max}(\kappa)$ determines the SNR at which this state is reached.



(a) Spectrogram of E_i^{opt} and W_i.

(b) Plot of E_i^{opt} and W_i for three exemplary SNRs.

Figure 4.3: Optimum speech spectrum level E_i^{opt} and subband weights W_i for maximum allowed output audio power $10\log\{\frac{\mathfrak{P}^{\max}}{P_0}\} = 90\,\mathrm{dB_{SPL}}$ based on speech spectrum level at normal voice effort U_i and perfect white noise (at the ear).
low SNR: bandpass characteristic of weights W_i
medium SNR: noise-like spectral shape of output speech
high SNR: no modification

enhancement is necessary, but on the other hand good enough, that not all power must be used to maximize the SII.

If $\mathfrak{P}^{\max}(\kappa)$ is sufficiently high, the power constraint (4.2) is not active anymore and OptSIIrecur (A4) turns into OptSIIbound (A1) (see Section 3.2.1).

If $\mathfrak{P}^{\max}(\kappa)$ is tighter, (3.23) is exact only in a very small SNR range. With decreasing distortion spectrum levels, the released power budget is used to reduce the change of tone color. However, due to the tight power limit, the optimum speech spectrum level can only reach $E_i^{\text{in}}(\kappa) - \varepsilon$, with the uniform attenuation ε decreasing to 0 dB with further increasing SNR.

High SNR: No Modification

At high SNR, the noise becomes less dominant and $E_i^{\text{adm}}(\kappa)$ becomes smaller than the input speech spectrum level. Then, the optimum speech spectrum level turns for all $\mathfrak{P}^{\max}(\kappa)$ to

$$E_i^{\text{opt}}(\kappa) = E_i^{\text{in}}(\kappa) \tag{4.29}$$

to prevent attenuation in noise-free environments. This results in 0 dB subband weights

$$W_i(\kappa) = 1 \tag{4.30}$$

and an unmodified speech signal.

4.1.5 Limited Bounded Optimization (LimOptSIIbound (A5))

The OptSIIbound (A1) algorithm presented in Section 3.2.1 can also be modified to obey the audio power constraint (4.2) by a subsequent frequency independent weight limitation (Sauert & Vary 2011):

$$W_i''(\kappa) = \begin{cases} W_i'(\kappa) & \text{if } \displaystyle\sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} W_i'^2(\kappa) \cdot \hat{P}_{s,i}^{\text{in}}(\kappa) \leq \mathfrak{P}^{\max}(\kappa) \\ \sqrt{\dfrac{\mathfrak{P}^{\max}(\kappa)}{\displaystyle\sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} W_i'^2(\kappa) \cdot \hat{P}_{s,i}^{\text{in}}(\kappa)}} \cdot W_i'(\kappa) & \text{otherwise.} \end{cases} \tag{4.31}$$

For sufficiently high SNRs, the constraint is not active and this *limited bounded SII-based optimization* (LimOptSIIbound (A5)) as well as OptSIIrecur (A4) behave the same, i. e., identical to the original OptSIIbound (A1) as shown above. If the constraint is active, the results of this algorithm are suboptimal compared to OptSIIrecur (A4) but require a lower complexity, which makes this algorithm especially interesting if power limitation is just a seldomly used safety measure.

4.2 Constraint 1: *No* Increase of Total Power

In this section, the power-constrained SII-based optimizations developed in Section 4.1 are evaluated under the stricter Constraint 1, which forbids any increase of total audio power.

Additionally, Section 4.2.1 presents a previously published algorithm which is designed for this constraint and based on a different model of human speech understanding. The simulation results of the algorithms proposed so far are given in Section 4.2.2.

In Section 4.2.3, the OptSIIrecur (A4) scheme is analyzed with respect to narrow bandpass noises and two modifications are presented in Sections 4.2.4 and 4.2.5. Section 4.2.6 finally presents the simulation results of these modifications.

4.2.1 Method of Maximal Power Transfer (MaxTransfer (A6))

The approach of this section was first presented in (Sauert et al. 2006) using a DFT analysis-synthesis filterbank. It is based on a simple model of human speech understanding which is depicted in Figure 4.4. In principle, the far-end speech signal $s^{\mathrm{in}}(k)$ is filtered by NELE in order to assist the speech understanding process of the listener. In the model, speech understanding is deteriorated by the acoustical channel from loudspeaker to eardrum[4], which adds background noise $n(k)$ to the emitted speech signal $s^{\mathrm{out}}(k)$.

A reasonable, still simple model of human hearing is assumed which consists of two cognitive stages:

- a noise reduction pre-processing $H_{\mathrm{nr}}(f)$ which is applied directly to the mixture signal $s^{\mathrm{out}}(k) + n(k)$ resulting in $\tilde{s}^{\mathrm{out}}(k)$ and
- an independent process which performs the actual speech understanding.

This decomposition is justified by the fact that the basilar membrane of the inner ear performs a frequency analysis (Zwicker & Fastl 1999). It is therefore assumed that cognitive signal processing could "easily" reduce the distorting noise before the actual speech understanding will happen. This is represented by the noise reduction filter $H_{\mathrm{nr}}(f)$ which is, of course, not exactly known and might not even be a linear filter. However, in a first attempt it is assumed that it acts like a Wiener filter, i. e., $H_{\mathrm{nr}}(f)$ will attenuate the signal $s^{\mathrm{out}}(k) + n(k)$ at frequencies where the SNR is low and preserve the signal where the SNR is high.

The aim of improving speech intelligibility in noisy acoustical environments motivates a method of *maximal power transfer* (MaxTransfer (A6)) from source $s^{\mathrm{in}}(k)$ to sink $\tilde{s}^{\mathrm{out}}(k)$. The key idea of MaxTransfer (A6) is to emit a signal $s^{\mathrm{out}}(k)$ primarily at those frequencies where the acoustical channel is (almost) clean, i. e., where the noise is low. This strategy will avoid a "waste" of energy for speech components on frequency channels which will be attenuated by the presence of

[4] This model assumes that the colorations which are caused by pinna and ear channel are the same for the speech and the noise signal and can therefore be neglected.

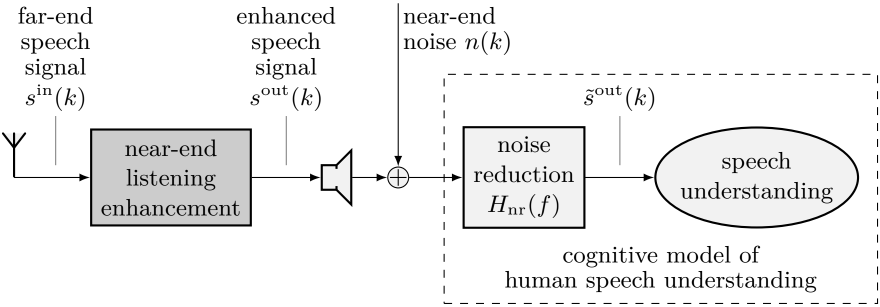

far-end speech signal $s^{\mathrm{in}}(k)$
enhanced speech signal $s^{\mathrm{out}}(k)$
near-end noise $n(k)$
$\tilde{s}^{\mathrm{out}}(k)$
near-end listening enhancement
noise reduction $H_{\mathrm{nr}}(f)$
speech understanding
cognitive model of human speech understanding

Figure 4.4: A simple cognitive model of human speech understanding in noisy acoustical environments.

noise reduction $H_{\mathrm{nr}}(f)$ in the model of hearing. This has a particular relevance if the power of the loudspeaker signal $s^{\mathrm{out}}(k)$ is constrained to the power of the original signal $s^{\mathrm{in}}(k)$.

The (heuristic) filter structure to assist the maximal power transfer from source to sink consists of the following two steps:

1. A frequency dependent attenuation

$$W_i(\kappa) = \frac{\dfrac{\hat{\overline{P}}_s(\kappa)}{K_1}}{\dfrac{\hat{\overline{P}}_s(\kappa)}{K_1} + \max\left\{\hat{P}_{n,i}(\kappa), \hat{P}_n^{\min}(\kappa)\right\}} \tag{4.32}$$

is applied to the far-end speech signal $s^{\mathrm{in}}(k)$. This scheme in general weights the subbands with the reciprocal of their noise subband power $\hat{P}_{n,i}(\kappa)$ but reverts to 0 dB in environments with no or very low noise compared to the average speech subband power

$$\hat{\overline{P}}_s(\kappa) = \frac{1}{i_{\mathrm{l}} - i_{\mathrm{f}} + 1} \sum_{i=i_{\mathrm{f}}}^{i_{\mathrm{l}}} \hat{P}_{s,i}(\kappa)\,. \tag{4.33}$$

The constant $10\log\{K_1\} = 20\,\mathrm{dB}$ denotes the "cut-off SNR" with $W_i(\kappa) = 0.5$ and is adjusted to deliver the best possible speech intelligibility.

A noise floor $\hat{P}_n^{\min}(\kappa)$ is applied to the estimated noise subband power $\hat{P}_{n,i}(\kappa)$, which limits the linear distortions produced by $W_i(\kappa)$ to a reasonable degree. It is chosen adaptively w. r. t. the average noise subband power as

$$\hat{P}_n^{\min}(\kappa) = K_2 \cdot \frac{1}{i_{\mathrm{l}} - i_{\mathrm{f}} + 1} \sum_{i=i_{\mathrm{f}}}^{i_{\mathrm{l}}} \hat{P}_{n,i}(\kappa), \tag{4.34}$$

with $10\log\{K_2\} = -7\,\mathrm{dB}$.

2. A frequency independent amplification is appended in order to match the power of the loudspeaker output signal $s^{\text{out}}(k)$ and the original speech signal $s^{\text{in}}(k)$:

$$W_i''(\kappa) = W_i(\kappa) \cdot \sqrt{\frac{\sum\limits_{i=i_\text{f}}^{i_\text{l}} \hat{P}_{s,i}^{\text{in}}(\kappa)}{\sum\limits_{i=i_\text{f}}^{i_\text{l}} W_i^2(\kappa) \cdot \hat{P}_{s,i}^{\text{in}}(\kappa)}}\,. \tag{4.35}$$

In this way, an amplification of the signal $s^{\text{in}}(k)$ at "audible" frequencies and an attenuation of the signal $s^{\text{in}}(k)$ at "inaudible" (i.e., strongly noise distorted) frequencies is achieved.

4.2.2 Simulation Results

Figure 4.5 shows the performance of OptSIIrecur (A4) presented in Section 4.1 and MaxTransfer (A6) under Constraint 1, i. e., without increase of total audio power.

It can be seen, that OptSIInum (A3) and OptSIIrecur (A4) have identical performance, which is to be expected since the concave approximation of OptSIInum (A3) and the linear approximation of OptSIIrecur (A4) are quite similar. However, it takes OptSIInum (A3) between 150 and 200 times longer[5] than OptSIIrecur (A4) to perform the optimization and calculate the subband weights $W_i(\kappa)$. Especially at medium SNRs the numerical optimization requires three times as much iterations to converge as at low SNRs.

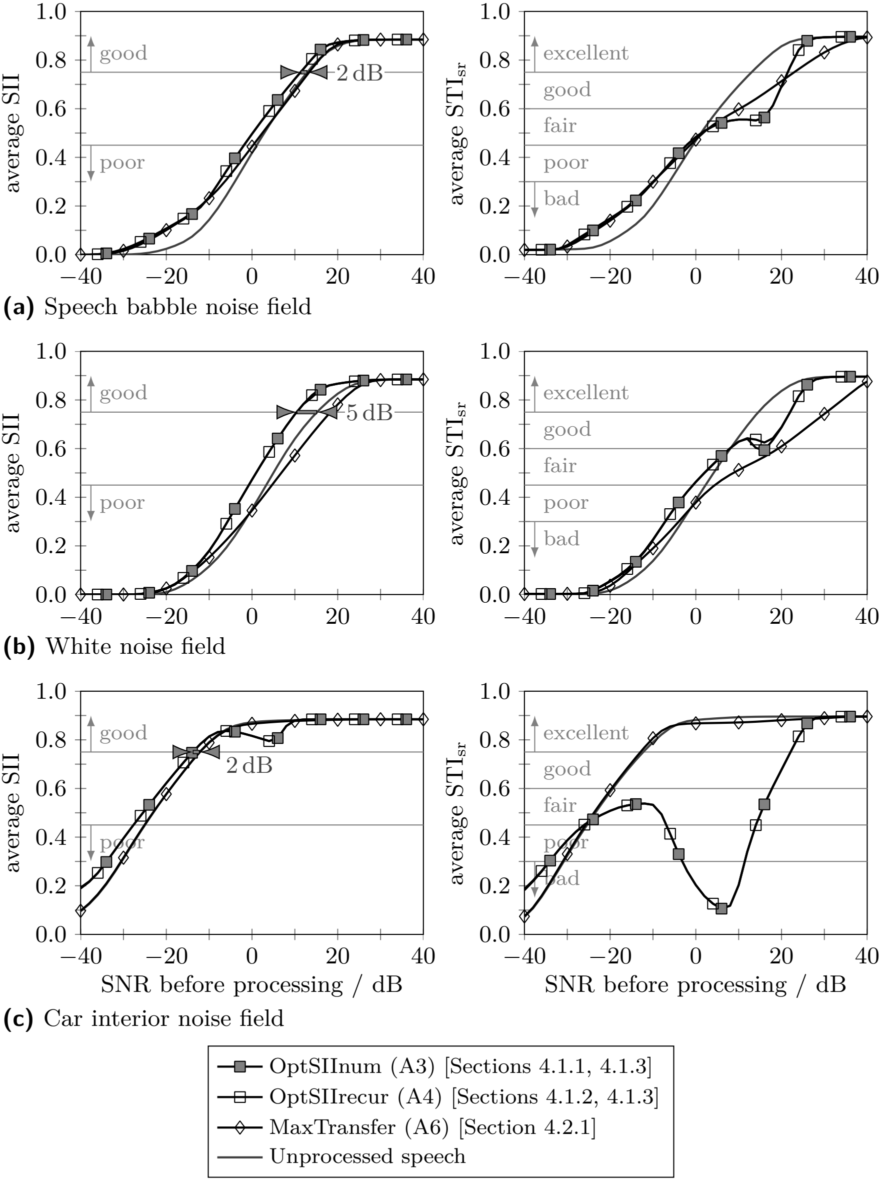
Both algorithms yield an SII gain of 5 dB for white noise and of 2 dB for speech and car interior noise (see Section 2.3.3), which is indicated in Figure 4.5 with arrows.

For the speech babble noise as well as especially the car interior noise, the STI_sr rating is deteriorated at a medium SNR range. This effect as well as countermeasures are discussed in the following.

The MaxTransfer (A6) method does not exhibit this deterioration in STI_sr but otherwise shows ambivalent results. For babble noise, the SII gain is slightly negative, for white noise, it is clearly negative. With the car interior noise, MaxTransfer (A6) is mostly inactive since all frequencies above 0.4 kHz are practically noise-free.

Apart from the deteriorated STI_sr ratings at medium SNR, OptSIIrecur (A4) outperforms MaxTransfer (A6) is all conditions and is therefore to be preferred, especially with the countermeasures discussed below.

[5] The algorithms were compared on an *Intel Pentium IV* with 2800 MHz and 2 GB RAM using reasonably optimized *Matlab* implementations.



(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 4.5: Comparison of OptSIIrecur (A4) and MaxTransfer (A6) without increase of total audio power. See Section 2.4 for simulation parameters. The arrows indicate the SII gain of OptSIIrecur (A4).

4.2.3 Analysis for Narrow Bandpass Noises

Although OptSIIrecur (A4) results in an optimized SII of the output speech and demonstrably increases intelligibility in various noise environments, the resulting optimum speech spectrum level can lead in special noise scenarios with a narrow bandpass characteristic to frequency weights which have a disadvantageous or even destructive effect on listening experience and speech intelligibility (which is, however, not really covered by the SII).

As derived above in Sections 3.2.2 and 4.1.4, OptSIIrecur (A4) simplifies for mid-range SNRs (and some other prerequisites depending on the algorithm) to

$$E_i^{\mathrm{opt}}(\kappa) = D_i(\kappa) + 15\,\mathrm{dB}\,. \qquad [3.23,\ \mathrm{p.}\,47]$$

If the noise signal has a very narrow bandpass disturbance spectrum, this leads to accordingly large weight factors in the corresponding frequency bands, which alone is at least annoying for the listener. If there is additionally a tight audio power constraint, all other frequency bands will be attenuated to allow amplification in the few noisy bands within the power constraint, which can compromise intelligibility.

The “narrow bandpass effect” becomes very apparent for the car interior noise, as this extreme noise signal accumulates almost all its energy in the three frequency bands below 0.4 kHz (see Figure 2.13), where the speech energy is weak. This can also be seen in the bottom right diagram of Figure 4.5, where the average STI$_{\mathrm{sr}}$

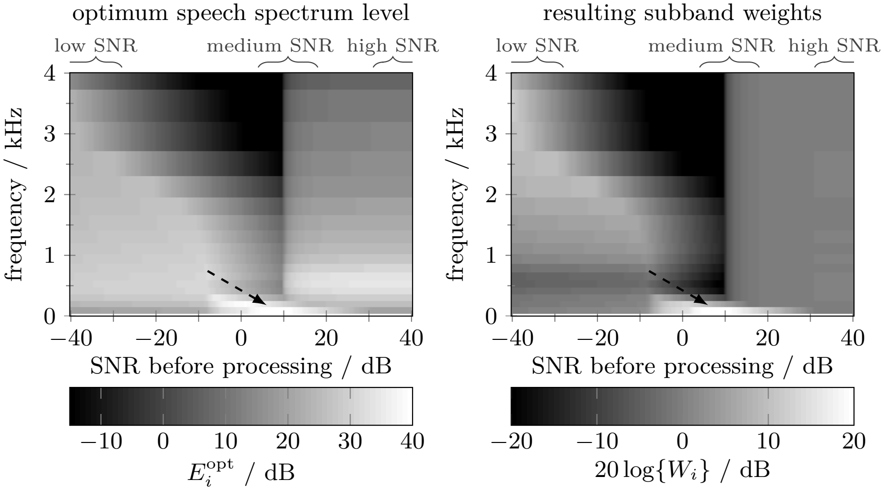


Figure 4.6: Effect of narrow bandpass noise on OptSIIrecur (A4). Optimum speech spectrum level and resulting subband weights after optimization without increase of total audio power based on average spectrum level of TIMIT database and car interior noise.

rating has a deep “notch” at a medium SNR range and reaches its minimum of below 0.1 at an SNR before processing of +5 dB.

In order to illustrate the deterioration of the $\mathrm{STI}_{\mathrm{sr}}$, the optimum speech spectrum level as well as the resulting weight factors are exemplarily plotted in Figure 4.6 for the average disturbance spectrum level of the car interior noise. At SNRs of −5 dB to 10 dB the lowest three frequency bands (marked by an arrow in Figure 4.6) are amplified by partially more than 20 dB, all other bands are attenuated by 15 dB to 20 dB. Thus, almost all energy of the output signal is concentrated at the lowest three frequency bands. To make things even worse, the input signal may have more noisy content in these bands than useful speech information, due to a high pitch of the far-end speaker or an unfavorable transfer characteristic of some element in the communication system chain.

Although this noise example has its narrow bandpass at very low frequencies where speech energy is low, a similar problem arises for other mono-frequent or bandpass noise sources with a peak at higher frequencies, like alarm signals or brake squeal of trains. Therefore, any solution which only copes with this special problem of car noise at low frequencies would not be sufficient.

4.2.4 A Priori Limitation of Disturbance (OptSIIrecurDist (A7))

The main cause for the problem of narrow bandpass noises discussed in the previous section is that the disturbance spectrum level has in few subbands much higher values than in all others. As a countermeasure to this problem, the disturbance spectrum levels can be restricted to be not larger than the threshold distance D_Δ above their average (calculated over frequency in decibel):

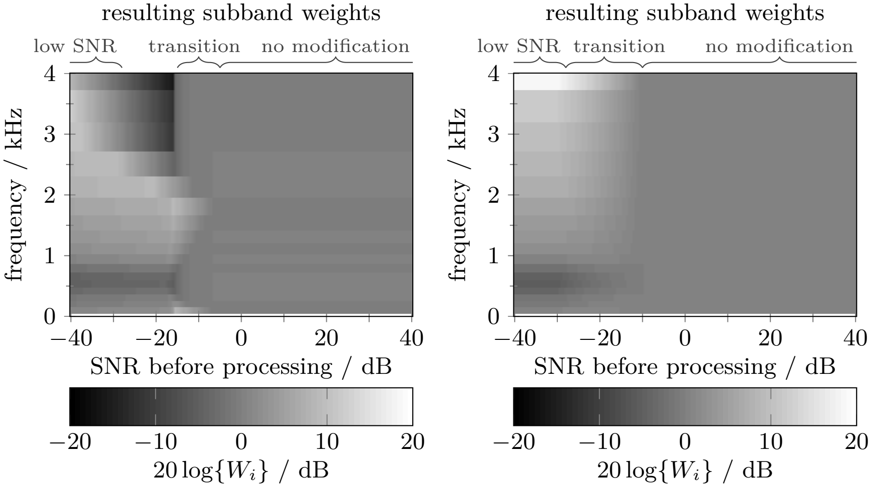
$$D_i'(\kappa) = \min\left\{ D_i(\kappa),\ \frac{1}{i_\mathrm{l} - i_\mathrm{f} + 1} \sum_{\zeta=i_\mathrm{f}}^{i_\mathrm{l}} D_\zeta'(\kappa) + D_\Delta \right\}. \tag{4.36}$$

This prevents that a single or few disturbance spectrum levels are much larger than the others, but does not restrict the overall dynamic since the limit is relative to the average. But, since (4.36) describes a non-linear function and $D_i'(\kappa)$ appears on the left-hand as well as right-hand side, it can not be calculated in a closed-form. Instead the restricted disturbance spectrum levels can be found recursively as follows.

In step $\upsilon = 0$, the average of the disturbance spectrum levels $D_i(\kappa)$ is taken as a starting point

$$\overline{D}^{(0)}(\kappa) = \frac{1}{i_\mathrm{l} - i_\mathrm{f} + 1} \sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} D_i(\kappa)\,. \tag{4.37}$$

If all frequency bands fulfill $D_i(\kappa) \leq \overline{D}^{(\upsilon)}(\kappa) + D_\Delta$, no further limitation is necessary which results in $D_i'(\kappa) = D_i(\kappa)$. In case of a narrow bandpass noise, there will be a certain number $I^{(\upsilon)}$ of frequency bands with $D_i(\kappa) > \overline{D}^{(\upsilon)}(\kappa) + D_\Delta$.



(a) OptSIIrecurDist (A7), $D_\Delta = 7\,\mathrm{dB}$ **(b)** OptSIIone (A8)

Figure 4.7: Resulting subband weights after algorithms OptSIIrecurDist (A7) and OptSIIone (A8) without increase of total audio power based on average spectrum level of TIMIT database and car interior noise. Compare with Figure 4.6.

During the next steps $\upsilon = 1, 2, \ldots$ of calculating the average disturbance spectrum level, all levels above the old restriction threshold $\overline{D}^{(\upsilon-1)}(\kappa) + D_\Delta$ are replaced by the new threshold $\overline{D}^{(\upsilon)}(\kappa) + D_\Delta$:

$$\overline{D}^{(\upsilon)}(\kappa) = \frac{1}{i_\mathrm{l} - i_\mathrm{f} + 1} \sum_{i=i_\mathrm{f}}^{i_\mathrm{l}} \begin{cases} D_i(\kappa) & \text{if } D_i(\kappa) < \overline{D}^{(\upsilon-1)}(\kappa) + D_\Delta \\ \overline{D}^{(\upsilon)}(\kappa) + D_\Delta & \text{otherwise} \end{cases} \tag{4.38}$$

$$= \frac{1}{i_\mathrm{l} - i_\mathrm{f} + 1} \left(\sum_{\substack{i_\mathrm{f} \le i \le i_\mathrm{l} \,\wedge \\ D_i(\kappa) < \overline{D}^{(\upsilon-1)}(\kappa) + D_\Delta}} D_i(\kappa) + I^{(\upsilon-1)} \cdot \left(\overline{D}^{(\upsilon)}(\kappa) + D_\Delta\right) \right), \tag{4.39}$$

which resolves to

$$\overline{D}^{(\upsilon)}(\kappa) = \frac{1}{i_\mathrm{l} - i_\mathrm{f} + 1 - I^{(\upsilon-1)}} \left(\sum_{\substack{i_\mathrm{f} \le i \le i_\mathrm{l} \,\wedge \\ D_i(\kappa) < \overline{D}^{(\upsilon-1)}(\kappa) + D_\Delta}} D_i(\kappa) + I^{(\upsilon-1)} \cdot D_\Delta \right). \tag{4.40}$$

If in at least one subband $\overline{D}^{(\upsilon)}(\kappa) + D_\Delta < D_i(\kappa) < \overline{D}^{(\upsilon-1)}(\kappa) + D_\Delta$ is true, i. e., the disturbance spectrum level $D_i(\kappa)$ passed the old threshold but not the new one, (4.40) is repeated recursively. After $\Upsilon \leq i_\mathrm{l} - i_\mathrm{f}$ recursion steps this leads to the final disturbance spectrum levels

$$D'_i(\kappa) = \min\Big\{D_i(\kappa),\ \overline{D}^{(\Upsilon)}(\kappa) + D_\Delta\Big\}\,. \tag{4.41}$$

In most cases, only $\Upsilon \leq 3$ recursion steps are necessary to find the final disturbance spectrum levels.

Figure 4.7a depicts the resulting subband weights of the *recursive closed-form power-constrained SII-based optimization with a priori limitation of disturbance spectrum level* (OptSIIrecurDist (A7)) with the threshold distance $D_\Delta = 7\,\mathrm{dB}$ for the example of Figure 4.6. At SNRs below $-15\,\mathrm{dB}$, the optimum weights are independent of the spectral shape of the noise as explained in Section 4.1.4 and are thus the same as of OptSIIrecur (A4) without limitation of the disturbance spectrum level. After a short transition range, OptSIIrecurDist (A7) reaches 0 dB weights for SNRs above $-5\,\mathrm{dB}$, where OptSIIrecur (A4) starts to exhibit the narrow bandpass weights.

Simulations show that $D_\Delta = 7\,\mathrm{dB}$ is a good compromise for an optimization without increase of audio power at sampling rate $f_\mathrm{s} = 8\,\mathrm{kHz}$. At $f_\mathrm{s} = 16\,\mathrm{kHz}$, D_Δ should be chosen to 8 dB. In case of the less tight Constraint 2, a higher threshold distance D_Δ of 12 dB is advisable.

4.2.5 One-Step Closed-Form Optimization (OptSIIone (A8))

The simulations of Section 4.2.2 showed, that OptSIIrecur (A4) performs well at low and high SNRs for all noise types. Only at medium SNRs the $\mathrm{STI_{sr}}$ rating is deteriorated for extreme bandpass noises, which is discussed in Section 4.2.3.

This motivates the new approach to replace the optimum subband weights of OptSIIrecur (A4) for medium SNRs by an interpolation between the weights at low SNR and the weights at high SNR.

As derived in Section 4.1.4, OptSIIrecur (A4) stops in low SNR situations after one recursion step with the first closed-form solution

$$E_i^{\mathrm{opt}}(\kappa) = E_i^{(1)} = 10 \log \left\{ \frac{\mathit{\Gamma}_i}{\sum\limits_{i_\mathrm{f} \leq \zeta \leq i_\mathrm{l}} \mathit{\Gamma}_\zeta} \cdot \frac{\mathfrak{P}^{\max}(\kappa)}{f_{\Delta,i}} \right\}. \qquad \text{[cf. 4.21, p. 58]}$$

With the audio power constraint (4.3) to the power of the input speech (Constraint 1, which is discussed in this section), (3.4), and (2.33), this results in the subband weights

$$W_i(\kappa) = \sqrt{\frac{\mathit{\Gamma}_i}{\sum\limits_{\zeta=i_\mathrm{f}}^{i_\mathrm{l}} \mathit{\Gamma}_\zeta} \cdot \frac{\sum\limits_{\zeta=i_\mathrm{f}}^{i_\mathrm{l}} \hat{P}^{\mathrm{in}}_{s,\zeta}(\kappa)}{\hat{P}^{\mathrm{in}}_{s,i}(\kappa)}} \tag{4.42}$$

in low SNRs. For high SNRs, 0 dB weights are approximately optimal as also shown in Section 4.1.4.

Among the different evaluated interpolation strategies, the weighting rule

$$W_i(\kappa) = \sqrt{\frac{\mathit{\Gamma}_i^{1-\gamma(\kappa)} \cdot \hat{P}^{\mathrm{in}}_{s,i}(\kappa)^{\gamma(\kappa)}}{\sum\limits_{\zeta=i_\mathrm{f}}^{i_\mathrm{l}} \mathit{\Gamma}_\zeta^{1-\gamma(\kappa)} \cdot \hat{P}^{\mathrm{in}}_{s,\zeta}(\kappa)^{\gamma(\kappa)}} \cdot \frac{\sum\limits_{\zeta=i_\mathrm{f}}^{i_\mathrm{l}} \hat{P}^{\mathrm{in}}_{s,\zeta}(\kappa)}{\hat{P}^{\mathrm{in}}_{s,i}(\kappa)}} \tag{4.43}$$

yielded the best results. The interpolation parameter $\gamma(\kappa)$ is time-varying with $0 \leq \gamma(\kappa) \leq 1$. For $\gamma(\kappa) = 1$, it simplifies to 0 dB weights, whereas the characteristic of (4.42) is obtained for $\gamma(\kappa) = 0$.

Figure 4.8 shows the performance of the *one-step closed-form power-constrained SII-based optimization* (OptSIIone (A8)) for some fixed parameters γ. As expected, OptSIIone (A8) yields for $\gamma = 0$ at very low SNR the same SII and STIsr as OptSIIrecur (A4) for all noise signals as well as for $\gamma = 1$ at very high SNR. In between, the STI_sr rating does not show the deep "notch" at the medium SNR range for all parameters γ, which could also be expected as the resulting weights of the presented optimization scheme do not exhibit a narrow bandpass characteristic. However, the changed weight characteristic of OptSIIone (A8) leads to lower SII ratings compared to OptSIIrecur (A4), especially for white noise.

Between very low and very high SNR there exists a transition range, in which the STI_sr ratings of smaller parameters γ saturate and are out-performed by higher ones (see Figure 4.8). Interestingly, the general shape of this transition range as well as its width of about 20 dB is the same for all evaluated noise signals, only its absolute position on the SNR scale varies.

In the next step, the parameter $\gamma(\kappa)$ is chosen adaptively based on the current speech and disturbance spectrum levels. In heuristic experiments it turned out that the transition between low and high SNR is well indicated by the *signal-to-disturbance ratio* (SDR) $\psi_i(\kappa)$ of the first closed-form solution (4.21) in the $\mathit{\Psi}$ "best" subbands. Therefore, the parameter $\gamma(\kappa)$ is derived as follows:

1. Calculate the speech spectrum level $E_i^{(1)}(\kappa)$ of the first closed-form solution according to (4.21).

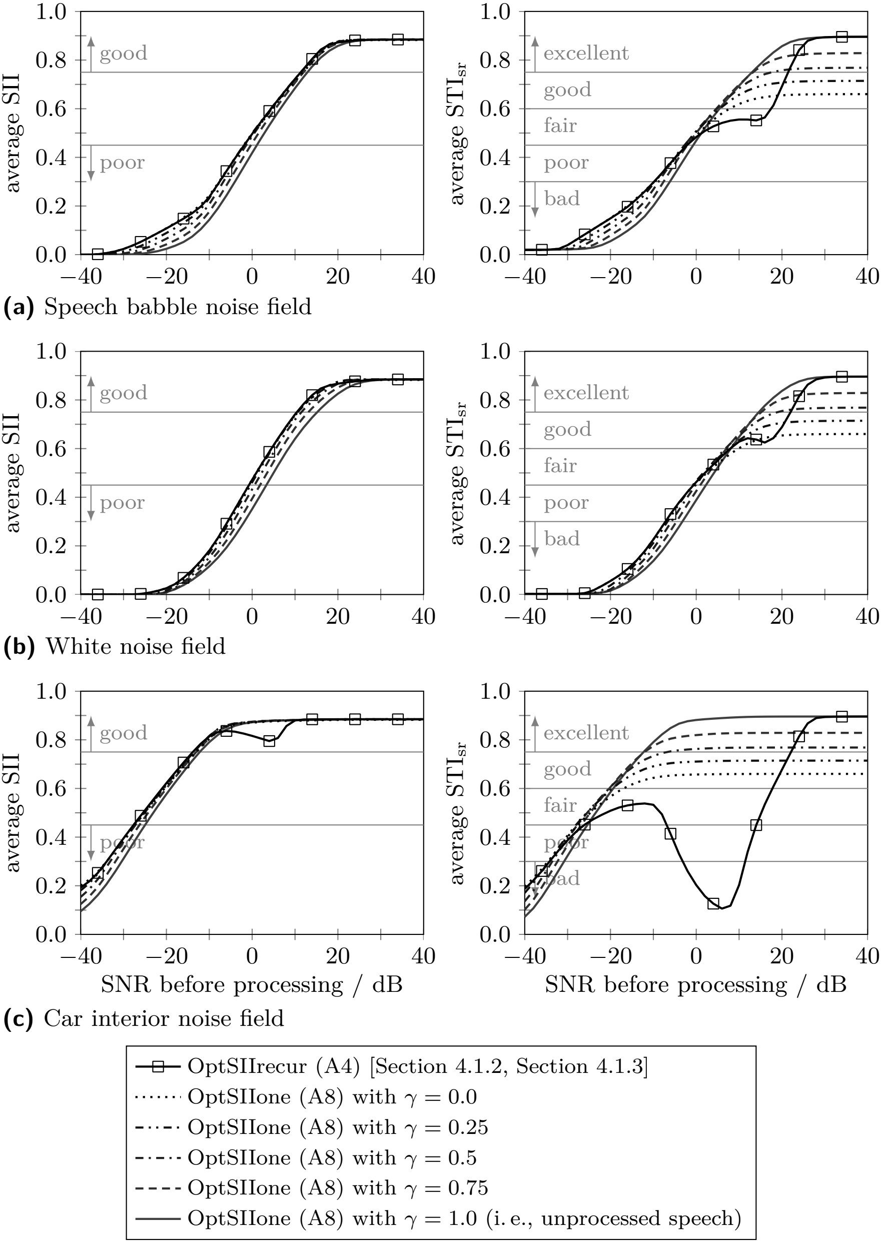
2. Calculate the SDR in decibel of the first closed-form solution:

$$\psi_i(\kappa) = E_i^{(1)}(\kappa) - D_i(\kappa)\,. \tag{4.44}$$

3. Calculate the average SDR $\bar{\psi}(\kappa)$ as the arithmetic mean (in decibel) of the $\mathit{\Psi}$ largest SDRs $\psi_i(\kappa)$.

4. Calculate the parameter $\gamma(\kappa)$ as

$$\gamma(\kappa) = \min\left\{ \max\left\{ \frac{\bar{\psi}(\kappa) - \psi_\mathrm{b}}{\psi_\mathrm{e} - \psi_\mathrm{b}}, 0 \right\}, 1 \right\}, \tag{4.45}$$



(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 4.8: Comparison of OptSIIone (A8) without increase of power for some fixed parameters γ. See Section 2.4 for simulation parameters.

where ψ_{b} and ψ_{e} denote the beginning and end of the transition range in decibel, respectively.

In the simulations, the settings $\Psi = 2$, $\psi_{\mathrm{b}} = 30\,\mathrm{dB}$, and $\psi_{\mathrm{e}} = 50\,\mathrm{dB}$ have provided a good compromise between $\mathrm{STI}_{\mathrm{sr}}$ and SII rating over the whole transition range and all noise signals.

Figure 4.7b sketches the resulting weights of OptSIIone (A8) for the example presented in Figure 4.6. The weights are similar to OptSIIrecurDist (A7) in the sense that OptSIIone (A8) yields 0 dB weights for SNRs above −10 dB. However, the transition range starts at lower SNRs and consists of a smooth and direct crossover to 0 dB weights, which is different for OptSIIrecurDist (A7).

OptSIIone (A8) was presented in (Sauert & Vary 2012b).

4.2.6 Simulation Results of Countermeasures

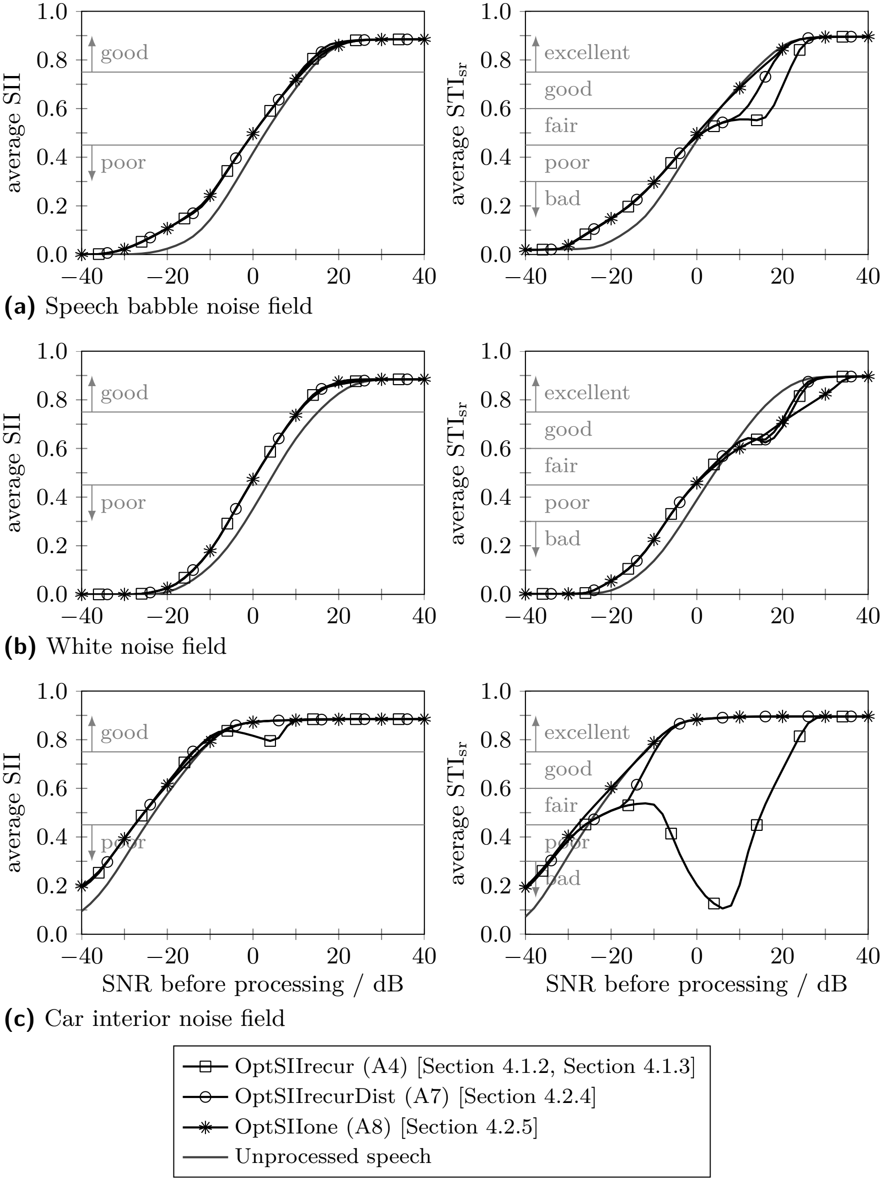
In this section, the success of the presented countermeasures to the problem of narrow bandpass weights is evaluated under the constraint that the short-term audio power in contributing subbands of the output signal is less than or equal to the short-term audio power in contributing subbands of the input signal (Constraint 1). Figure 4.9 compares OptSIIrecur (A4), OptSIIrecurDist (A7) with $D_{\Delta} = 7\,\mathrm{dB}$ and OptSIIone (A8) with $\Psi = 2$, $\psi_{\mathrm{b}} = 15\,\mathrm{dB}$, and $\psi_{\mathrm{e}} = 35\,\mathrm{dB}$.

For speech babble and especially for car interior noise, OptSIIrecurDist (A7) yields a dramatically better $\mathrm{STI}_{\mathrm{sr}}$ rating than OptSIIrecur (A4) at the medium SNR range: the "notches" are reduced to small "coves". For white noise, the $\mathrm{STI}_{\mathrm{sr}}$ ratings of OptSIIrecurDist (A7) and OptSIIrecur (A4) are very similar, which is to be expected as the disturbance spectrum level is almost spectrally flat and thus does not exceed the threshold often. The SII ratings of OptSIIrecurDist (A7) are in general identical to OptSIIrecur (A4) or even better in case of car interior noise and medium range SNR.

OptSIIone (A8) eliminated the "notches" in $\mathrm{STI}_{\mathrm{sr}}$ completely and still has a very comparable SII rating. Just for white noise between 21 dB and 35 dB SNR the $\mathrm{STI}_{\mathrm{sr}}$ rating is worse than for OptSIIrecur (A4) but still "excellent". It thus has the better performance of the two but is only applicable for the very strict power constraint to the input power as the calculation of the parameter $\gamma(\kappa)$ is based on equal input and output audio powers.

In contrast, OptSIIrecurDist (A7) works for any power constraint, only D_{Δ} should be adjusted depending on sampling rate and available total audio power budget.

A closing discussion of these results is given in Section 4.7



(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 4.9: Comparison of OptSIIrecurDist (A7) and OptSIIone (A8) without increase of total audio power. See Section 2.4 for simulation parameters.

4.3 Constraint 2: Increase of Total Power *Up To* Thermal Limit

The performances of the power-constrained SII-based optimizations developed in Section 4.1 are shown in Figure 4.10 under the constraint that the maximum allowed total audio power $\mathfrak{P}^{\max}$ is determined by the thermal limit of the loudspeaker (Constraint 2). For this evaluation, $10\log\left\{\frac{\mathfrak{P}^{\max}}{P_0}\right\} = 90\,\mathrm{dB_{SPL}}$ is chosen as motivated above.

A comparison with Figure 3.2 shows that the average SII and $\mathrm{STI_{sr}}$ rating of OptSIIrecur (A4) for speech babble and white noise drop below the average ratings of the (unlimited) OptSIIbound (A1) scheme at SNRs below $-12\,\mathrm{dB}$. In this SNR range, the power limitation becomes active and is a stricter constraint than the prevention of listener's hearing damage. At SNRs above $-12\,\mathrm{dB}$, all algorithms are basically unconstrained and thus lead to the same subband weights and, consequently, the same objective scores. Accordingly, the SII gains of the OptSIIrecur (A4) algorithm are with 23 dB to 26 dB similar to the gains of OptSIIbound (A1), cf., Section 3.3.

In comparison to LimOptSIIbound (A5), OptSIIrecur (A4) provides consistently better objective scores at SNRs below $-12\,\mathrm{dB}$.

The SII and $\mathrm{STI_{sr}}$ ratings of OptSIIrecur (A4) and OptSIIrecurDist (A7) with $D_{\triangle} = 12\,\mathrm{dB}$ are in most cases identical. For car interior noise at low SNRs, the OptSIIrecurDist (A7) algorithm yields even higher $\mathrm{STI_{sr}}$ scores. This is in accordance with informal listening experiments as OptSIIrecurDist (A7) avoids the annoyingly large weight factors in only few frequency bands.

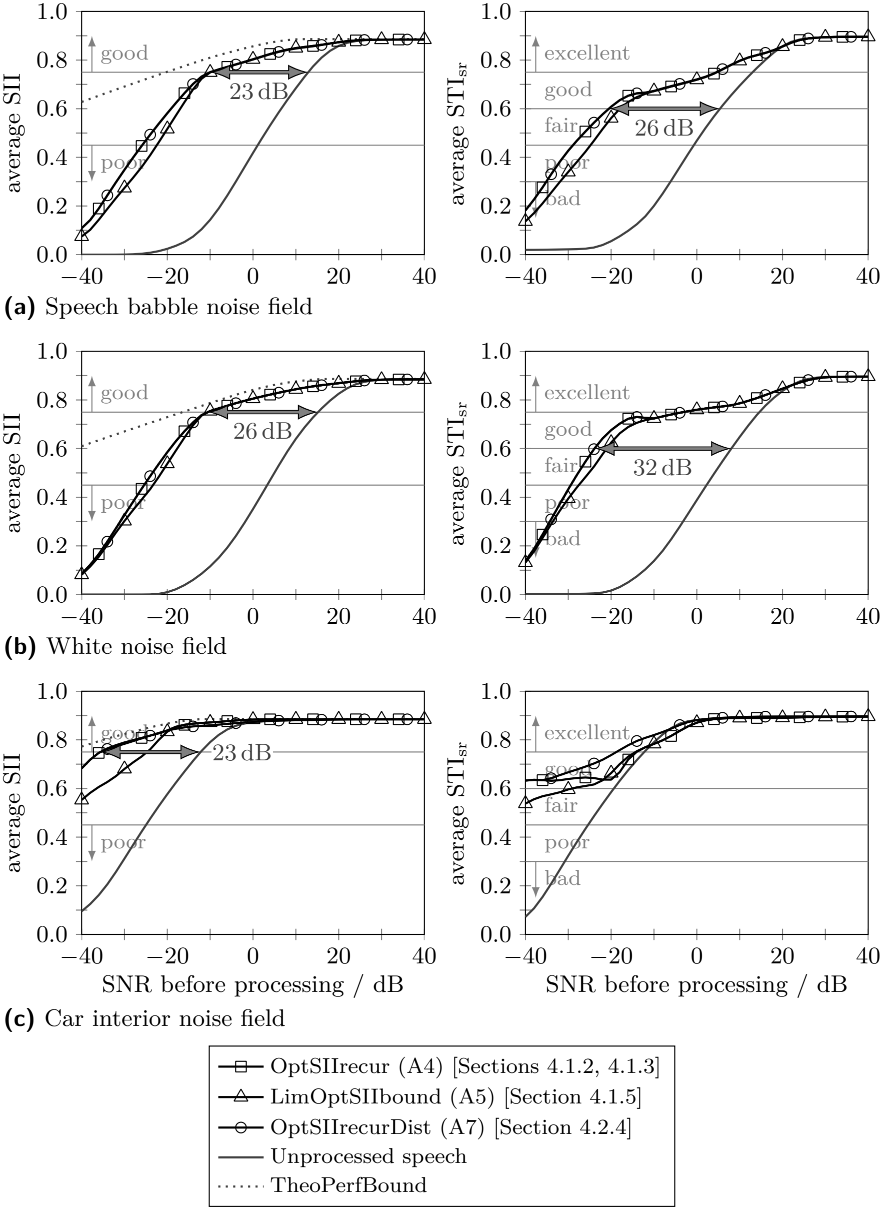
4.4 Simulation Results for 16 kHz Sampling Rate

Although all simulations so far were carried out with a sampling rate of $f_{\mathrm{s}} = 8\,\mathrm{kHz}$ for the sake of clarity, the presented algorithms work as described for arbitrary sampling rates.

Figure 4.11 depicts the performance of OptSIIrecur (A4) with an increase of total audio power up to $10\log\left\{\frac{\mathfrak{P}^{\max}}{P_0}\right\} = 90\,\mathrm{dB_{SPL}}$ and of OptSIIrecurDist (A7) without increase of total audio power, both for 16 kHz sampling rate. A comparison with Figures 4.9 and 4.10 shows that the overall characteristics of the objective measure curves are the same regardless of the sampling rate.

The OptSIIrecur (A4) algorithm with increase up to thermal limit yields an SII gain of 25 dB to 27 dB and an STI gain of 25 dB to 32 dB.

At 16 kHz sampling rate, the speech signal contains (almost) all frequencies considered by the SII and the $\mathrm{STI_{sr}}$. Consequently, the plain speech signal in silence reaches an SII of 1.0 and an $\mathrm{STI_{sr}}$ of 0.99 as opposed to a speech signal at 8 kHz sampling rate.

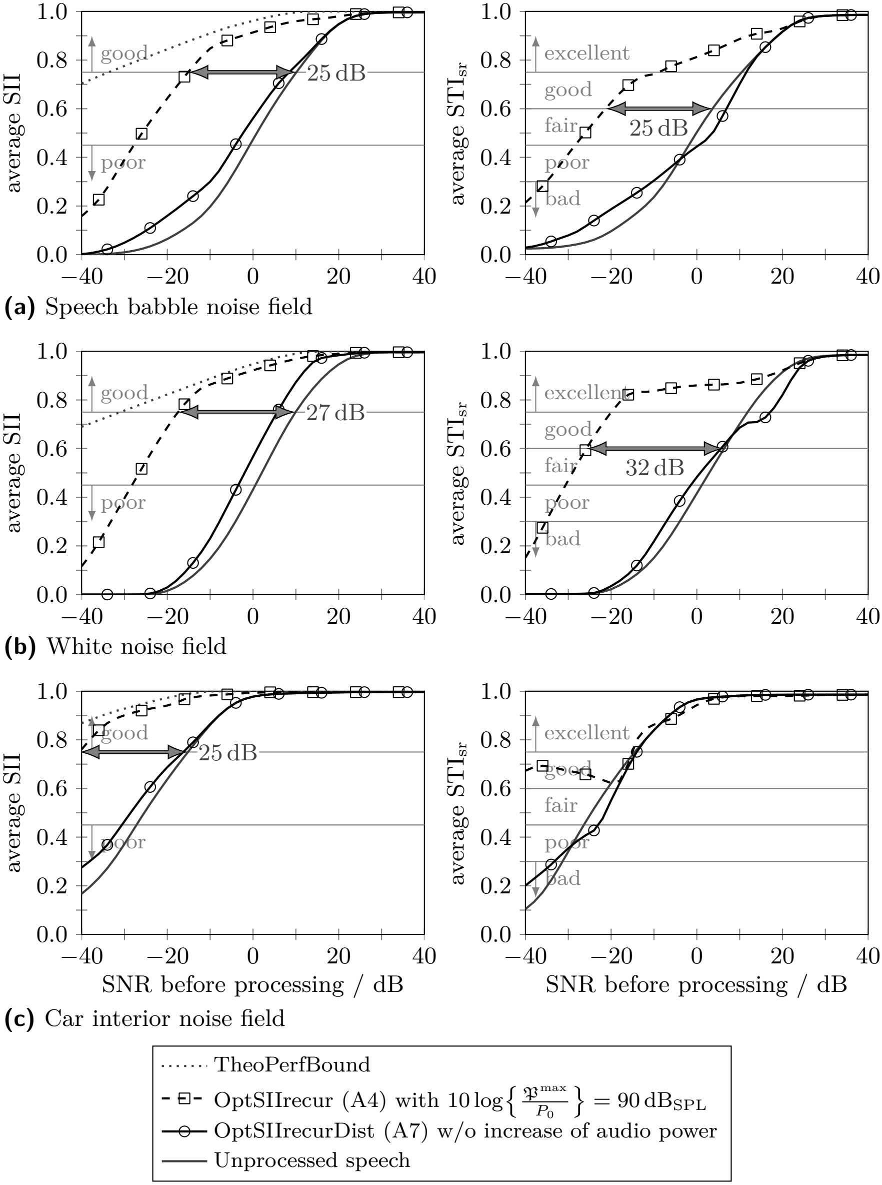


(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 4.10: Comparison of SII-based optimizations for max. output power $10 \log\left\{\frac{\mathfrak{P}^{\max}}{P_0}\right\} = 90\,\mathrm{dB}_{\mathrm{SPL}}$. See Section 2.4 for simulation parameters. Arrows indicate the SII and STI gain of OptSIIrecur (A4).



(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 4.11: Comparison of power-constrained SII-based optimizations at sampling rate $f_s = 16\,\mathrm{kHz}$. See Section 2.4 for simulation parameters. The arrows indicate the SII and STI gain of OptSIIrecur (A4).

4.5 Comparison with Literature

In this section, OptSIIone (A8) is compared with other NELE algorithms from literature in terms of average SII and $\mathrm{STI}_{\mathrm{sr}}$. Section 2.5 presents more details about these algorithms. None of the methods added audio power to the speech signal, which corresponds to Constraint 1. It should be noted, that generally reimplementations based in the published papers have been used.

Boosting of Consonant-Vowel-Ratio and Formant Enhancement

Figure 4.12 shows the comparison with algorithms from Thomas and Niederjohn, and colleagues. In (Thomas & Niederjohn 1970) and (Niederjohn & Grotelueschen 1976), highpass filtering is followed by "infinite amplitude clipping" and "rapid amplitude compression", respectively. With both approaches, speech *quality* is deteriorated, but in terms of SII, they perform almost identical to OptSIIone (A8) over the whole SNR range. The $\mathrm{STI}_{\mathrm{sr}}$ ratings of both approaches are also about the same but much worse than the rating of OptSIIone (A8) and the rating of unprocessed speech. They saturate for high SNR at "poor" (0.3). After all, the processing is noise-independent and thus introduces distortions in noise-free environments.

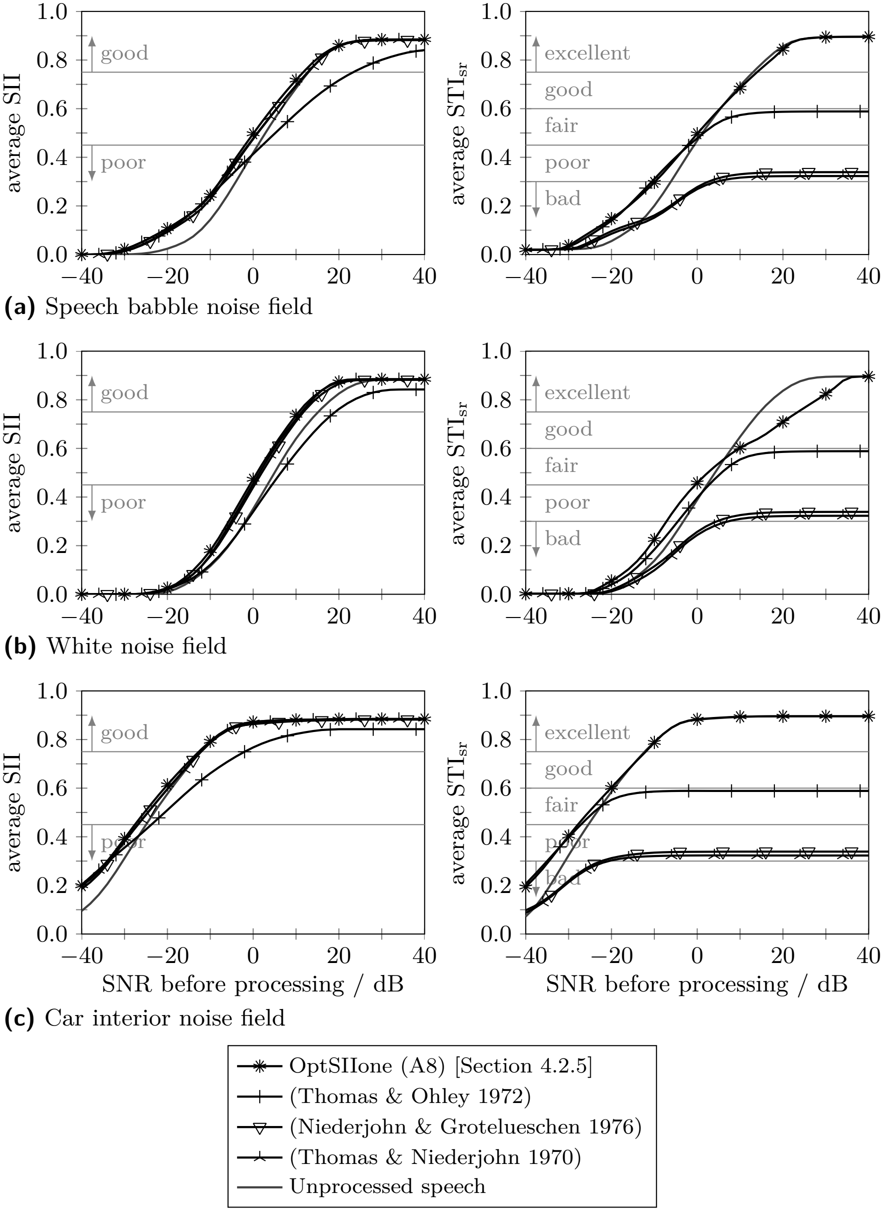
The approach of Thomas and Ohley (1972), which applies only a highpass filter, exhibits a different behaviour. The SII rating is lower than the rating of OptSIIone (A8) and of unprocessed speech, especially at higher SNRs. For white noise, the performance is always slightly worse than without processing. In all cases, the SII rating levels off at 0.84, which is below the rating of unprocessed speech. The $\mathrm{STI}_{\mathrm{sr}}$ ratings behave about the same as the SII, besides that they level off at 0.59 ("fair").

Obviously, infinite amplitude clipping and rapid amplitude compression improve the SII but decrease the $\mathrm{STI}_{\mathrm{sr}}$.

A comparison with algorithms from Skowronski and Harris as well as Chanda and S. Park is shown in Figure 4.13. Both algorithms have SII and $\mathrm{STI}_{\mathrm{sr}}$ ratings that are similar to the rating of unprocessed speech. However, the $\mathrm{STI}_{\mathrm{sr}}$ ratings degrade to "good" (0.62 resp. 0.73) at high SNR.

Enhancement of Pitch and Temporal Envelope

Figure 4.14 presents the performance of the algorithm from H. Park et al. (2010) with three enhancement levels L1, L2, and L3. The performance of this algorithm depends on the enhancement level and the SNR as concluded in (H. Park et al. 2010). For low SNR, the enhancement level L3 is better than L2 which is better than L1, while the reverse order is true for high SNR. This is especially true for the $\mathrm{STI}_{\mathrm{sr}}$, where enhancement level L3 reaches the performance of OptSIIone (A8) at low SNR, while level L1 is not a big improvement over the unprocessed speech. At high SNR, enhancement level L3 yields an $\mathrm{STI}_{\mathrm{sr}}$ of only 0.58 ("fair"), while level L1 achieves 0.71 ("good"), which is, however, still below the rating of unprocessed

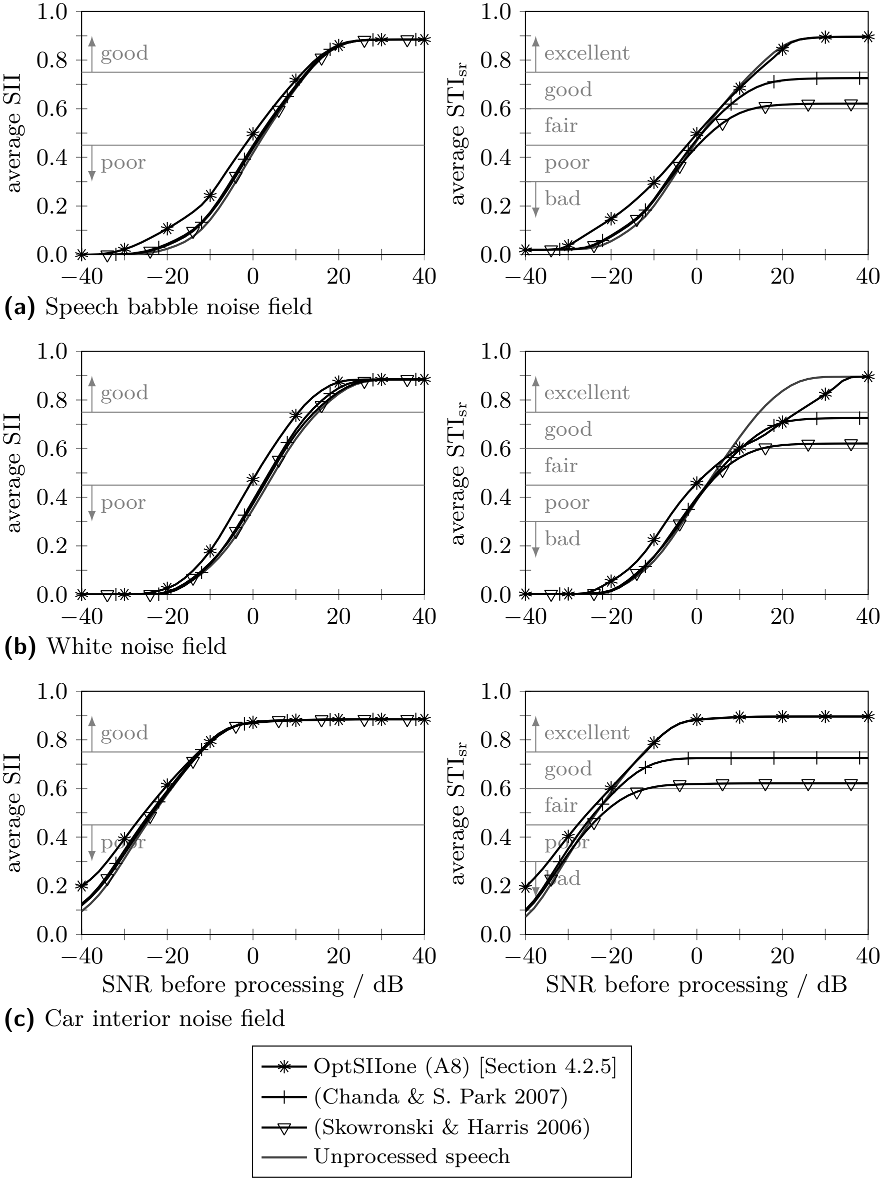


(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 4.12: Comparison of OptSIIone (A8) with algorithms from Thomas and Niederjohn, and colleagues without increase of total audio power. See Section 2.4 for simulation parameters.

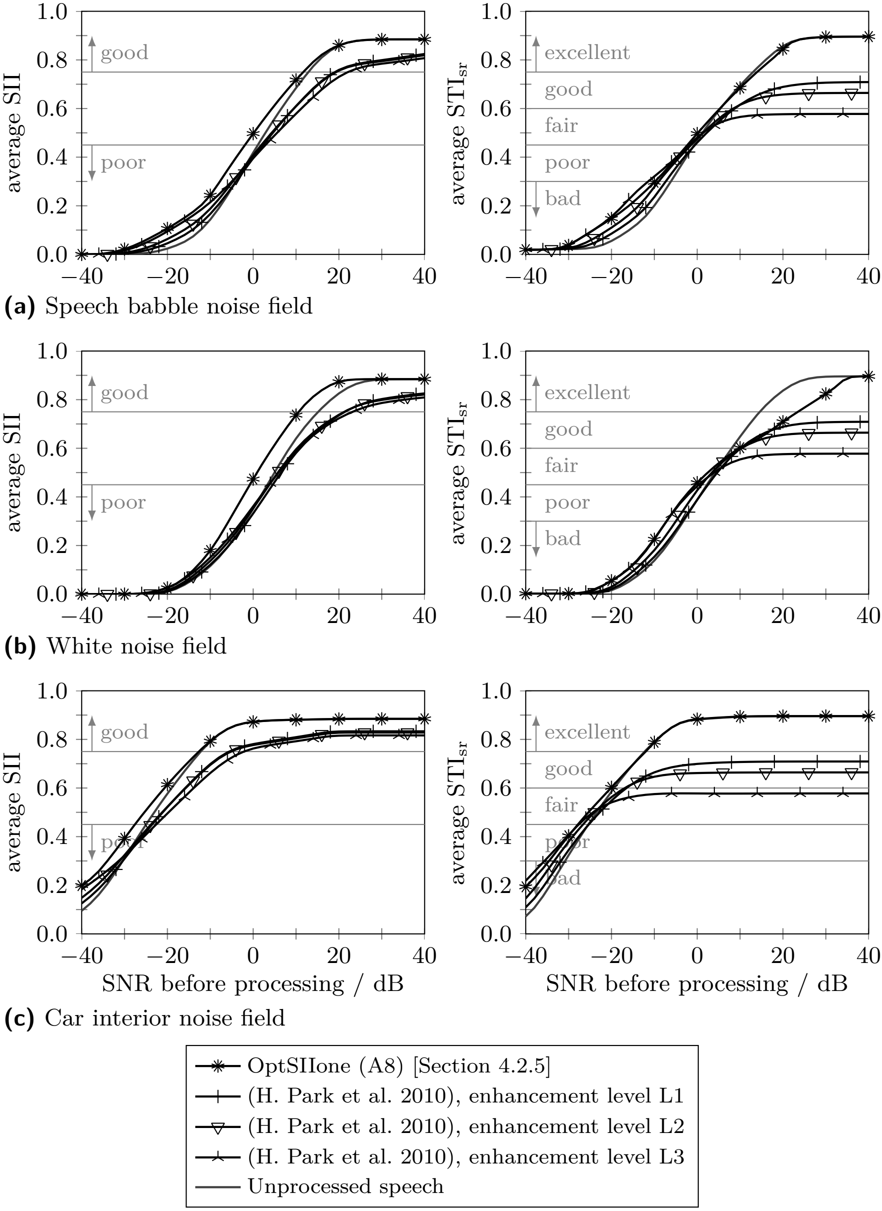


(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 4.13: Comparison of OptSIIone (A8) with algorithms from Skowronski and Harris as well as Chanda and S. Park without increase of total audio power. See Section 2.4 for simulation parameters.



(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 4.14: Comparison of OptSIIone (A8) with algorithm from H. Park et al. without increase of total audio power. See Section 2.4 for simulation parameters.

speech. The SII performance depends only marginally on the enhancement levels. For high SNR the rating clearly below the rating of unprocessed speech.

4.6 Listening Tests

In contrast to the preceding sections, which contain only evaluations with instrumental measure, this section shows the functionality of the developed NELE algorithms with three listening tests which compare them with speech modification algorithms from literature. Thereby, each test has a different focus and scale.

4.6.1 Evaluation of Modifications of Natural Speech

A large-scale speech intelligibility evaluation which compares speech modification algorithms under energy and duration constraints is presented in (Cooke et al. 2013).

The listening test was initiated to evaluate the progress within "The Listening Talker" consortium (LISTA) at the end of the second year of the project (Cooke et al. 2012), which is funded under the Seventh Framework Programme for Research of the European Union (FP7) beginning in May 2010.

"Apart from one system [OptSIIrecur (A4)], chosen since it represents the pre-LISTA state-of-the-art, only those techniques developed within the LISTA project have been evaluated" (Cooke et al. 2012).

In total, ten different "types"[6] of modified and unmodified speech were tested in this study at the same speech signal energy, cf. Table 4.1. Three are unmodified natural or synthetic speech ("plain", "Lombard", "TTS"), five modifications are applied to the natural plain speech ("OptSIIrecur (A4)", "OptGP", "SelBoost", "SSDRC", "TMDRC"), and for the final two types, the generation process of a text-to-speech (TTS) system has been modified ("TTSLomb", "TTSGP"). As the generation of synthetic speech is of minor focus in this thesis, the results of the "TTS" types are not reported in the following.

The listening test was intended to explore the best possible performance under ideal circumstances and not to describe a realistic application scenario. Therefore, the modification algorithms were allowed to enlarge the duration of the utterance by up to one second and to redistribute the signal energy over time but not to increase it in total. In fact, the recursive closed-form power-constrained SII-based optimization (OptSIIrecur (A4)), presented in Sections 4.1.2 and 4.1.3, was the only evaluated modification type which is suitable for a real-world scenario as it estimates all noise information blindly from the noise signal and does not modify the speech signal for high SNR. All other modified types, although being more recently developed, use either perfect knowledge about the masker signal or apply the same processing independent of the noise, i. e., also modify the speech signal in silence.

[6] The term "speech type" is used as a collective name for unmodified speech styles and the outputs of speech modification algorithms.

type	approach	reference	mode	noise dependency offline	 online
plain	unmodified neutral speech	–	natural	–	–
Lombard	unmodified Lombard speech	–	natural	–	–
OptSIIrecur (A4)	SII-optimized spectral reweighting	Sections 4.1.2, 4.1.3	natural	no	blind estimation of short-term subband power
OptGP	glimpse-optimized spectral reweighting	(Tang et al. 2012)	natural	yes	perfect noise type & SNR
SelBoost	boost just audible regions	(Tang et al. 2010)	natural	no	perfect SNR in time-frequency regions
TMDRC	harmonic model tilt modification followed by DRC	(Erro et al. 2012)	natural	yes	perfect noise type & SNR
SSDRC	spectral shaping followed by dynamic range compression (DRC)	(Zorilă et al. 2012)	natural	no	no
TTS	unmodified HMM-based text-to-speech (TTS)	–	synth.	–	–
TTSLomb	TTS adapted to Lombard	–	synth.	yes	no
TTSGP	glimpse-optimized TTS	(Val.-Bot. et al. 2012)	synth.	no	perfect short-term PSD

Table 4.1: Speech modification types tested in (Cooke et al. 2013). Results of TTS types are not presented in the following.

Methodology

A subset of the Harvard sentence materials (IEEE 1969) consisting of 180 phonetically balanced sentences, uttered by a male native British English talker, were presented in a stationary speech-shaped noise[7] at the three SNRs $-9\,\mathrm{dB}$, $-4\,\mathrm{dB}$, and $+1\,\mathrm{dB}$ as well as in competing speech from a female talker as fluctuating masker. The SNRs were selected from pilot tests in order to produce recognition rates of approximately 25 %, 50 %, and 75 %. Note, that a single competing speaker as near-end background noise does not constitute a realistic communication scenario and that therefore these results are not discussed in the following.

154 listeners aged predominantly between 19 and 25 years with English as native language participated in the test. 15 listeners were removed from the analysis because of failures in the audiological screening.

The listeners were asked to identify keywords[8] in speech in the six aforementioned noise conditions and the percentage of keywords identified correctly by listeners was scored. Additionally, the concept of *equivalent intensity change* (EIC) is introduced in (Cooke et al. 2013), which describes the amount in decibels by which plain speech would need to be amplified/attenuated to acquire the same intelligibility as the evaluated speech modification type.

Results

Figure 4.15 shows the change of the keyword scores in percentage points as well as the EIC relative to the scores of "plain" for the speech-shaped noise masker at three SNR levels. The noise dependency is indicated with different fillings.

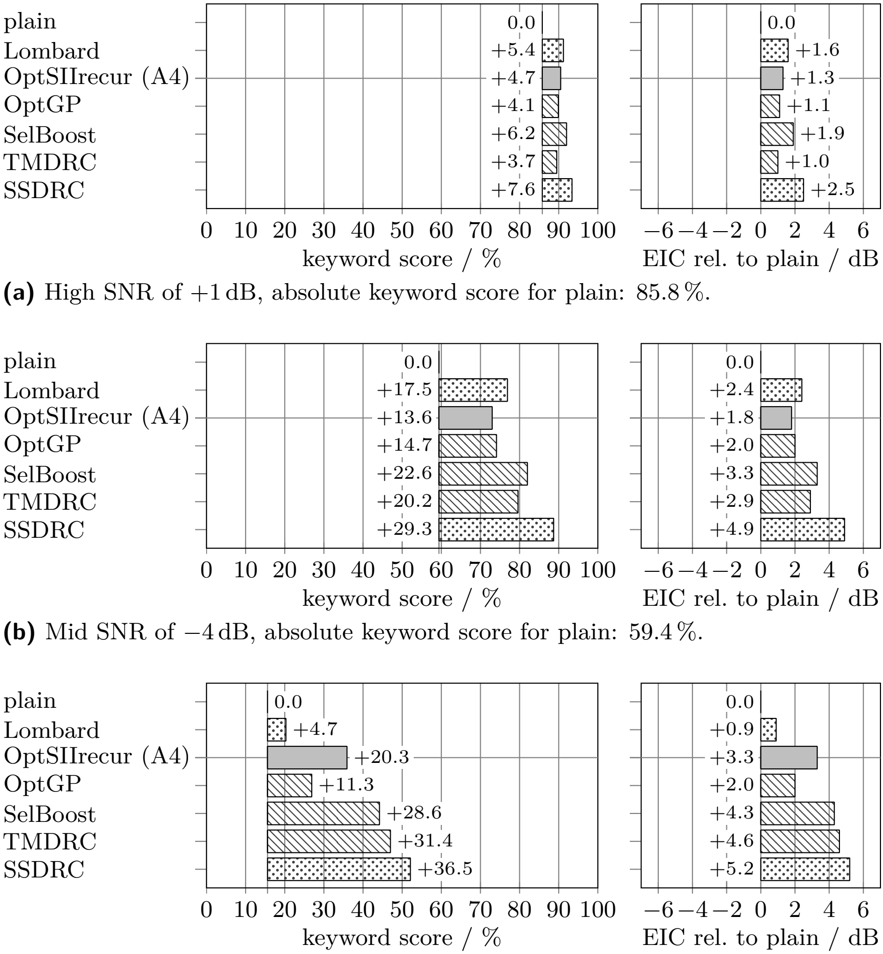
In all SNR conditions, OptSIIrecur (A4) was more intelligible than plain speech with an up to 20 percentage points higher keyword score. The intelligibility gain increases with decreasing SNR. For the highest SNR, the gain is limited by saturation effects.

OptSIIrecur (A4) yields around 1.5 dB of EIC at mid and high SNR, and around 3 dB at low SNR. For high and mid SNR, its gain is comparable to the Lombard speech, for low SNR, OptSIIrecur (A4) is much more intelligible.

OptGP, the second evaluated type which uses an objective intelligibility model, performs comparable to OptSIIrecur (A4) for high and mid SNR but is clearly inferior at low SNR. The SelBoost, the TMDRC, and the SSDRC method show a better performance than OptSIIrecur (A4) in most evaluated SNR conditions. However, all these more recent methods are either noise independent or rely on perfect knowledge of the disturbing noise which is only available in this listening test context. In contrast, the OptSIIrecur (A4) approach estimates all noise information blindly, can cope with double-talk situations, behaves transparent in noise-free environments, and prevents hearing damage of the listener.

[7] The "acoustical leakage" from sound source to the ear, cf. Section 2.1, was not considered in this evaluation.

[8] The term "keyword" means all words excluding the short common words 'a', 'the', 'in', 'to', 'on', 'is', 'and', 'of', and 'for'.



(a) High SNR of +1 dB, absolute keyword score for plain: 85.8 %.

(b) Mid SNR of −4 dB, absolute keyword score for plain: 59.4 %.

(c) Low SNR of −9 dB, absolute keyword score for plain: 15.6 %.

Figure 4.15: Change in keyword scores in percentage points and EIC relative to natural plain speech for the speech-shaped noise masker without increase of total audio power, cf. (Cooke et al. 2013, Figures 2 and 4).

▬ types with noise estimation
▧ types with perfect noise knowledge
▒ noise independent types

While it is a perfectly reasonable approach to mark the best performance possible under idealized circumstances, these algorithms can be expected to show a degraded intelligibility in less ideal or noise-free contexts, all the more when they are made usable for real-world scenarios.

4.6.2 Evaluation of Modifications of Synthetic Speech

Another related large-scale study is presented in (Valentini-Botinhao et al. 2013), which evaluates eight speech "types" in the context of synthetic speech (see Table 4.2).

Three speech modification algorithms, which were originally proposed for natural speech and were compared in the natural speech context in (Cooke et al. 2013), are here applied to synthetic speech. The performance of these methods ("TTS-OptSIIrecur (A4)", "TTS-SSDRC", "TTS-SSEDRC") is compared with a modification algorithm for producing optimized synthetic speech ("TTSGP") and two combinations of both ("TTSGP-DRC", "TTSGP-SSDRC"). The results of the three latter ones are again not reported in the following, as the generation of synthetic speech is of minor focus of this thesis.

Opposed to the preceding study, the OptSIIrecur (A4) here uses a simple moving average with $\tau_n = 2\,\mathrm{s}$ memory as noise subband power estimator (see Section 2.2.4).

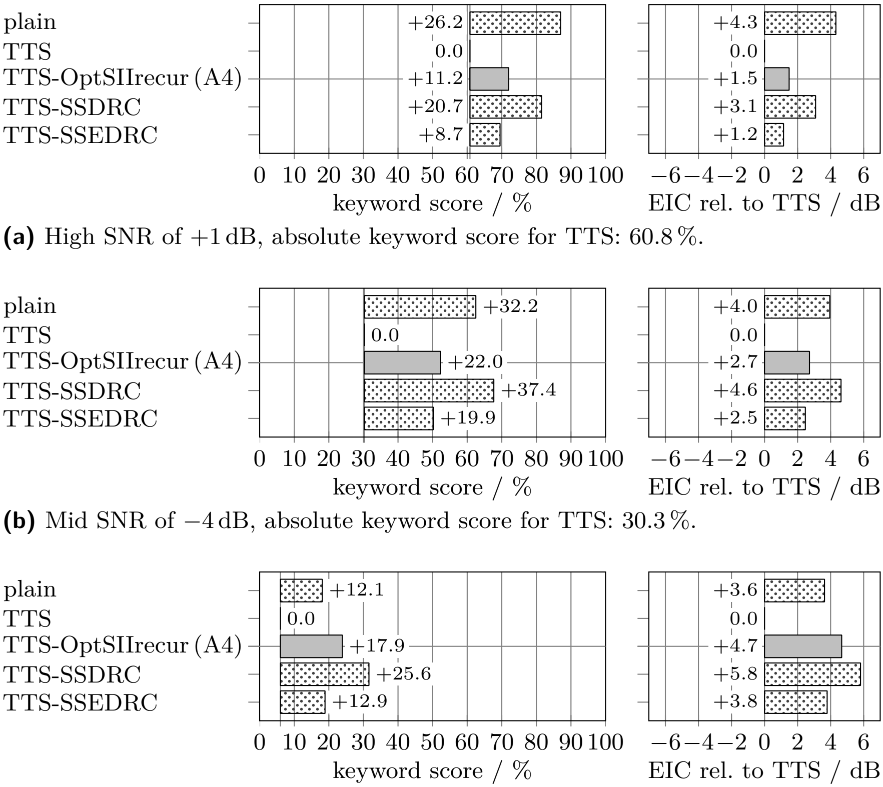
Sentence material, noise signals, SNRs, and all other methodology are the same as in (Cooke et al. 2013), cf. Section 4.6.1. 88 native English speakers participated in this test.

Results

Figure 4.16 shows the change of the keyword scores in percentage points as well as the EIC relative to the TTS type for the speech-shaped noise masker at three SNR

type	approach	noise dependency
plain	unmodified natural speech	–
TTS	unmodified HMM-based text-to-speech (TTS)	–
TTS-OptSIIrecur (A4)	SII-optimized spectral reweighting applied to TTS	blind estim.
TTS-SSDRC	spectral shaping followed by dynamic range compression (DRC) applied to TTS	no
TTS-SSEDRC	extended version of spectral shaping followed by DRC applied to TTS	no
TTSGP	glimpse-optimized TTS	perfect PSD
TTSGP-DRC	DRC applied to glimpse-optimized TTS	perfect PSD
TTSGP-SSDRC	SSDRC applied to glimpse-optimized TTS	perfect PSD

Table 4.2: Speech modification types tested in (Valentini-Botinhao et al. 2013). Results of TTSGP types are not presented in the following.



(a) High SNR of +1 dB, absolute keyword score for TTS: 60.8 %.

(b) Mid SNR of −4 dB, absolute keyword score for TTS: 30.3 %.

(c) Low SNR of −9 dB, absolute keyword score for TTS: 6.0 %.

Figure 4.16: Change in keyword scores in percentage points and EIC relative to TTS for the speech-shaped noise masker without increase of total audio power, cf. (Valentini-Botinhao et al. 2013, Figure 2).

▭ types with noise estimation
▨ noise independent types

levels. The noise dependency is again indicated with different fillings.

In all SNR conditions, TTS-OptSIIrecur (A4) was more intelligible than TTS with an up to 22 percentage points higher keyword score. For low SNRs, TTS-OptSIIrecur (A4) is also more intelligible than plain natural speech with a 5.8 percentage points higher score. TTS-OptSIIrecur (A4) yields between 1.5 dB and 4.7 dB of EIC relative to TTS with increasing EIC gain for decreasing SNR.

As in (Cooke et al. 2013), TTS-SSDRC has a better performance than TTS-OptSIIrecur (A4) in most evaluated conditions although the lead tends to be tighter. Again, it is to be expected that the intelligibility of TTS-SSDRC degrades in a noise-free context.

4.6.3 Speech Recognition Threshold for Numerals

The Bayesian adaptive speech intelligibility estimation (BASIE) method of Gaubitch et al. (2010) allows a rapid estimation of the speech recognition threshold (SRT), i. e., the minimum SNR at which an individual can recognize 50 % of the speech material (Plomp & Mimpen 1979). This adaptive estimation method chooses the probe SNR of the next trial based on the information of all previous trials and thus reduces the number of trials necessary to estimate the threshold.

In a third evaluation, BASIE was used to estimate the SRT for triplets of English numerals. This technique was chosen instead of a more sophisticated measure to allow comparing more algorithms with more subjects within the available time, although it represents a word recognition task with only a small vocabulary of phonetically unbalanced words. Due to this restriction, the absolute SRT values are not directly comparable with other studies, but can be at least an indication of the tendency for the algorithms.

Four speech "types", two methods presented in this thesis ("OptSIIrecur (A4)", "OptSIIone (A8)"), one method from literature ("Chanda"), and plain speech, are compared in the presence of speech babble noise (see Table 4.3).

The method of Chanda and S. Park was chosen for comparison since it represents a recent contribution to the popular class of algorithms which boost the consonant-vowel-ratio.

type	approach	reference	noise depend.
plain	unprocessed speech	–	–
OptSIIrecur (A4)	recursive closed-form power-constrained SII-based optimization	Sections 4.1.2, 4.1.3	blind estim.
OptSIIone (A8)	one-step closed-form power-constrained SII-based optimization	Section 4.2.5	blind estim.
Chanda	tunable equalization filter	(Chanda et al. 2007)	no

Table 4.3: Speech modification types tested with speech babble noise.

Methodology

Anechoic recordings of digit triplets from the TIDigits database (Leonard 1984; Leonard & Doddington 1993) were used as speech data. Triplets containing "oh" instead of "zero" were excluded from the dataset, leaving 973 triplets from female talkers and 934 triplets from male talkers.

The speech files from the TIDigits database were downsampled to 8 kHz sampling rate and scaled according to the active speech level (ITU-T P.56 1993). Similar to the simulations, each speech file was replicated twice and the three copies were concatenated. After processing, the first two thirds are cropped to avoid transient effects of the processing. The last third is combined with speech babble noise at 8 kHz at the required SNR. A lead and lag of 0.5 s noise was added at the beginning and at the end. In accordance with the listening test of (Cooke et al. 2013), the "acoustical leakage" $\overline{H}_{\text{leak}}$ from sound source to the ear (see Section 2.1) is not considered in this evaluation.

The samples were presented diotically in the sound booth at the Institute of Communication Systems and Data Processing at the RWTH Aachen University through *Sennheiser HD600* headphones. The headphones were calibrated using *HEAD acoustics*' *PEQ V* such that the active speech level of the unprocessed speech samples corresponds to a sound pressure level of about 62.35 dB_{SPL} as specified in (ANSI S3.5 1997) for normal voice effort, which is the same scaling as used for the simulations (see Section 2.4).

The method of Brookes (2012) was chosen as implementation of BASIE. It makes use of the standard keyboard instead of a graphical numeric keypad in (Gaubitch et al. 2010).

14 subjects, aged predominantly between 28 and 38, participated in the experiment. Ten listeners judged themselves as fluent in English language, four as good. None was aware of a significant hearing loss.

All speech modification types were compared in one listening test session per subject. Beforehand, the subjects received a short practice session to familiarize themselves with the experiment and the user interface. It consisted of five samples in speech babble noise, both starting at an SNR of 0 dB.

During the listening test, each type had on average 40 trials, resulting in 160 trials per session. The listeners had no information about the probe SNR and were not allowed to repeat the noisy speech sample. For each trial, the listener were forced to input three digits and the next trial did not start before the input for the preceding trial was finished. The listening test session was in this sense self-paced and took on average 16 minutes. Upon completion, chocolate was given away as gratification.

For each subject and modification type, the SRT is calculated as the average of the last ten trials of that type as suggested by (Gaubitch et al. 2010). The SRT gain, which is the difference between the plain SRT and the processed SRT, is of special interest as it indicates the benefit of processing. Positive gains imply an improvement in intelligibility while a negative gain implies a degradation.

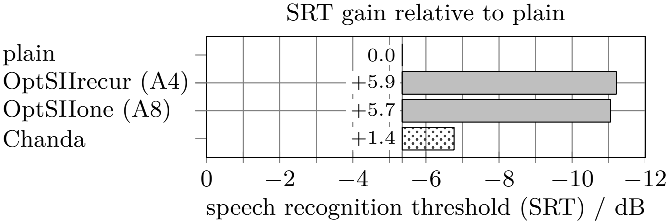


Figure 4.17: Speech recognition threshold (SRT) gain for each modification type without increase of total audio power relative to plain for speech babble noise.

▭ types with noise estimation
▨ noise independent types

Results

Figure 4.17 depicts the SRT gain relative to plain for speech babble noise. The underlying SRTs are averaged over all subjects, since there the results showed no difference between the groups with self-judged English skills “fluent” and “good”.

Both SII-based optimizations have an SRT gain of about +5.8 dB, i. e., they achieve the same intelligibility as plain speech at a 5.7 dB to 5.9 dB worse SNR. While OptSIIrecur (A4) is slightly ahead, the lead is not significant. The SRT gain of Chanda is +1.4 dB.

In general, these listening test results confirm the simulation results shown in Figure 4.12 and recommend the SII-based optimizations for application.

4.7 Discussion

The speech recognition test with human listeners show, that the SII-based optimizations effectively improve speech intelligibility in noisy environments. The same intelligibility is obtained at a +1.3 dB to +5.9 dB higher noise level, depending on type of noise, SNR, and recognition task.

The comparison with other state-of-the-art NELE algorithms shows – under idealized circumstances – for some algorithms a better performance than OptSIIrecur (A4) and a worse performance for others. However, all these methods have unrealistic requirements, which are only available in an evaluation context. In fact, the SII-based optimizations developed in this thesis are uniquely suitable for real-world applications, which includes a blind noise estimation from the microphone signal, consideration of double-talk situations, transparent behaviour in noise-free environments, and prevention of hearing damage of the listener as well as equipment damage.

In applications with a tight audio power constraint for the output signal, OptSIIone (A8) and OptSIIrecurDist (A7) are best suited. OptSIIone (A8) has slightly better instrumental measure scores at medium SNRs, but is, in the presented design, restricted to applications, where the audio power of the output signal is strictly confined to the power of the input signal. In contrast, OptSIIrecurDist (A7) also works well with less tight constraints.

In most practical applications, e. g., in mobile phones, the sound reproduction system imposes a certain constant maximum audio power, which the output signal may not exceed. Also in this case, OptSIIrecurDist (A7) is suitable for use if the threshold distance $D_\triangle$ is adjusted to a higher value. It has very similar instrumental measure scores than the plain OptSIIrecur (A4) without limitation of the disturbance spectrum level, but sounds more natural and thus yields a better listening experience especially for extreme bandpass noises, like the car interior noise.

Chapter 5

Loudspeaker Distortions and Protection

In mobile phones, usually no high-end loudspeakers are used but micro-loudspeakers with comparably low characteristic sensitivities[1] and limited capabilities for power handling, which then must be pushed to their limits to produce the needed sound pressure levels. These devices have basically two failure modes which can be caused by playing an audio signal and must therefore be controlled by a preceding block in the processing chain (Hsu & Poornima 2001):

1. the voice coil can break due to overheating caused by a too high electric current and
2. the excursion of moving parts including the diaphragm becomes too large. In this case, the suspension may tear or the voice coil may be forced out of the magnetic gap, which both destroys the loudspeaker. In a less dramatic case, the voice coil hits the back plate, which still causes acoustical distortions.

Concerning the first threat of overheating, the manufacturer of a loudspeaker usually specifies (see, e. g., Knowles 2011; NXP 2010a,b) that the loudspeaker can stand a specific noise signal with crest factor 2

- at maximum short-term power $\mathfrak{P}_x^{\text{short}}$ for, e. g., 60 cycles of 1 second on and 1 minute off,
- at maximum long-term power $\mathfrak{P}_x^{\text{long}}$ for, e. g., 10 cycles of 1 minute on and 2 minutes off (not always given), and
- at maximum continuous power $\mathfrak{P}_x^{\text{cont}}$ for, e. g., 500 hours.

If the change of voice coil resistance with temperature (Hsu & Poornima 2000) is neglected, the electric power in the voice coil can be assumed to be a linear function of the audio signal power. In this case, an overheating of the voice coil can be prevented by a fullband time-domain limiter as described in Section 5.3.3 as well as by NELE with a constant audio power constraint as described in Chapter 4.

Some loudspeaker specifications furthermore specify the maximum linear excursion and the excursion of the membrane for a single sine signal as a function of the frequency. Unfortunately, these specifications cannot directly be used for the design of loudspeaker protection algorithms. Therefore, a measurement campaign which is described in Section 5.1 was performed to derive a simple model suitable for

[1] Sound pressure level measured at, e. g., 1 m distance with an input of, e. g., 1 W.

algorithm design. Section 5.2 presents the results of this measurement campaign, which are used to derive a loudspeaker protection algorithm in Section 5.3.

5.1 Measurement Procedure

Usually two types of transducers[2] are used in mobile phones:

- The *speaker* is commonly placed on the back side of the phone and is used for hands-free telephony, music playback, and ring tones. It operates towards an open front volume, i. e., free field, and against a sealed back cavity.
- The *receiver*, in contrast, is embedded in the front side above the display and is used in the usual handset telephone situation, i. e., it is held right on the ear to receive the far-end speech signal. It has a specific front volume with one or more sound port holes which are, in operation, covered by the ear of the listener. The back volume is "semi-open" through various holes and slits in the mobile phone, e. g., at keypad or data connection jacks.

A comparison of the characteristics of speaker and receiver is given in Table 5.1.

	Speaker	Receiver
Use cases	hands-free, music playback, ring tone	handset
Main dimension	"bigger"	"smaller"
Front volume	open	defined volume with sound port hole(s), covered by ear
Back volume	closed, typ. 0.75–2 cm^3	semi-open
Typ. rated impedance	8 Ω	32 Ω
Typ. resonance frequency (1 cm^3 back cavity)	800–950 Hz	700-750 Hz
Typ. max. short-term power $\mathfrak{P}_x^{\text{short}}$	700–1000 mW	75–100 mW
Typ. max. continuous power $\mathfrak{P}_x^{\text{cont}}$	300–500 mW	40–50 mW
Typ. characteristic sensitivity (max. cont. power, 10 cm dist.)	84–90 $\mathrm{dB_{SPL}}$ (average 2–5 kHz)	74 $\mathrm{dB_{SPL}}$ (average 1–3 kHz)

Table 5.1: Comparison of the characteristics of speaker and receiver

[2] To avoid confusion between "speaker" and "loudspeaker", the term "transducer" is used in the following to denote the general class of micro-loudspeakers containing both, speakers and receivers.

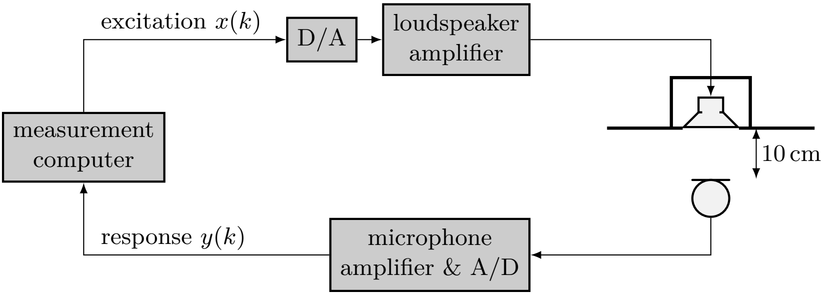


Figure 5.1: Setup for measurement with speaker.

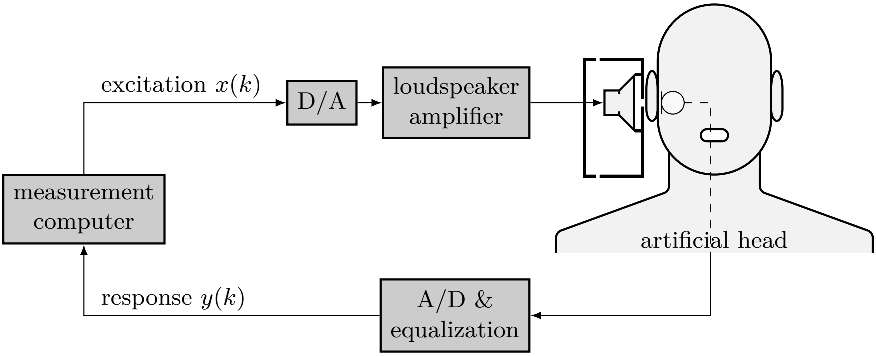


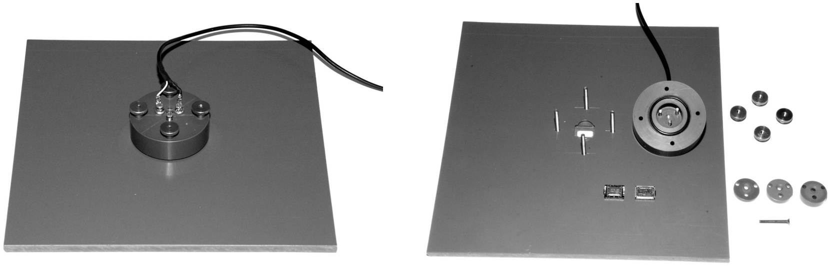
Figure 5.2: Setup for measurement with receiver.

5.1.1 Measurement Setup

Resulting from the different configurations, i. e., front and back volumes, different setups for acoustical measurement with speakers and receivers are needed, which are depicted in Figures 5.1 and 5.2. In both cases, the excitation signals, which are described below, are created in the measurement computer, digital-analog converted using a *Big DAADI* device by *Tracer Technologies Inc.*, amplified with custom built loudspeaker amplifiers, and played back with the transducer under test. After recording with a microphone and analog-digital conversion, the response signal is evaluated in the measurement computer.

Setup for Measurements with Speaker

The custom built amplifier used in the setup for speakers is calibrated such that a digital full-scale sine signal, i. e., a sine signal with $0\,\mathrm{dB_{FS}}$, yields an electric power of $1000\,\mathrm{mW}$ at the speaker, which is the largest maximum short-term power $\mathfrak{P}_x^{\mathrm{short}}$ of all tested speakers. This calibration allows to set all required power settings



(a) Mounted back cavity. **(b)** Dismounted back cavity. Disks are used to change size of back cavity.

Figure 5.3: Housing for the *NXP 13.6x9.6x2.9 speaker*.

via script in the measurement computer without any tuning at the hardware. The *Big DAADI*/amplifier bundle has a total harmonic distortion (THD) of about 0.03 % at $0\,\mathrm{dB_{FS}}$ signal power.

The speaker itself is placed in a custom built housing which is depicted in Figure 5.3 for the *NXP 13.6x9.6x2.9 speaker*. It consists of a 20 cm × 20 cm front plate with a hole that exposes the whole speaker membrane and a closed back cavity, which is tightened with four nuts. This front plate size was chosen for practical reasons after pilot tests showed that its influence on the measurement is negligible. The size of the back cavity can be chosen with small disks to approximately[3] $0.75\,\mathrm{cm}^3$, $1.0\,\mathrm{cm}^3$, $1.5\,\mathrm{cm}^3$, and $2.0\,\mathrm{cm}^3$. Three screws in the back cavity hold the speaker in place and serve also as electric contact.

The measurement microphone, a *Beyerdynamic MM 1*, is placed perpendicular to the front plate in 10 cm distance and attached to the microphone amplifier *RME OctaMic II* with activated 80 Hz highpass filter. The measurement chain of microphone and microphone amplifier including analog-digital conversion is calibrated with a *Voltcraft Schallpegelkalibrator 326*.

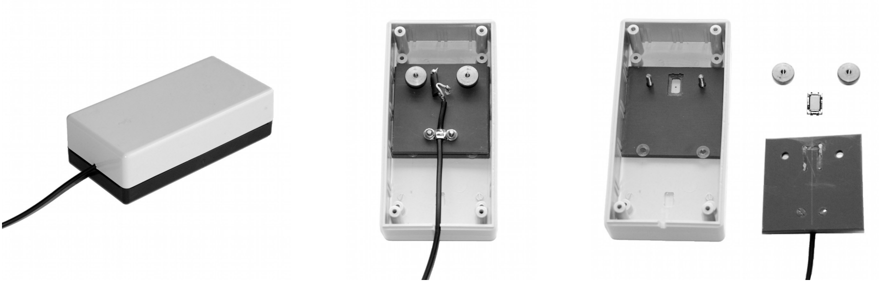
All measurements with the *NXP 13.6x9.6x2.9 speaker* were performed in the anechoic chamber at the Institute for Communications Engineering at the RWTH Aachen University.

Setup for Measurements with Receiver

The custom built amplifier for the receiver is tuned to yield the maximum short-term power $\mathfrak{P}_x^{\text{short}}$ of the tested receiver, in this case 75 mW, when loaded with a $0\,\mathrm{dB_{FS}}$ sine signal.

The receiver itself is placed in a custom built mockup which is depicted in Figure 5.4 for the *NXP 8x12x2 receiver*. It consists of a 5.9 cm × 11.9 cm × 4.6 cm

[3] The size of the back cavity is calculated using some minor approximations. For the sake of clarity, the term “approximately” is nevertheless omitted in the following.



(a) Mounted mockup with slit along the perimeter.

(b) Inner side of front case.

(c) Dismounted front case.

Figure 5.4: Housing for the *NXP 8x12x2 receiver*.

box with a slit along the perimeter. The front case has a hollow with a front cavity and a sound port hole. Using small spacer and a back plate, the receiver is connected and held in place in the hollow without closing the back volume.

For data acquisition, *HEAD acoustics*' artificial head measurement system *HMS II.3* with ear simulator and anatomically shaped pinna according to (ITU-T P.57 2009, Type 3.3) is used. The mockup is connected to the *HMS II.3* using the handset positioner *HHP III* clamped with the 5° positioning jig and spatially positioned at ear reference point (ERP) and 0° rotation in all three axes.[4] The measurement frontend *MFE VI* performs the analog-digital conversion as well as the binaural equalization with filter setting "linear" and a first-order highpass filter with 180 Hz cut-off frequency.

All measurements with the *NXP 8x12x2 receiver* were performed in the sound booth at the Institute of Communication Systems and Data Processing at the RWTH Aachen University.

5.1.2 Measures

At low amplitudes, i. e., if the excursion of the diaphragm is below the maximum linear excursion, the transducer is expected to behave approximately linearly. That is, the response of the transducer should be a linearly filtered version of the excitation. If the excursion exceeds this limit, non-linear, mostly harmonic distortions occur.

A commonly used and well known measure for harmonic distortions is the *total harmonic distortion* (THD). The THD, however, is only defined for single sine waves as input. Therefore, two other measures, the *total intermodulation distortion* (TID) for mixtures of two sines and the *total non-linear distortion* (TND) for bandpass signals, are used in this chapter.

[4] A similar setup with *HMS II.4* and *HHP II* is depicted in Figure 2.4.

Total Harmonic Distortion (THD)

If the device under test exhibits a non-linear behaviour and a single pure sine wave with frequency f is used as excitation for the device, the response signal will contain additional components at the harmonics $\eta \cdot f$, $\eta \in \mathbb{N}$, of the fundamental frequency of the excitation signal. The THD is commonly defined as the square root of the ratio of the power of all higher harmonic components of the response signal to the power at the fundamental frequency:

$$THD(f) = 100\,\% \cdot \sqrt{\frac{P_{y,2} + P_{y,3} + P_{y,4} + \dots}{P_{y,1}}}\,, \tag{5.1}$$

with $P_{y,\eta}$ denoting the power of the η-th harmonic of the response signal.

With this definition, $THD(f) = 0\,\%$ means that the device under test is perfectly linear at frequency f, whereas with $THD(f) = 100\,\%$ all harmonics together have the same power as the fundamental frequency. While speech has a strongly harmonic structure by itself, which is "just" intensified by harmonic distortions, music signals are much more sensitive to non-linear transformations. In general, a THD up to 10 % is usually acceptable for micro-loudspeakers (Behler 2010).

The THD can also be determined with the DFT using digital signal processing, if the excitation frequency f is the center frequency of a DFT bin and if at least one second of the response signal $y(k)$ is recorded at a sufficiently high sampling rate f_s to cover all relevant harmonics. The THD can then be calculated as

$$THD_{\mathfrak{P}_x}(f) = 100\,\% \cdot \sqrt{\frac{\sum\limits_{\eta=2}^{\min\left\{\left\lfloor \frac{f_\mathrm{s}}{2f} \right\rfloor,\, 11\right\}} |\mathcal{Y}_{\eta\cdot\mu}|^2}{|\mathcal{Y}_\mu|^2}} \qquad \text{with } \mu = \frac{f}{f_\mathrm{s}} \cdot M \in \mathbb{N} \tag{5.2}$$

and the DFT coefficients $\mathcal{Y}_\mu$ of the response signal $y(k)$ with a rectangular window. The subscript $\mathfrak{P}_x$ denotes the electric power of the excitation signal.

Preliminary tests have shown that in most cases the powers of the eighth and higher harmonics are more than 50 dB below the overall response power and that they contain more noise than response signal. Therefore only the first ten overtones are considered when calculating the THD, which is why the sum in (5.2) is restricted to $\eta \leq 11$.

For stationary signals and environments as well as uncorrelated noise sources, the SNR can be improved by averaging over several repetitions of the response signal. In this case, the SNR increases by 3 dB for every doubling of the duration.

For the measurements in this chapter, the sampling rate $f_\mathrm{s} = 48\,\mathrm{kHz}$, the DFT size $M = {}^{f_\mathrm{s}}\!/_{\mathrm{Hz}} = 48000$, and "integer" excitation frequencies ${}^{f}\!/_{\mathrm{Hz}} \in \mathbb{N}$ were chosen.

Total Intermodulation Distortion (TID)

If a non-linear device under test is excited with a mixture of two sine waves with frequencies f_1 and $f_2 \neq f_1$, the response signal will not only contain the harmonic components at $\eta_1 \cdot f_1$ and $\eta_2 \cdot f_2$ but in general intermodulation products at the frequencies $\eta_1 \cdot f_1 + \eta_2 \cdot f_2$ with $\eta_1, \eta_2 \in \mathbb{Z}$.

Corresponding to the THD, the *total intermodulation distortion* (TID) is defined as the amplitude ratio of the power of all intermodulation components of the response signal to the average power at both fundamental frequencies. Analogously to the THD, the TID can be calculated using the DFT for excitation frequencies f_1 and f_2 which are center frequencies of DFT bins:

$$TID_{\mathfrak{P}_{x_1}+\mathfrak{P}_{x_2}}(f_1, f_2) = 100\,\% \cdot \sqrt{\frac{\sum\limits_{\mu \in \mathbb{M}_{f_1,f_2}} |\mathcal{Y}_\mu|^2}{\frac{|\mathcal{Y}_{\mu_1}|^2 + |\mathcal{Y}_{\mu_2}|^2}{2}}} \tag{5.3}$$

with $\mu_1 = \frac{f_1}{f_\mathrm{s}} \cdot M \in \mathbb{N}_0$, $\mu_2 = \frac{f_2}{f_\mathrm{s}} \cdot M \in \mathbb{N}_0$, and the set of intermodulation indices

$$\begin{aligned} \mathbb{M}_{f_1,f_2} = \Big\{ \eta_1 \cdot \mu_1 + \eta_2 \cdot \mu_2 \;\Big|\; & \eta_1, \eta_2 \in \mathbb{Z} \wedge |\eta_1| \le 10 \wedge |\eta_2| \le 10 \\ & \wedge \eta_1 \cdot f_1 + \eta_2 \cdot f_2 \neq f_1 \wedge \eta_1 \cdot f_1 + \eta_2 \cdot f_2 \neq f_2 \\ & \wedge 60\,\mathrm{Hz} \le \eta_1 \cdot f_1 + \eta_2 \cdot f_2 < \frac{f_\mathrm{s}}{2} \Big\}. \end{aligned} \tag{5.4}$$

The subscript $\mathfrak{P}_{x_1}+\mathfrak{P}_{x_2}$ denotes the electric power of the two excitation sine waves. With the same reasoning as above, the number of considered intermodulation components is restricted to $|\eta_1| \le 10$, $|\eta_2| \le 10$. Furthermore, all components below 60 Hz are discarded as the response signal cannot contain useful signal in this range due to the active highpass filter in the microphone pre-amplifier and the *MFE VI*. As for the THD, the SNR of the TID can be improved by averaging over several seconds of the response signal.

If f_1 and f_2 have large common divisors as, e. g., $f_1 = \frac{3}{2} f_2$, relevant intermodulation products of low order fall on one of the fundamental frequencies ($3f_2 - f_1 = f_1$ in the example) and are thus missed by (5.3). Due to that, f_1 and f_2 should be chosen such that the greatest common divisor of f_1 and f_2 is as small as possible, preferably one. For this reason, “strange” second frequencies f_2 were used during the measurements, which are only close to multiples of 100 Hz (see Figures 5.9 and 5.14).

With the TID, the distortion power is evaluated in relation to the average of the powers of the fundamental frequencies, which gives the most meaningful results if $|\mathcal{Y}_{f_1}|^2 = |\mathcal{Y}_{f_2}|^2$. Otherwise, it could be the case that the distortion is mostly caused by one of the two frequency components but the evaluation is dominated by the other one. For this reason, the two sine components in the excitation signal are equalized as explained in Section 5.1.3, which will be denoted by the subscript $\mathcal{H}^2_{f_1}\mathfrak{P}_x{+}\mathcal{H}^2_{f_2}\mathfrak{P}_x$.

Total Non-Linear Distortion (TND)

The third measure, the *total non-linear distortion* (TND), is related to the so-called *Rauschklirrmessung* (e. g., Kammeyer 2004), but uses a “reverse” approach. Instead of a broadband excitation with a narrow gap caused by a bandstop filter, a narrow-band noise signal with center frequency f_c and bandwidth $f_\triangle$ is used as excitation.

The excitation is derived from a Gaussian white noise signal using a Chebyshev Type I bandpass filter[5] with the passband $f_\mathrm{c} - \frac{f_\triangle}{2}$ to $f_\mathrm{c} + \frac{f_\triangle}{2}$ and an attenuation of 60 dB below $f_\mathrm{c} - f_\triangle$ and above $f_\mathrm{c} + f_\triangle$.

Using this excitation, intermodulation products occur not at discrete frequencies but (as all other non-linear distortions) everywhere in the spectrum. Corresponding to the THD and TID, the TND is defined as the amplitude ratio of the distortion power to the power of the response signal in the excitation band, which can be calculated using a DFT as

$$\mathit{TND}_{\mathfrak{P}_x}(f_\mathrm{c}, f_\triangle) = 100\,\% \cdot \sqrt{\frac{\sum\limits_{\mu=0}^{\mu_\mathrm{f}-1} |\mathcal{Y}_\mu|^2 + \sum\limits_{\mu=\mu_\mathrm{l}+1}^{M/2} |\mathcal{Y}_\mu|^2}{\sum\limits_{\mu=\mu_\mathrm{f}}^{\mu_\mathrm{l}} |\mathcal{Y}_\mu|^2}} \tag{5.5}$$

with $\mu_\mathrm{f} = \left\lceil (f_\mathrm{c} - f_\triangle) \cdot \frac{M}{f_\mathrm{s}} \right\rceil$ and $\mu_\mathrm{l} = \left\lfloor (f_\mathrm{c} + f_\triangle) \cdot \frac{M}{f_\mathrm{s}} \right\rfloor$ being the first and last DFT bin of the excitation band, respectively. In order to account for the finite slope of the bandpass filter and the spectral leakage of the DFT, this measure assumes that all response components within the frequency range $f_\mathrm{c} - f_\triangle$ to $f_\mathrm{c} + f_\triangle$ are useful whereas everything outside is distortion.[6] This, of course, is only an approximation as intermodulation products might also fall in this range, but it turns out to still give reasonable results.

Compared to the TID, the TND has an increased sensitivity to background noise since the whole spectrum and not only some discrete frequencies are considered. Besides discarding all components of the response signal below 60 Hz and averaging over several seconds of the response signal, the background noise problem is tackled by a very basic form of spectral subtraction which is applied to $\mathcal{Y}_\mu$ before calculation of the TND. The noise estimate $\hat{\mathcal{N}_\mu}$ is measured and averaged beforehand the same way as $\mathcal{Y}_\mu$ but with silenced excitation signal, assuming that the background noise is sufficiently stationary during the measurement. The PSD of the noise-free microphone signal is estimated as

$$\hat{\Phi}_{yy,\mu} = \max\left\{ |\mathcal{Y}_\mu|^2 - |\hat{\mathcal{N}_\mu}|^2,\, 0 \right\}. \tag{5.6}$$

[5] The filter was designed with the *Matlab Signal Processing Toolbox*.

[6] The signal power which is leaked by the bandpass and the DFT outside this frequency range is about −70 dB of the useful response power, equivalent to a TND of less than 0.05 %.

This leads to the final implementation

$$TND_{\mathcal{H}^2\mathfrak{P}_x}(f_\mathrm{c}, f_\Delta) = 100\,\% \cdot \sqrt{\frac{\sum\limits_{\mu=\left\lceil 60\,\mathrm{Hz}\cdot\frac{M}{f_\mathrm{s}}\right\rceil}^{\mu_\mathrm{f}-1} \hat{\varPhi}_{yy,\mu} + \sum\limits_{\mu=\mu_\mathrm{l}+1}^{M/2} \hat{\varPhi}_{yy,\mu}}{\sum\limits_{\mu=\mu_\mathrm{f}}^{\mu_\mathrm{l}} \hat{\varPhi}_{yy,\mu}}}\,. \tag{5.7}$$

Again, the TND gives the most meaningful results, if the signal radiated from the transducer is spectrally flat. Therefore, the excitation signal is equalized before digital-analog conversion with a time-domain equalization filter as described in Section 5.1.3. This is indicated by the subscript $\mathcal{H}^2\mathfrak{P}_x$.

5.1.3 Equalization

As described above, the TID and TND measures require that the radiated signal is approximately spectrally flat. This is achieved with one equalization transfer function $\mathcal{H}_f$ for each transducer, which equalizes the overall frequency response of the transducer.

The derivation of $\mathcal{H}_f$ is based on the DFT coefficients $\mathcal{Y}_{\mathfrak{P}_x,f}$ of the microphone signal $y(k)$ measured in response to the sine excitation signals $x(k)$ with

- frequencies $100\,\mathrm{Hz} \le f \le \frac{f_\mathrm{s}}{4} = 12\,\mathrm{kHz}$ in up to 1/16-th octave steps and
- excitation powers $-40\,\mathrm{dB} \le 10\log\left\{\frac{\mathfrak{P}_x}{\mathfrak{P}_x^\mathrm{short}}\right\} \le -15\,\mathrm{dB}$ in 0.5 dB steps.

Within this wide power range, the normalized magnitude response of the transducer is roughly constant and neither influenced by the measurement noise at low power levels nor by a non-linear behaviour at high power levels.

The equalization transfer function $\mathcal{H}_f$ is then calculated at the measurement frequencies f as the inverse of the average in decibel over all excitation power levels of the response power at frequency f, normalized to the response power at 1 kHz:

$$20\log\{\mathcal{H}_f\} = -\operatorname*{mean}_{\mathfrak{P}_x} 10\log\left\{\frac{|\mathcal{Y}_{\mathfrak{P}_x,f}|^2}{|\mathcal{Y}_{\mathfrak{P}_x,1\,\mathrm{kHz}}|^2}\right\}. \tag{5.8}$$

In case of the TND, a time-domain equalization filter is needed. For this purpose, the equalization transfer function is linearly interpolated in decibel between the measured frequencies as well as extrapolated with a 10 dB/octave decade towards lower frequencies and flat towards higher frequencies. The filter coefficients are finally calculated using the inverse DFT and truncated to a linear-phase filter of degree 1000, which is large enough to achieve negligible truncation errors.

5.2 Measurement Results

5.2.1 Measurement Results of a Typical Speaker

Speakers are commonly placed on the back side of the phone and are used for hands-free telephony, music playback, and ring tones, cf. Table 5.1.

This section presents the results of measurements with the *13.6x9.6x2.9 speaker*[7] series built by *NXP Semiconductors*. According to its specification (Knowles 2011)[8], this speaker can withstand the maximum short-term power $\mathfrak{P}_x^{\text{short}} = 700\,\text{mW}$ for 1 second and the maximum continuous power $\mathfrak{P}_x^{\text{cont}} = 300\,\text{mW}$ for 500 hours.

Magnitude Response

Figure 5.5 plots the measured "linear" magnitude response of the *NXP 13.6x9.6x2.9 speaker* with different back cavities. In detail, it shows the emitted SPL at excitation frequency in 10 cm distance for a sine excitation with 250 mW power and frequencies between 200 Hz and 22.6 kHz in 1/16-th octave steps below 1 kHz, 1/8-th octave steps between 1 kHz and 2 kHz, and 1/4-th octave steps above 2 kHz. All measurements were conducted with a sampling rate of $f_\text{s} = 48\,\text{kHz}$.

The specification (Knowles 2011) states a "tolerance window" for $1\,\text{cm}^3$ back cavity, which is also depicted in Figure 5.5. It can be seen, that the measured

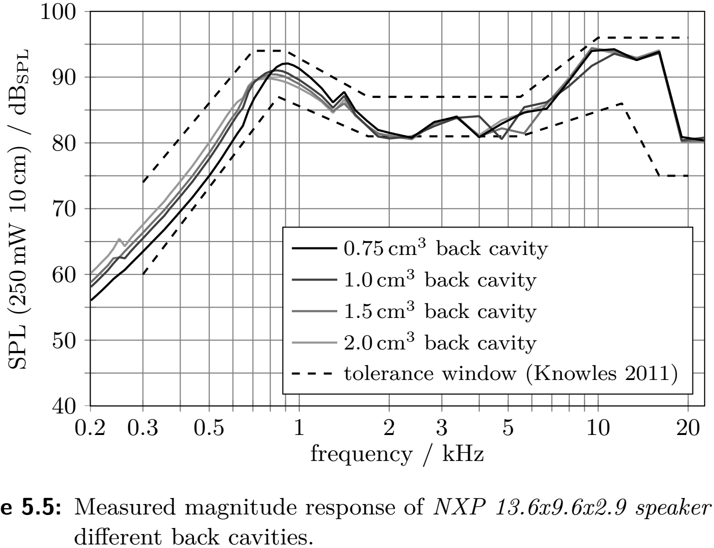


Figure 5.5: Measured magnitude response of *NXP 13.6x9.6x2.9 speaker* with different back cavities.

[7] The numbers in the product name indicate the main dimensions of the speaker in mm.

[8] Knowles Electronics acquired NXP's Sound Solutions Business in July 2011.

response fits in the tolerance window, which gives an indication that the housing presented in Section 5.1.1 works well.

The magnitude response of the speaker exhibits a strong resonance rise between 700 Hz and 900 Hz with a decay of about 17 dB per octave towards lower frequencies. As to be expected, the resonance frequency increases with smaller back cavity. The magnitude response stays about 10 dB below resonance peak for frequencies between 1.7 kHz and 5.5 kHz and increases afterwards again. Besides the resonance frequency, the general shape of the magnitude response is independent of the size of the back cavity; only above 3.5 kHz the fluctuations of the magnitude responses deviate slightly.

Further measurements showed the reproducibility of the results with a deviation of less than 2 dB below 5 kHz and less than 3 dB above.

Total Harmonic Distortion (THD)

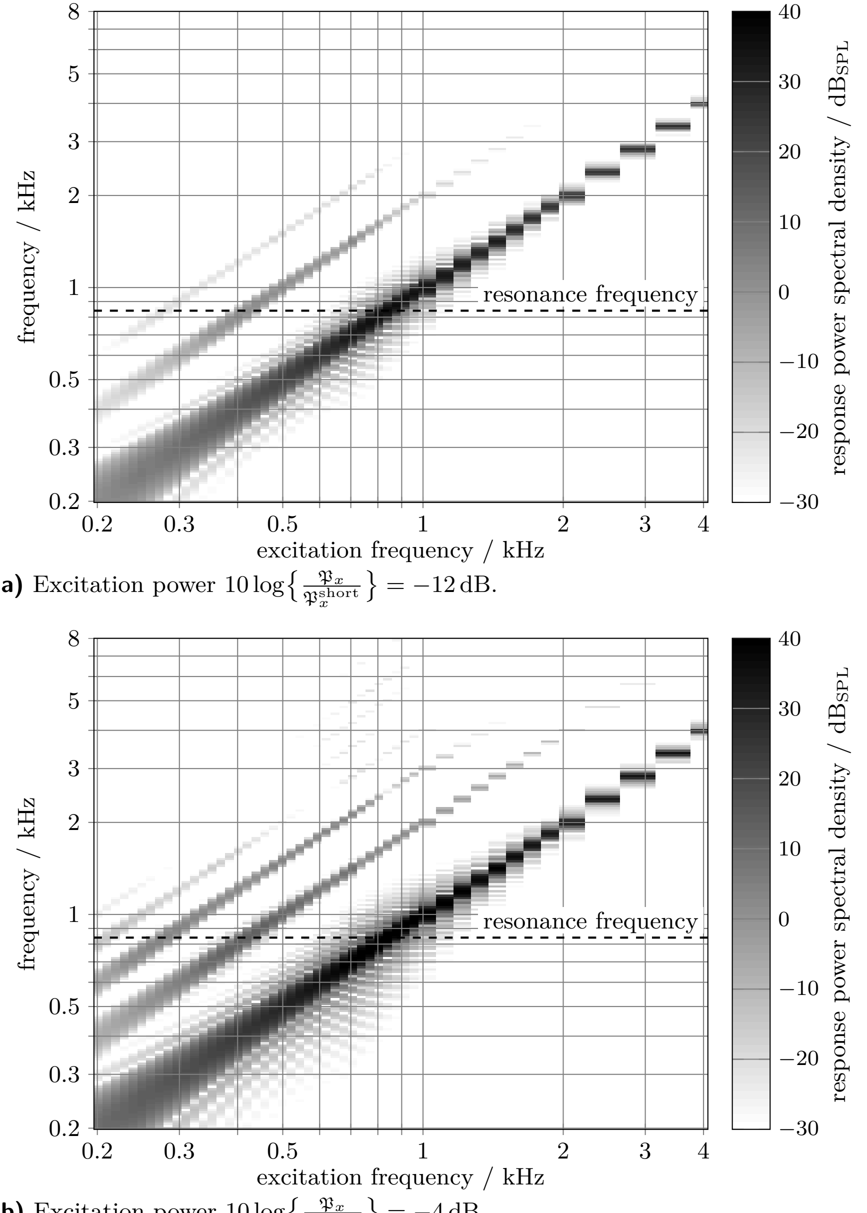
Figure 5.6 depicts the spectral response power without equalization of the *NXP 13.6x9.6x2.9 speaker* as a spectrogram over the frequency of the sine excitation for two excitation powers $10\log\left\{\frac{\mathfrak{P}_x}{\mathfrak{P}_x^{\text{short}}}\right\} = -12\,\text{dB}$ and $10\log\left\{\frac{\mathfrak{P}_x}{\mathfrak{P}_x^{\text{short}}}\right\} = -4\,\text{dB}$. The linear response to the fundamental frequency can be seen as the main diagonal and each harmonic distortion as a parallel line above it.

For the lower excitation power level of −12 dB, only the second harmonic is substantially stimulated and has its maximum around the resonance frequency of 840 Hz, i. e., for an excitation frequency of half the resonance frequency. At a high excitation power level of −4 dB, the third harmonic is also significantly excited, again with a maximum around resonance frequency. In both cases, the harmonic distortions become insignificant (in relation to the linear response) for excitation frequencies above 800 Hz to 900 Hz.

In Figure 5.7, the measured THD without equalization of the *NXP 13.6x9.6x2.9 speaker* with different back cavities is plotted over excitation power and frequency. It should be noted, that the very slight increase in THD in the lower left corner for low excitation power and low frequencies is not a characteristic of the speaker but caused by a decreased SNR due to a very low response power.

In accordance with the above analysis of Figure 5.6, the THD is largest just below half the resonance frequency, as the second harmonic is then around resonance frequency. At very high excitation power levels, the third harmonic is also stimulated, leading to an increase of THD around a third of the resonance frequency. As can further be seen, that the THD is basically negligibly small above 700 Hz. In general, the THD is higher for larger back cavities, which is again to be expected as the excursion of the membrane also increases with larger back cavities.

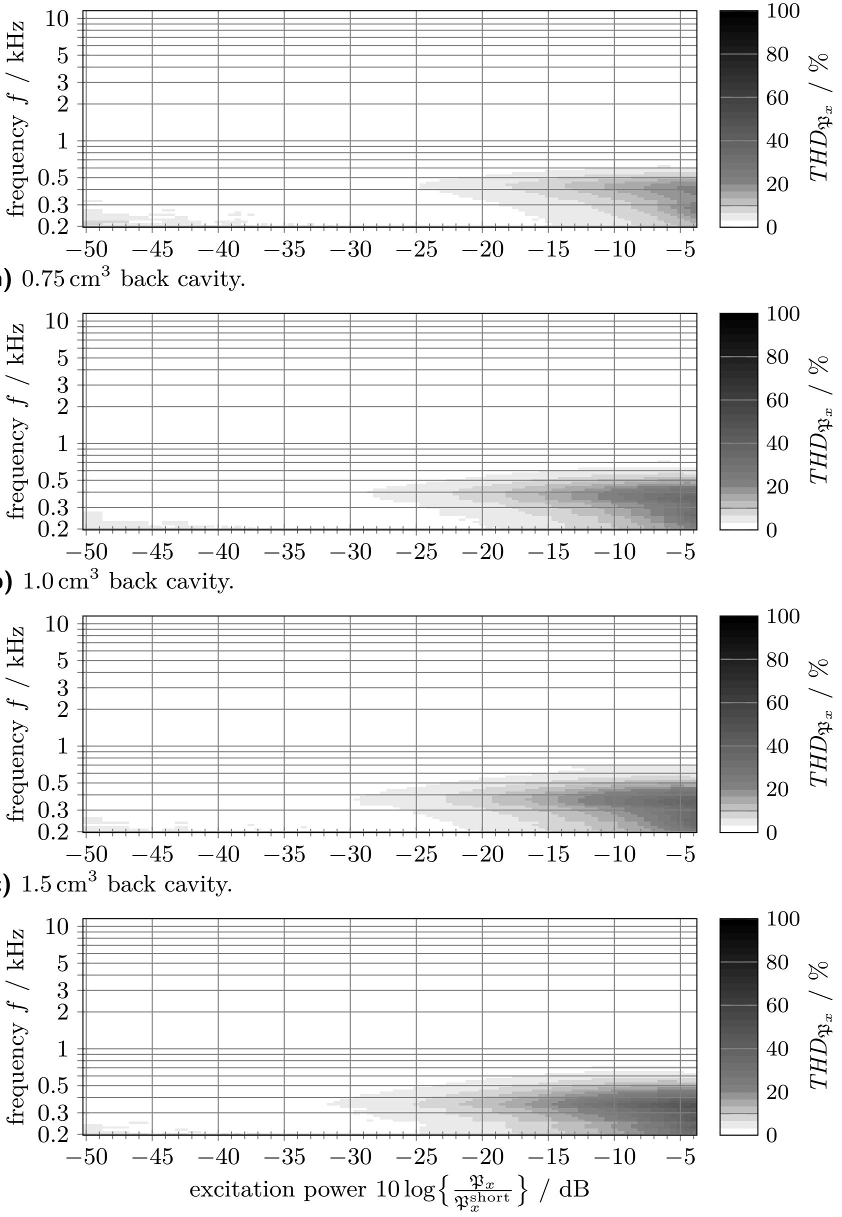
In the implementation of a mobile phone, the transfer function of the transducer and its housing is usually equalized, which, in turn, must be considered to successfully minimize distortions and protect the transducer. For speech telephony, the equalization should achieve the transfer characteristic which is demanded, e. g., by (3GPP TS 26.131 2011) or (ITU-T G.712 2001), whereas for music playback it



(a) Excitation power $10\log\left\{\frac{\mathfrak{P}_x}{\mathfrak{P}_x^{\text{short}}}\right\} = -12\,\text{dB}$.

(b) Excitation power $10\log\left\{\frac{\mathfrak{P}_x}{\mathfrak{P}_x^{\text{short}}}\right\} = -4\,\text{dB}$.

Figure 5.6: Spectrogram without equalization of *NXP 13.6x9.6x2.9 speaker* with $1.0\,\text{cm}^3$ back cavities for different excitation powers.



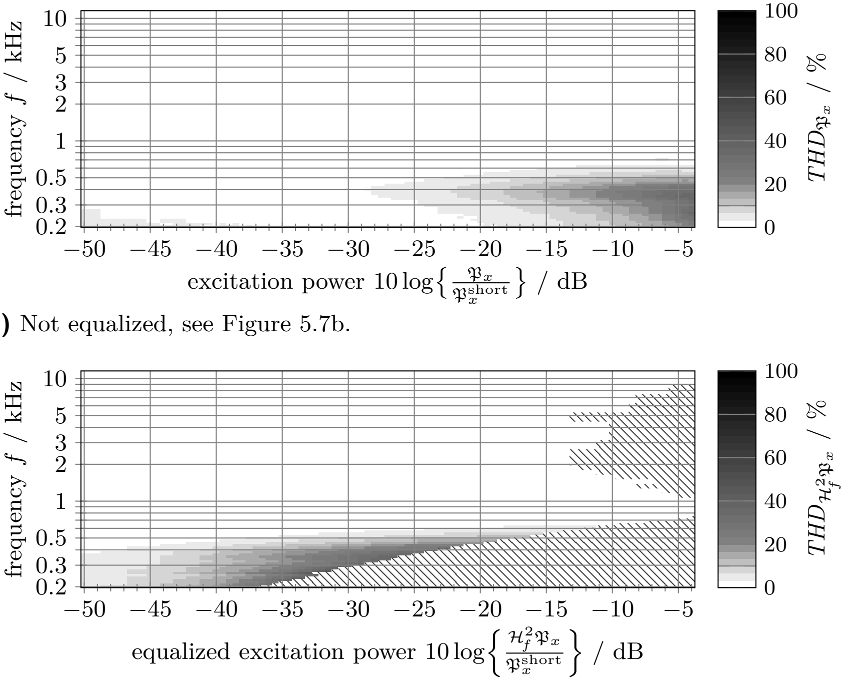
(a) $0.75\,\text{cm}^3$ back cavity.

(b) $1.0\,\text{cm}^3$ back cavity.

(c) $1.5\,\text{cm}^3$ back cavity.

(d) $2.0\,\text{cm}^3$ back cavity.

Figure 5.7: Total harmonic distortion (THD) without equalization of *NXP 13.6x9.6x2.9 speaker* for different back cavities.



(a) Not equalized, see Figure 5.7b.

(b) Equalized according to Section 5.1.3.

Figure 5.8: Influence of equalization on total harmonic distortion (THD) of *NXP 13.6x9.6x2.9 speaker* with 1.0 cm^3 back cavity.

“just” yields a flatter magnitude response.

Therefore, in a second measurement, the excitation signal is equalized, i. e., weighted with the inverse transfer function, as described in Section 5.1.3. Figure 5.8b shows the result of this measurement with 1.0 cm^3 back cavity in terms of THD over equalized excitation power and frequency. As an effect of the equalization, excitation signals with the same equalized excitation power emit approximately the same SPL, but have different electrical signal powers for different frequencies. This means that for some frequencies the maximum continuous power is reached at lower radiated sound pressure levels (SPLs) than for others, leading to the hatched areas in Figure 5.8b.

The equalization, however, makes clear that distortions are a more severe problem at low frequencies below the cut-off frequency of the magnitude response than to be expected from Figure 5.7. Essentially, all frequency components below 500 Hz to 600 Hz must be attenuated to about 50 dB$_{\mathrm{SPL}}$ if the THD should be below 10 %, whereas frequencies above 600 Hz can be played up to the maximum continuous power.

Total Intermodulation Distortion (TID)

The measured TID of the *NXP 13.6x9.6x2.9 speaker* with 1.0 cm^3 back cavity is plotted in Figure 5.9 for different frequencies over equalized excitation power. The ordinate shows the first excited frequency f_1, whereas the second excited frequency f_2 is given as a parameter. Both sine components are equalized.

A more thorough evaluation shows, that for the measured power range and $\min\{f_1, f_2\} \geq 400\,\mathrm{Hz}$, the approximation

$$TID_{\mathcal{H}^2_{f_1}\mathfrak{P}_x+\mathcal{H}^2_{f_2}\mathfrak{P}_x}(f_1, f_2) \approx THD_{\mathcal{H}^2_{f}\mathfrak{P}_x}(f_1) + THD_{\mathcal{H}^2_{f}\mathfrak{P}_x}(f_2) \tag{5.9}$$

holds reasonably well, i. e., the TID of two sines is approximately the sum of the individual THDs at the two frequencies. For lower frequencies, however, this approximation underestimates the true TID.

As $THD_{\mathcal{H}^2_{f}\mathfrak{P}_x}(f) \approx 0$ for $f \geq 800\,\mathrm{Hz}$, it follows for $\max\{f_1, f_2\} \geq 800\,\mathrm{Hz}$ that

$$TID_{\mathcal{H}^2_{f_1}\mathfrak{P}_x+\mathcal{H}^2_{f_2}\mathfrak{P}_x}(f_1, f_2) \approx THD_{\mathcal{H}^2_{f}\mathfrak{P}_x}\big(\min\{f_1, f_2\}\big)\,, \tag{5.10}$$

i. e., frequency components above 800 Hz do not contribute to the TID.

Total Non-Linear Distortion (TND)

Figure 5.10 shows the measured TND of the *NXP 13.6x9.6x2.9 speaker* with 1.0 cm^3 back cavity for different bandwidths over equalized excitation power. The ordinate shows the center frequency f_c of the excited band, whereas the bandwidth f_Δ is given in the caption. The final bandpass signal is equalized.

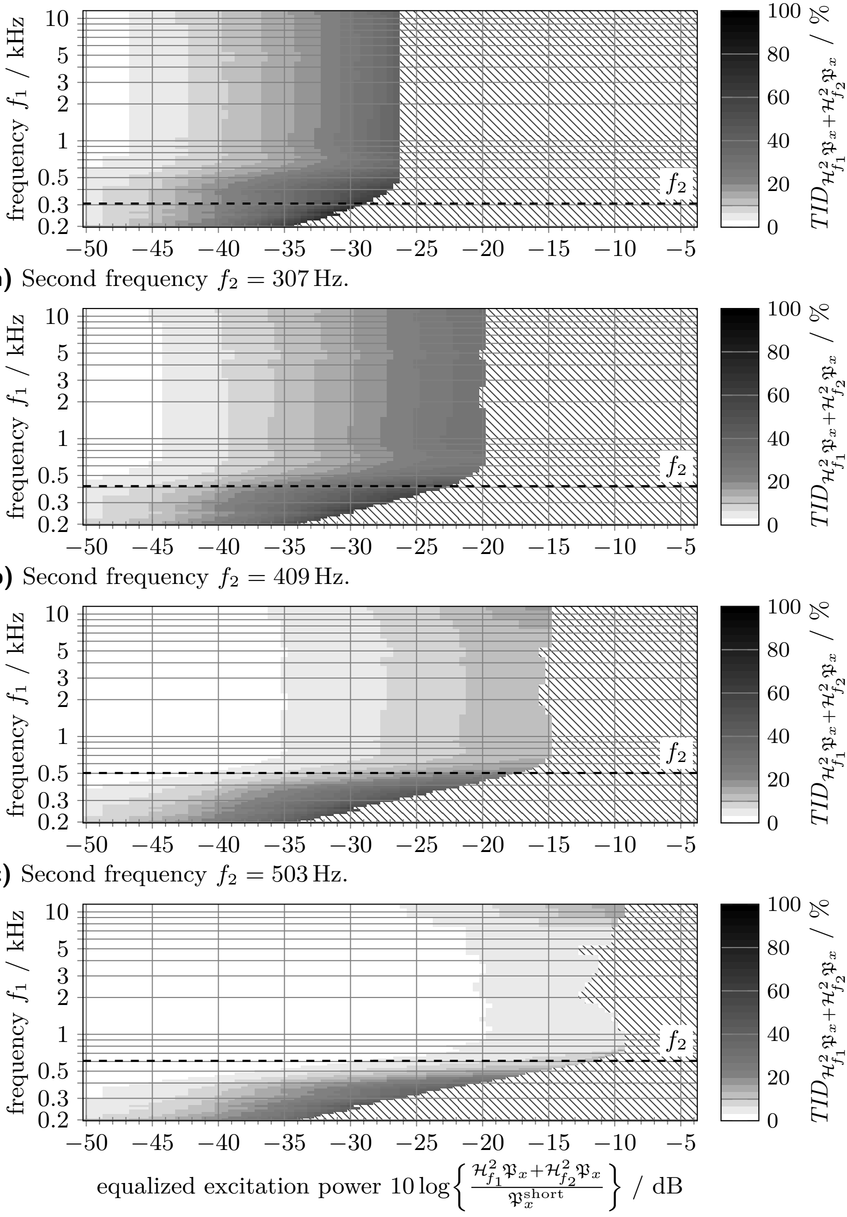
It should be noted, that although the speaker is embedded in a rubber gasket, it vibrates strongly enough at very high power levels to still hit its housing. This results in a strong rattle and high TND values as, e. g., at about 700 Hz above -10 dB equalized excitation power. However, since the speaker is required to be exchangeable during the measurements, it cannot be glued to the custom built measurement housing. This effective solution would in contrast be applicable for the manufacturer of a real device. Therefore, these high TND values are considered a measurement artifact.

As a result of the measurement, the measured TND becomes “smoother” for higher bandwidths, which can be expected as more frequency components with different harmonic behaviours are excited. Besides that, the general shape and behaviour is similar to that of the measured THD of Figure 5.8b.

However, the ratio between distortion power and linear response power is higher for the TND than for the THD. Interestingly, the average increase of 2.3 dB is about the same for all evaluated bandwidths and might be caused by the longer tail of the probability density function of the bandpass noise excitation compared to the single sine excitation. This leads to the approximation

$$TND_{\mathcal{H}^2\mathfrak{P}_x}(f_\mathrm{c}, f_\Delta) = 1.3 \cdot THD_{\mathcal{H}^2_{f}\mathfrak{P}_x}(f_\mathrm{c})\,, \tag{5.11}$$

which is depicted for comparison in Figure 5.10d.



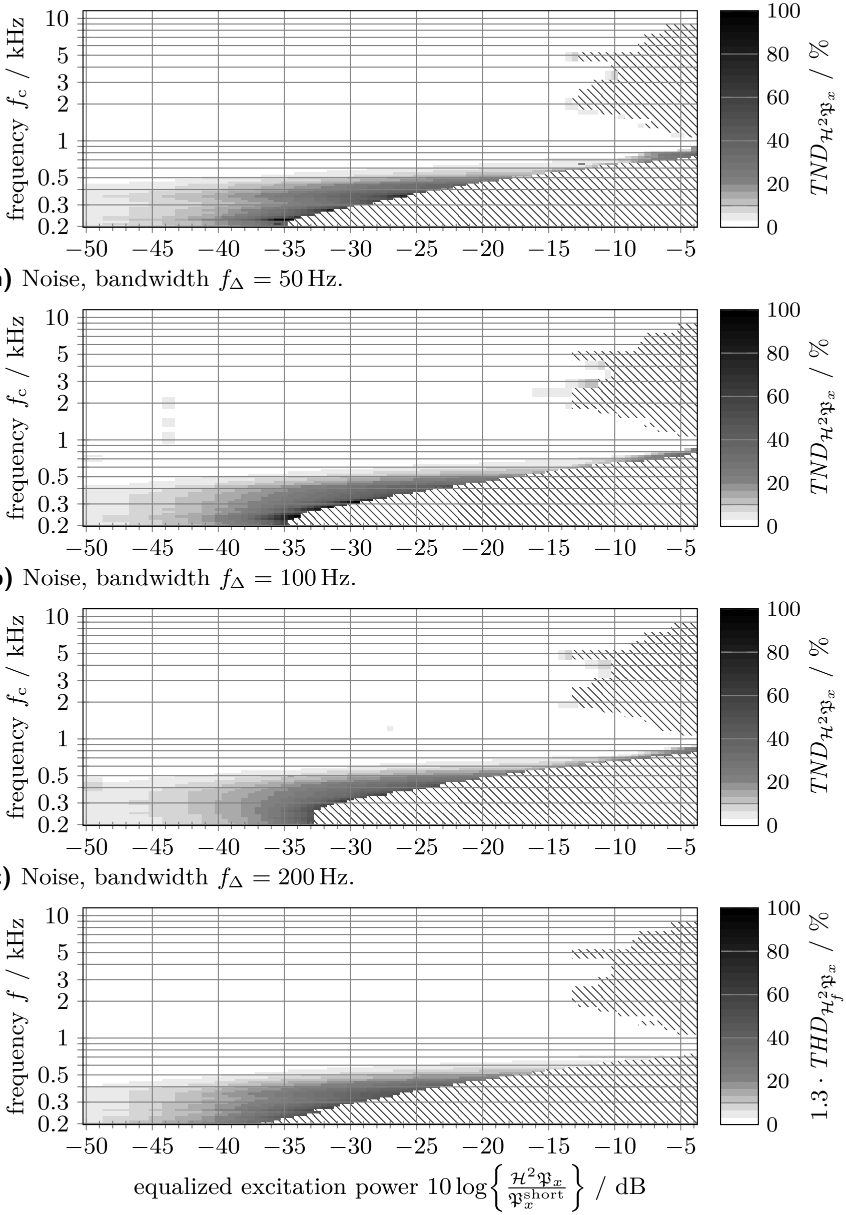
(a) Second frequency $f_2 = 307\,\text{Hz}$.

(b) Second frequency $f_2 = 409\,\text{Hz}$.

(c) Second frequency $f_2 = 503\,\text{Hz}$.

(d) Second frequency $f_2 = 607\,\text{Hz}$.

Figure 5.9: Total intermodulation distortion (TID) of *NXP 13.6x9.6x2.9 speaker* with $1.0\,\text{cm}^3$ back cavity for different second frequencies.



(a) Noise, bandwidth $f_\Delta = 50\,\text{Hz}$.

(b) Noise, bandwidth $f_\Delta = 100\,\text{Hz}$.

(c) Noise, bandwidth $f_\Delta = 200\,\text{Hz}$.

(d) Single sine, cf. Figure 5.8b, THD multiplied with 1.3.

Figure 5.10: Total non-linear distortion (TND) of *NXP 13.6x9.6x2.9 speaker* with $1.0\,\text{cm}^3$ back cavity for different bandwidths.

5.2.2 Measurement Results of a Typical Receiver

Receivers are embedded in the front side of the phone above the display and are used in the usual handset telephone situation, cf. Table 5.1.

This section presents the results of measurements with the *NXP 8x12x2 receiver*[9] series (NXP 2010b). This receiver can be considered "typical" for handset telephony and is specified to withstand the maximum short-term power $\mathfrak{P}_x^{\text{short}} = 75\,\text{mW}$ for 1 second and the maximum continuous power $\mathfrak{P}_x^{\text{cont}} = 40\,\text{mW}$ for 500 hours.

Magnitude Response

Figure 5.11 plots the measured "linear" magnitude response of the *NXP 8x12x2 receiver*, i. e., the emitted SPL at excitation frequency at the ear simulator of the *HMS II.3* for a sine excitation with 5 mW power and frequencies between 100 Hz and 22.6 kHz in 1/8-th octave steps below 200 Hz, 1/16-th octave steps between 200 Hz and 1 kHz, 1/8-th octave steps between 1 kHz and 2 kHz, and 1/4-th octave steps above 2 kHz. The handset receiving sensitivity mask for narrow-band transmission as given in (3GPP TS 26.131 2011) is also depicted in Figure 5.11 even though the testing conditions do not match perfectly.

The magnitude response of the receiver is rather flat between 350 Hz and 3.5 kHz with a resonance rise around 1.4 kHz and a drop between 1.8 kHz and 2.8 kHz. The response decreases towards lower frequencies with about 14 dB per octave and towards higher frequencies with about 16 dB per octave.

As for the speaker, these results are reproducible with the same as well as with a different *NXP 8x12x2 receiver* with a deviation of less than 2.5 dB.

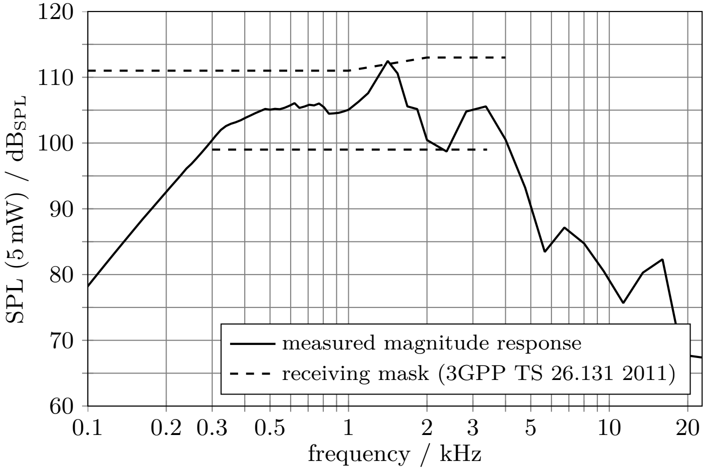


Figure 5.11: Measured magnitude response of *NXP 8x12x2 receiver*.

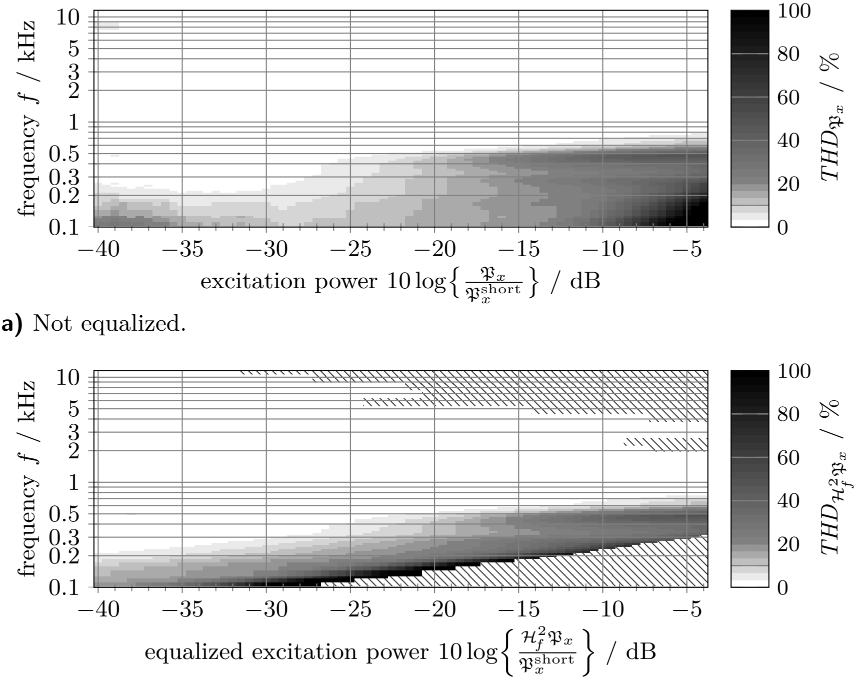
[9] The numbers in the product name again indicate the main dimensions in mm.

Total Harmonic Distortion (THD)

In Figure 5.12a, the measured THD of the *NXP 8x12x2 receiver* is plotted over excitation power and frequency. Please note, that again the increase in THD on the left side for low excitation power and especially for low frequencies is not a characteristic of the receiver but caused by a decreased SNR due to a very low response power.

In order to facilitate understanding of the measured THD, Figure 5.13 depicts the spectral response power of the *NXP 8x12x2 receiver* as a spectrogram over the frequency of the sine excitation for excitation powers $10\log\left\{\frac{\mathfrak{P}_x}{\mathfrak{P}_x^{\text{short}}}\right\} = -20\,\text{dB}$ and $10\log\left\{\frac{\mathfrak{P}_x}{\mathfrak{P}_x^{\text{short}}}\right\} = -10\,\text{dB}$. The main diagonal contains the linear response and parallel lines above it represent the higher harmonic distortion.

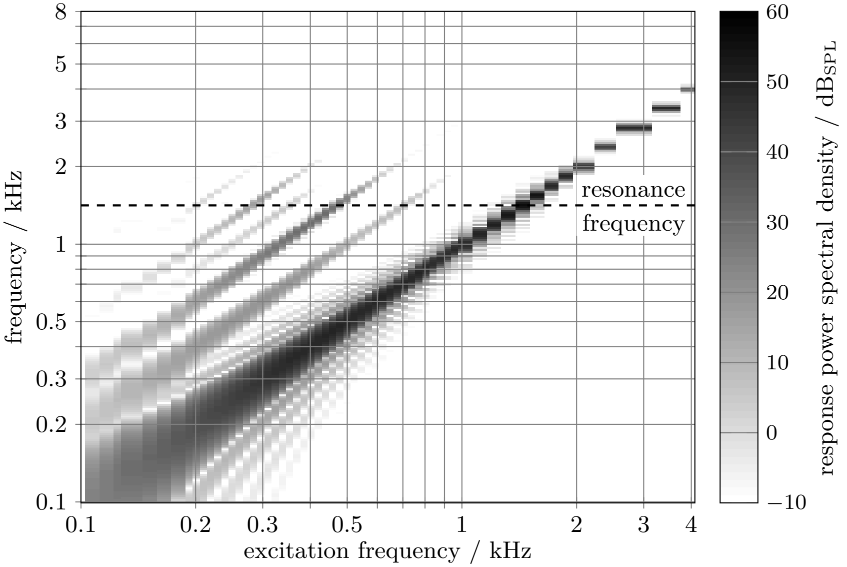
It can be seen, that the response of the *NXP 8x12x2 receiver* is dominated by the odd harmonics, i. e., mainly the third and fifth harmonic. This explains that the measured THD shows peaks at excitation frequencies around 470 Hz and 280 Hz, which are a third and fifth of the resonance frequency, respectively. As for the *NXP 13.6x9.6x2.9 speaker*, the harmonic distortions become insignificant (in



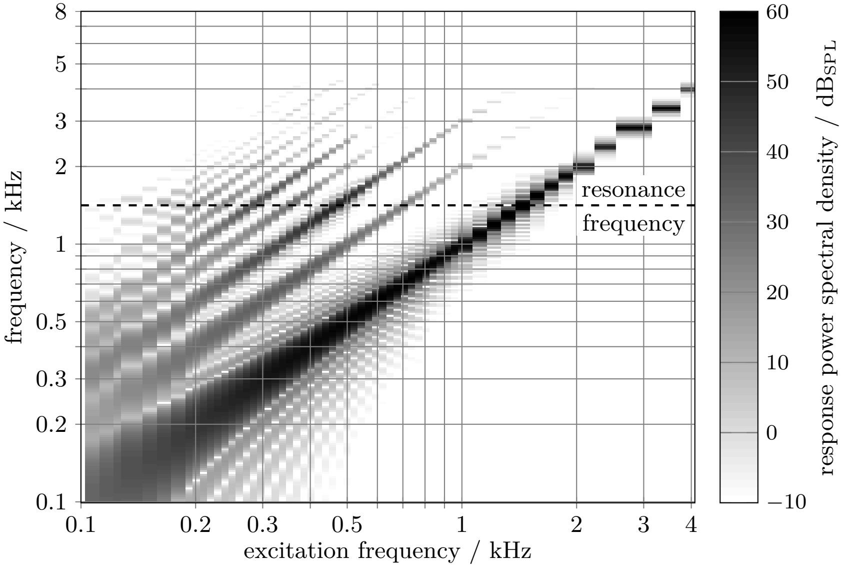
(a) Not equalized.

(b) Equalized according to Section 5.1.3.

Figure 5.12: Total harmonic distortion (THD) of *NXP 8x12x2 receiver*.



(a) Excitation power $10\log\{\frac{\mathfrak{P}_x}{\mathfrak{P}_x^{\mathrm{short}}}\} = -20\,\mathrm{dB}$.



(b) Excitation power $10\log\{\frac{\mathfrak{P}_x}{\mathfrak{P}_x^{\mathrm{short}}}\} = -10\,\mathrm{dB}$.

Figure 5.13: Spectrogram without equalization of *NXP 8x12x2 receiver* for different excitation powers.

relation to the linear response) for excitation frequencies above 800 Hz to 900 Hz.

With the same reason as in Section 5.2.1, the excitation signal is equalized in a second measurement, i. e., weighted with the inverse transfer function, as described in Section 5.1.3. Figure 5.12b shows the result of this measurement in terms of THD over equalized excitation power and frequency.

Due to the rather low cut-off frequency of the magnitude response of the receiver of about 350 Hz and the rather high sensitivity of the receiver, equalization makes the problem of distortions at low frequencies not as bad as for the *NXP 13.6x9.6x2.9 speaker*. If the THD should be below 10 %, frequency components between 320 Hz and 600 Hz must be attenuated to $10\log\left\{\frac{\mathfrak{P}_x}{\mathfrak{P}_x^{\text{short}}}\right\} = -21\,\text{dB}$, which corresponds to about 95 dB_{SPL}. Components below 320 Hz must be attenuated increasingly down to $10\log\left\{\frac{\mathfrak{P}_x}{\mathfrak{P}_x^{\text{short}}}\right\} = -40\,\text{dB}$ resp. 76 dB_{SPL} at 150 Hz. Frequencies above 600 Hz can be played basically up to the maximum continuous power.

Total Intermodulation Distortion (TID)

The measured TID of the *NXP 8x12x2 receiver* is plotted in Figure 5.14 for different frequencies over equalized excitation power. The ordinate shows the first excited frequency f_1, whereas the second excited frequency f_2 is given as a parameter. Both sine components are equalized as described in Section 5.1.2.

It can be seen from Figure 5.14, that the TID is approximately independent of frequency f_1 if it is above 800 Hz to 1000 Hz. Accordingly, the approximation

$$\mathit{TID}_{\mathcal{H}^2_{f_1}\mathfrak{P}_x+\mathcal{H}^2_{f_2}\mathfrak{P}_x}(f_1, f_2) \approx \mathit{THD}_{\mathcal{H}^2_f\mathfrak{P}_x}\big(\min\{f_1, f_2\}\big) \qquad \text{[5.10, p. 107]}$$

for $\max\{f_1, f_2\} \geq 800\,\text{Hz}$, which was originally derived in Section 5.2.1 for the *NXP 13.6x9.6x2.9 speaker*, also holds for the *NXP 8x12x2 receiver*.

Total Non-Linear Distortion (TND)

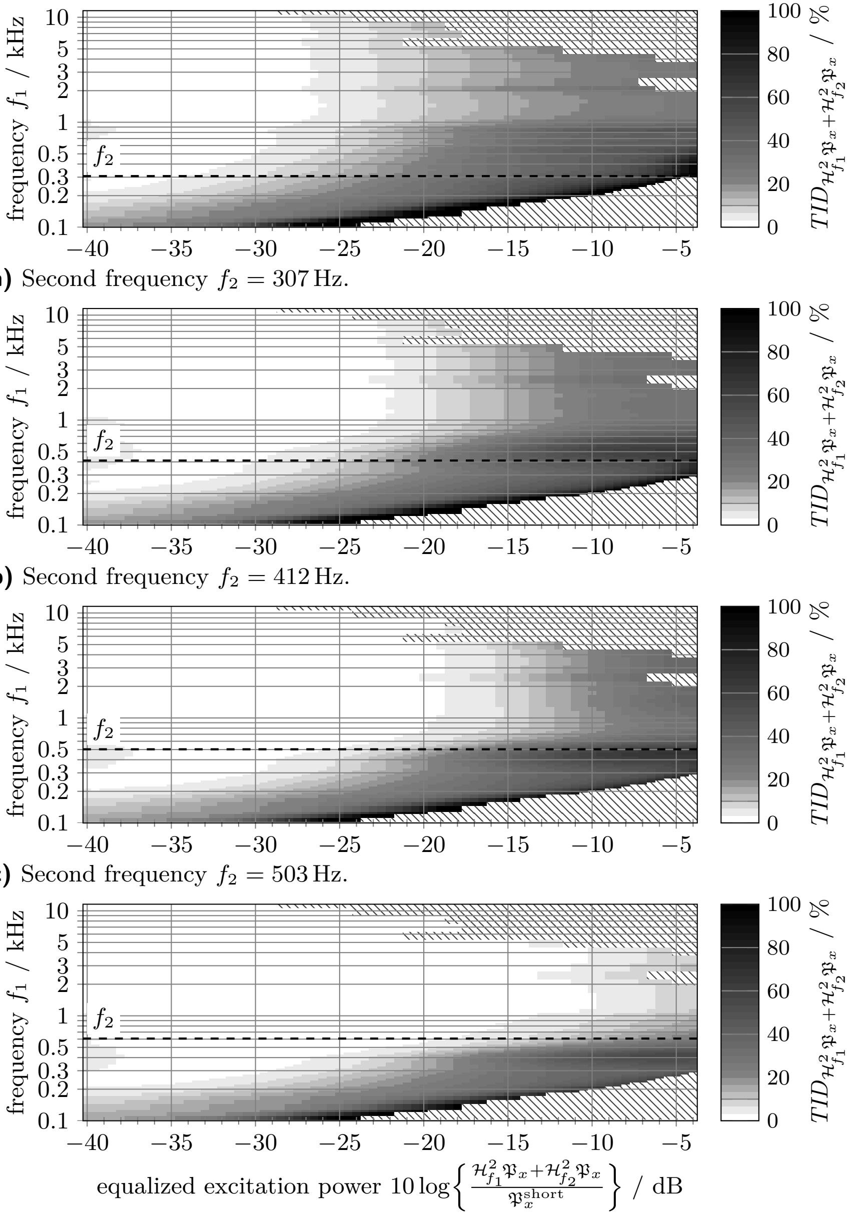
Figure 5.15 shows the measured TND of the *NXP 8x12x2 receiver* for different bandwidths over equalized excitation power. The ordinate shows the center frequency f_c of the excited band, whereas the bandwidth f_Δ is given in the caption. The final bandpass signal is equalized as described in Section 5.1.2.

As for the *NXP 13.6x9.6x2.9 speaker*, the measured TND of the receiver becomes "smoother" for higher bandwidths, which can again be expected as more frequency components with different harmonic behaviours are excited. Besides that, the general shape and behaviour is similar to that of the measured THD of Figure 5.12b.

Again, the ratio between distortion power and linear response power is higher for the TND than for the THD. The average increase of 2.3 dB is also about the same for all evaluated bandwidths, leading to the same approximation as for the *NXP 13.6x9.6x2.9 speaker*

$$\mathit{TND}_{\mathcal{H}^2\mathfrak{P}_x}(f_\text{c}, f_\Delta) = 1.3 \cdot \mathit{THD}_{\mathcal{H}^2_f\mathfrak{P}_x}(f_\text{c}) \, , \qquad \text{[5.11, p. 107]}$$

which is depicted for comparison in Figure 5.15d.



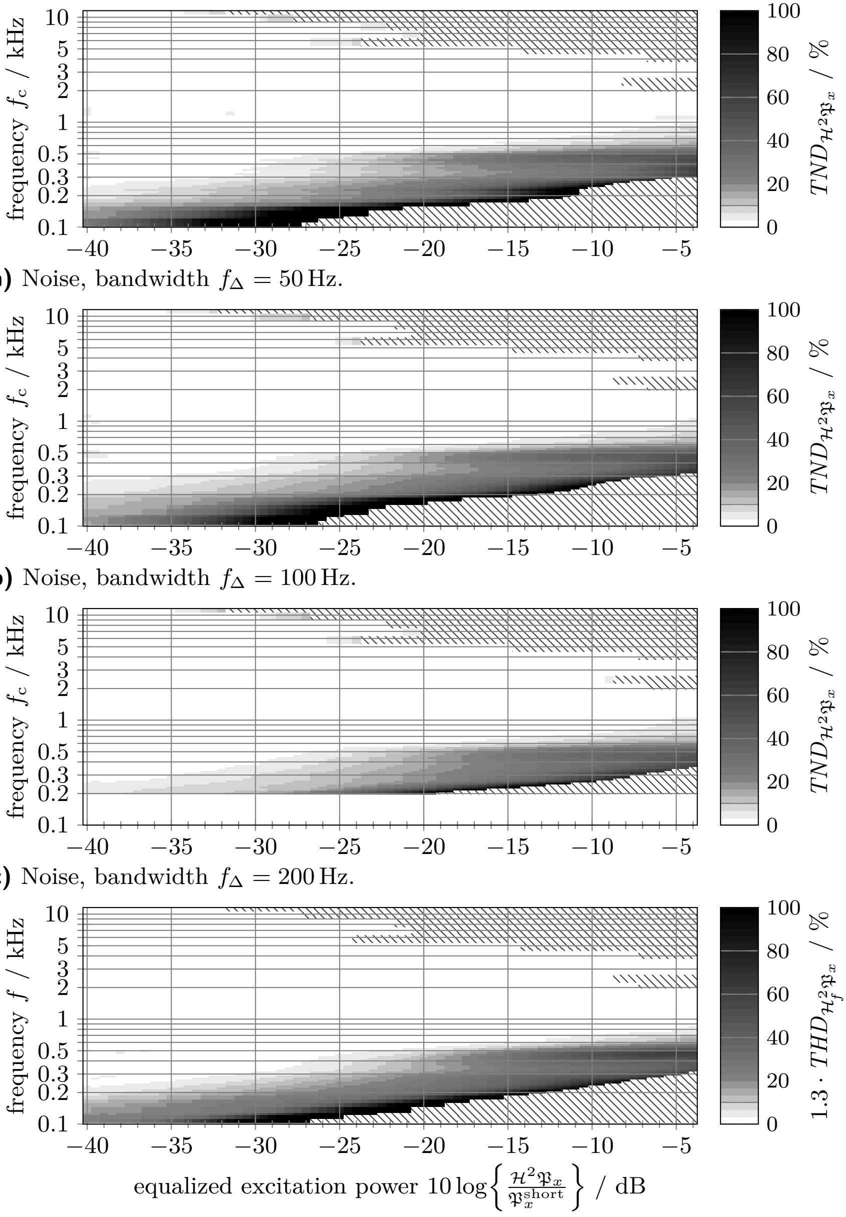
(a) Second frequency $f_2 = 307\,\mathrm{Hz}$.

(b) Second frequency $f_2 = 412\,\mathrm{Hz}$.

(c) Second frequency $f_2 = 503\,\mathrm{Hz}$.

(d) Second frequency $f_2 = 607\,\mathrm{Hz}$.

Figure 5.14: Total intermodulation distortion (TID) of *NXP 8x12x2 receiver* for different second frequencies.



(a) Noise, bandwidth $f_\Delta = 50\,\text{Hz}$.

(b) Noise, bandwidth $f_\Delta = 100\,\text{Hz}$.

(c) Noise, bandwidth $f_\Delta = 200\,\text{Hz}$.

(d) Single sine, cf. Figure 5.12b, THD multiplied with 1.3.

Figure 5.15: Total non-linear distortion (TND) of *NXP 8x12x2 receiver* for different bandwidths.

5.2.3 Discussion

For both, the *NXP 13.6x9.6x2.9 speaker* and the *NXP 8x12x2 receiver*, the measured TID as well as TND for all bandwidths show that frequency components above 800 Hz to 1000 Hz do not significantly contribute to the non-linear distortion of a multi-frequency or bandpass excitation. Therefore, it is feasible to conclude that excitation components above 1000 Hz, independent of their bandwidth and spectral shape, never cause significant non-linear distortions. Below 1000 Hz, non-linear distortions can reliably be kept within reasonable bounds if the excitation power does not exceed a frequency dependent limit, which is approximately independent[10] of the excitation bandwidth.

It can thus be concluded, that a decomposition of the loudspeaker signal into *one* subband *above* 1000 Hz and *some* subbands *below* 1000 Hz with a subsequent subband limitation is sufficient to combat non-linear distortions. Regarding the effectiveness of this scheme, the exact number and bandwidths of the subbands below 1000 Hz are of secondary importance. Evaluations show that reasonably tight power limits can already be found with three to five subbands.

A second finding of the above measurements with three measures THD, TID, and TND is, that measurements with only the THD are sufficient to describe the system reasonably well for frequencies above 400 Hz. Thus time-consuming measurements with TID and TND are not necessary to tune the subband limiters.

Smaller measurement series with other speakers and receivers as well as a specially prepared real mobile phone suggest that the found behaviour and the drawn conclusions are universal among most micro-loudspeakers.

5.3 Loudspeaker Protection

Loudspeaker protection (LOPRO) as described in this section represents a safety unit to prevent damage and failure of the transducer. Figure 5.16 shows the proposed general LOPRO framework.

The loudspeaker signal $x(k)$ is segmented in I real-valued bandpass signals $x_i(k)$ by means of an analysis filterbank. A suitable filterbank is derived in Sections 5.3.1 and 5.3.2 based on the generalized sliding DFT (GSDFT). Each subband signal is then limited by an individual subband time-domain limiter as described in Section 5.3.3 to consider the frequency dependent excursion of the membrane of the transducer as well as the frequency dependent amplification by the transducer equalization. A synthesis filterbank reconstructs the (frequency dependently limited) output signal and a final fullband time-domain limiter takes care of the maximum total power that the transducer coil can stand.

This concept is verified with a measurement campaign presented in Section 5.3.4. Section 5.3.5 finally evaluates the impact of LOPRO on speech intelligibility.

[10] The limit for bandpass excitation underestimates the limit for pure sine excitation by about 2.3 dB as shown in Sections 5.2.1 and 5.2.2.

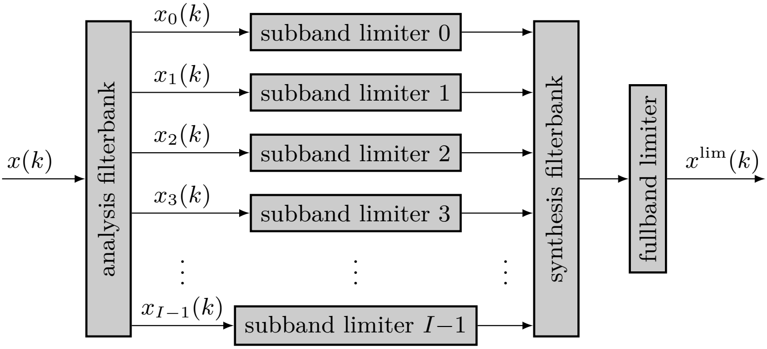


Figure 5.16: General framework for loudspeaker protection.

5.3.1 Filterbank Summation Method

In order to allow the subband limiters to react sufficiently fast, the analysis filterbank should exhibit no or only very little downsampling. Furthermore, the LOPRO framework should introduce no signal distortion in the normal case where no limitation is necessary, i. e., the analysis and synthesis filterbanks should be perfectly reconstructing in this case.

Both criteria are fulfilled by the filterbank summation method (FBSM), where a generalized DFT (GDFT) filterbank of even size M without downsampling is used as analysis filterbank. In every sample instant, the normalized GDFT coefficients

$$\mathcal{X}_\mu(k) = \frac{1}{M} \sum_{l=0}^{M-1} h(l) \cdot x(k+l-M+1) \cdot \mathrm{e}^{-\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)(l-l_0)} \tag{5.12}$$

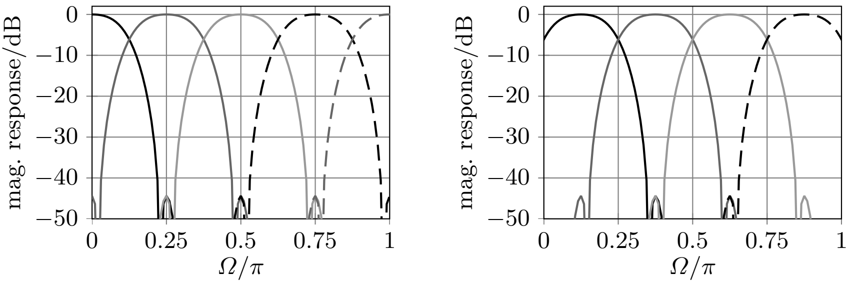
are calculated, where $h(l)$ can be a window function to reduce spectral leakage. A causal and approximately linear-phase FIR bandpass filter is obtained with $l_0 = \frac{M}{2}$. The common, *evenly-stacked* frequency bands are obtained for $\mu_0 = 0$, whereas $\mu_0 = \frac{1}{2}$ “shifts” all frequency bands by one half of their widths, leading to *oddly-stacked* frequency bands, cf. (Crochiere & Rabiner 1983). Evenly- and oddly-stacked frequency bands are exemplarily depicted in Figure 5.17.

The synthesis filterbank of the FBSM simplifies to the sum of all coefficients

$$x^{\lim}(k) = \sum_{\mu=0}^{M-1} \mathcal{X}_\mu(k). \tag{5.13}$$

It can be shown, that the FBSM is perfectly reconstructing, i. e., the output signal $x^{\lim}(k)$ is a delayed version of the input signal

$$x^{\lim}(k) = x\big(k - \tfrac{M}{2}\big)\,. \tag{5.14}$$



(a) Evenly-stacked frequency bands.

(b) Oddly-stacked frequency bands.

Figure 5.17: Magnitude responses of evenly- and oddly-stacked frequency bands with $M = 8$.

As motivated in Section 5.2.3, it is sufficient for LOPRO to split the frequency range below 1 kHz in three to five small subbands and one large subband above 1 kHz, leading to $I = 4$ to $I = 6$ subbands.

Without loss of generality, it is assumed in the following, that the $I - 1$ "small" subbands are consecutive starting with index 0, which eases the notation. Their $I-1$ real-valued bandpass signals $x_i(k)$ are calculated utilizing the complex conjugate symmetry $\mathcal{X}_\mu(\kappa) = [\mathcal{X}_{M-2\mu_0-\mu}(\kappa)]^*$ of the GDFT (with even size) of the real-valued signal $x(k)$, cf. (2.17):

$$x_i(k) = g_{\mathrm{sym},i} \cdot \mathrm{Re}\big\{\mathcal{X}_i(k)\big\}\,, \quad 0 \leq i \leq I-2\,, \tag{5.15}$$

with the symmetry factor

$$g_{\mathrm{sym},i} = \begin{cases} 1 & \text{if } i \in \big\{0,\ \frac{M}{2}\big\} \text{ and } \mu_0 = 0 \\ 2 & \text{otherwise}\,. \end{cases} \tag{5.16}$$

The "large" subband signal $x_{I-1}(k)$ is calculated as the sum of all remaining coefficients

$$x_{I-1}(k) = \sum_{i=I-1}^{\frac{M}{2}-2\mu_0} g_{\mathrm{sym},i} \cdot \mathrm{Re}\big\{\mathcal{X}_i(k)\big\}\,. \tag{5.17}$$

With the perfect reconstruction property (5.14), the "large" subband with all frequencies above 1 kHz can alternatively be calculated as the difference of the properly delayed input signal and the "small" subband signals below 1 kHz:

$$x_{I-1}(k) = x\big(k - \tfrac{M}{2}\big) - \sum_{i=0}^{I-2} x_i(k)\,. \tag{5.18}$$

Even though the FFT can be used to calculate the GDFT, the FBSM still has a considerable computation complexity as a full FFT has to be calculated in every sample instant.

5.3.2 Generalized Sliding DFT

The sliding DFT (SDFT) (Jacobsen & Lyons 2003, 2004) allows an efficient calculation of a DFT filterbank without downsampling especially if only few DFT coefficients are actually needed. In this section, a computationally efficient generalized sliding DFT (GSDFT) is derived.

Given the normalized GDFT coefficients of the preceding sample

$$\mathcal{X}_\mu(k-1) = \frac{1}{M}\sum_{l=0}^{M-1} x(k+l-M)\cdot \mathrm{e}^{-\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)(l-l_0)}, \tag{5.19}$$

it follows for the normalized GDFT coefficients of the current sample

$$\begin{aligned}
\mathcal{X}_\mu(k) &= \frac{1}{M}\sum_{l=0}^{M-1} x(k+l-M+1)\cdot \mathrm{e}^{-\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)(l-l_0)} && (5.20)\\
&= \frac{1}{M}\sum_{l'=1}^{M} x(k+l'-M)\cdot \mathrm{e}^{-\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)(l'-l_0-1)} \quad \text{with} \quad l'=l+1 && (5.21)\\
&= \left(\frac{1}{M}\sum_{l'=1}^{M} x(k+l'-M)\cdot \mathrm{e}^{-\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)(l'-l_0)}\right)\cdot \mathrm{e}^{\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)} && (5.22)\\
&= \Big(\mathcal{X}_\mu(k-1) + \frac{1}{M}\cdot x(k)\cdot \mathrm{e}^{-\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)(M-l_0)}\\
&\qquad - \frac{1}{M}\cdot x(k-M)\cdot \mathrm{e}^{-\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)(-l_0)}\Big)\cdot \mathrm{e}^{\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)} && (5.23)\\
&= \left(\mathcal{X}_\mu(k-1) + \big(x(k)\cdot \mathrm{e}^{-\mathrm{j}2\pi\mu_0} - x(k-M)\big)\cdot \frac{\mathrm{e}^{\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)l_0}}{M}\right)\cdot \mathrm{e}^{\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)}. && (5.24)
\end{aligned}$$

With $\mu_0 \in \{0,\, \frac{1}{2}\}$ and $l_0 = \frac{M}{2}$ as before, (5.24) simplifies to

$$\mathcal{X}_\mu(k) = \left(\mathcal{X}_\mu(k-1) + \big(x(k)\cdot(-1)^{2\mu_0} - x(k-M)\big)\cdot \frac{(-1)^\mu\cdot \mathrm{j}^{2\mu_0}}{M}\right)\cdot \mathrm{e}^{\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)}. \tag{5.25}$$

Finally, the subband signals $x_i(k)$ are calculated using

$$x_i(k) = g_{\mathrm{sym},i}\cdot \mathrm{Re}\big\{\mathcal{X}_i(k)\big\}, \quad 0 \le i \le I-2, \qquad [5.15,\ \text{p. } 118]$$

$$x_{I-1}(k) = x\big(k-\tfrac{M}{2}\big) - \sum_{i=0}^{I-2} x_i(k), \qquad [5.18,\ \text{p. } 118]$$

which leads to the efficient LOPRO scheme sketched in Figure 5.18.

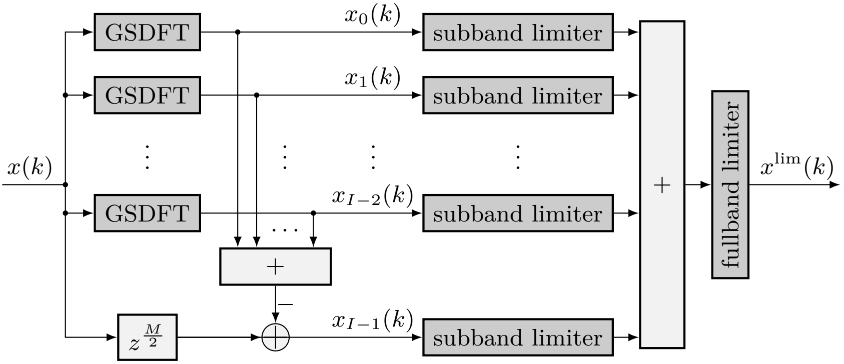


Figure 5.18: Framework for loudspeaker protection with GSDFT filterbank.

Windowing

The plain GDFT as described in (5.20) yields in a comparably high spectral leakage, which is commonly reduced by a time-domain multiplication of the input samples $x(k)$ with a window function $h(l)$, cf. (5.12). Unfortunately, this would break the derivation of the SDFT. An alternative, equivalent approach is the frequency-domain convolution of the output of the GSDFT with the DFT of the window function (Jacobsen & Lyons 2003). Accordingly, a Hann windowed GDFT coefficient $\tilde{\mathcal{X}}_\mu(k)$ is calculated by the three-point convolution

$$\tilde{\mathcal{X}}_\mu(k) = \underline{H} \cdot \begin{pmatrix} \mathcal{X}_{\mu-1}(k) \\ \mathcal{X}_\mu(k) \\ \mathcal{X}_{\mu+1}(k) \end{pmatrix} \tag{5.26}$$

with the window coefficients

$$\underline{H} = \begin{pmatrix} -0.25 & +0.50 & -0.25 \end{pmatrix}. \tag{5.27}$$

The window coefficients

$$\underline{H} = \begin{pmatrix} -0.46 & +0.54 & -0.46 \end{pmatrix} \tag{5.28}$$

yield a Hamming windowing and the rectangular window of (5.25) is included as special case with

$$\underline{H} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}. \tag{5.29}$$

Obviously, the two adjacent GDFT coefficients $\mathcal{X}_{\mu-1}(k)$ and $\mathcal{X}_{\mu+1}(k)$ must be additionally calculated before convolution, which adds, of course, some computational complexity. However, in the common use case that a continuous block of subbands $\tilde{\mathcal{X}}_{i_\mathrm{f}}(k), \tilde{\mathcal{X}}_{i_\mathrm{f}+1}(k), \ldots, \tilde{\mathcal{X}}_{i_\mathrm{l}}(k)$ is to be calculated, the overall extra effort compared to the rectangular window case (5.25) consists only of two additional GDFT coefficients $\mathcal{X}_{i_\mathrm{f}-1}(k)$ and $\mathcal{X}_{i_\mathrm{l}+1}(k)$.

Complexity Consideration

In order to reduce the number of multiplications, a modified GDFT coefficient

$$\begin{aligned}\mathcal{X}'_\mu(k) &= \tfrac{M}{(-1)^\mu\cdot \mathrm{j}^{2\mu_0}} \cdot \mathcal{X}_\mu(k) &(5.30)\\ &= \Big(\mathcal{X}'_\mu(k-1) + \underbrace{\big(x(k)\cdot(-1)^{2\mu_0} - x(k-M)\big)}_{\text{calculated once for all subbands}}\Big)\cdot \mathrm{e}^{\mathrm{j}\frac{2\pi}{M}(\mu+\mu_0)} &(5.31)\end{aligned}$$

is introduced and used for the windowed GDFT coefficient

$$\begin{aligned}\tilde{\mathcal{X}}_\mu(k) &= \underline{H}\cdot\begin{pmatrix}\frac{(-1)^{\mu-1}\cdot \mathrm{j}^{2\mu_0}}{M}\cdot\mathcal{X}'_{\mu-1}(k)\\ \frac{(-1)^{\mu}\quad\cdot \mathrm{j}^{2\mu_0}}{M}\cdot\mathcal{X}'_{\mu}(k)\\ \frac{(-1)^{\mu+1}\cdot \mathrm{j}^{2\mu_0}}{M}\cdot\mathcal{X}'_{\mu+1}(k)\end{pmatrix} &(5.32)\\ &= \underline{H}\cdot\tfrac{(-1)^\mu}{M}\cdot\begin{pmatrix}-1&0&0\\0&+1&0\\0&0&-1\end{pmatrix}\cdot \mathrm{j}^{2\mu_0}\cdot\begin{pmatrix}\mathcal{X}'_{\mu-1}(k)\\ \mathcal{X}'_{\mu}(k)\\ \mathcal{X}'_{\mu+1}(k)\end{pmatrix}. &(5.33)\end{aligned}$$

Using (5.15) it follows for $0 \le i \le I-2$, that

$$x_i(k) = \underline{H}'_i \cdot \mathrm{Re}\left\{(-\mathrm{j})^{2\mu_0}\cdot\begin{pmatrix}\mathcal{X}'_{i-1}(k)\\ \mathcal{X}'_{i}(k)\\ \mathcal{X}'_{i+1}(k)\end{pmatrix}\right\} \tag{5.34}$$

with $\underline{H}'_i$ denoting the constant and *real* vector of modified window coefficients for the i-th subband

$$\underline{H}'_i = \underline{H}\cdot g_{\mathrm{sym},i}\cdot\tfrac{(-1)^i\cdot(-1)^{2\mu_0}}{M}\cdot\begin{pmatrix}-1&0&0\\0&+1&0\\0&0&-1\end{pmatrix}. \tag{5.35}$$

With $\mathrm{Re}\big\{-\mathrm{j}\cdot\mathcal{X}'_\mu(k)\big\} = \mathrm{Im}\big\{\mathcal{X}'_\mu(k)\big\}$, (5.34) simplifies to the case distinction

$$x_i(k) = \underline{H}'_i\cdot\begin{cases}\mathrm{Re}\left\{\begin{pmatrix}\mathcal{X}'_{i-1}(k)\\ \mathcal{X}'_{i}(k)\\ \mathcal{X}'_{i+1}(k)\end{pmatrix}\right\} & \text{if } \mu_0 = 0\\ \mathrm{Im}\left\{\begin{pmatrix}\mathcal{X}'_{i-1}(k)\\ \mathcal{X}'_{i}(k)\\ \mathcal{X}'_{i+1}(k)\end{pmatrix}\right\} & \text{if } \mu_0 = 0.5\,,\end{cases} \tag{5.36}$$

which can be implemented as an index shift in most low-level programming languages.

This way, four (partly complex) multiplications in (5.15), (5.25), and (5.26) are substituted by *one real* multiplication at the expense of a very slight increase of static memory of up to $3\cdot(I-2)$ real values.

In Tables 5.2 and 5.3 the computational complexity of the GSDFT filterbank without windowing, i. e., with a rectangular window, as well as with a Hann or Hamming window is given. Table 5.4 shows the memory requirement of the GSDFT filterbank.

part of filterbank	eq.	operations per sample
difference of input signal	(5.31)	$1\,\mathrm{mult} + 1\,\mathrm{add}$
update of GDFT coefficient	(5.31)	$(4\,\mathrm{mult} + 3\,\mathrm{add}) \cdot (I-1)$
multiplication with coefficient	(5.36)	$1\,\mathrm{mult} \cdot (I-1)$
calculation of “large” subband	(5.18)	$1\,\mathrm{add} \cdot (I-1)$
sum		$(5\,\mathrm{mult} + 4\,\mathrm{add}) \cdot I - 4\,\mathrm{mult} - 3\,\mathrm{add}$
example: $I = 5$ subbands		$21\,\mathrm{mult} + 17\,\mathrm{add}$

Table 5.2: Computational complexity of GSDFT filterbank with rectangular window and $I-1$ “small” subbands.

part of filterbank	eq.	operations per sample
difference of input signal	(5.31)	$1\,\mathrm{mult} + 1\,\mathrm{add}$
update of GDFT coefficient	(5.31)	$(4\,\mathrm{mult} + 3\,\mathrm{add}) \cdot (I+1)$
convolution with coefficients	(5.36)	$(3\,\mathrm{mult} + 2\,\mathrm{add}) \cdot (I-1)$
calculation of “large” subband	(5.18)	$1\,\mathrm{add} \cdot (I-1)$
sum		$(7\,\mathrm{mult} + 6\,\mathrm{add}) \cdot I + 2\,\mathrm{mult} + 1\,\mathrm{add}$
example: $I = 5$ subbands		$37\,\mathrm{mult} + 31\,\mathrm{add}$

Table 5.3: Computational complexity of GSDFT filterbank with Hann or Hamming window and $I-1$ *consecutive* “small” subbands.

part of filterbank	eq.	real values
difference of input signal	(5.31)	M
update of GDFT coefficient	(5.31)	$2 \cdot$ number of GDFT coefficients $\mathcal{X}'_\mu$: $2 \cdot (I-1)$ for rectangular window $2 \cdot (I+1)$ for Hann/Hamming window

Table 5.4: Memory requirement of GSDFT filterbank with $I-1$ *consecutive* “small” subbands.

5.3.3 Time-Domain Limiter

For the limitation, a straight-forward and commonly known recursive smoothing approach is used. Without loss of generality, the time-domain limiter is presented in the following with input signal $x(k)$ and output signal $x^{\mathrm{lim}}(k)$, although it is actually applied to the subband signals $x_i(k)$ as well as the reconstructed fullband signal.

For power calculation, the input signal $x(k)$ with (original) sampling rate f_s is segmented in frames of length R with the sub-sampled time index $\kappa = \lfloor {}^{k}/_{R} \rfloor \cdot R$. A common frame length is $R = 1\,\mathrm{ms} \cdot f_\mathrm{s}$. For each frame κ, the root mean square (RMS) $\mathfrak{R}_x(\kappa)$ of the current frame is calculated as

$$\mathfrak{R}_x(\kappa) = \sqrt{\frac{\sum\limits_{\zeta=0}^{R-1} x^2(\kappa+\zeta)}{R}}\,. \tag{5.37}$$

The instantaneous RMS curve is smoothed according to

$$\overline{\mathfrak{R}}_x(\kappa) = \begin{cases} \alpha_{\mathfrak{R},\mathrm{a}} \cdot \mathfrak{R}_x(\kappa) + (1-\alpha_{\mathfrak{R},\mathrm{a}}) \cdot \overline{\mathfrak{R}}_x(\kappa-1) & \text{if } \mathfrak{R}_x(\kappa) > \overline{\mathfrak{R}}_x(\kappa-1) \\ \alpha_{\mathfrak{R},\mathrm{r}} \cdot \mathfrak{R}_x(\kappa) + (1-\alpha_{\mathfrak{R},\mathrm{r}}) \cdot \overline{\mathfrak{R}}_x(\kappa-1) & \text{otherwise} \end{cases} \tag{5.38}$$

with the smoothing coefficients

$$\alpha_{\mathfrak{R},\mathrm{a}} = 1 - \exp\left\{-\frac{1}{2} \cdot \frac{R}{f_\mathrm{s} \cdot \tau_{\mathfrak{R},\mathrm{a}}}\right\}, \tag{5.39}$$

$$\alpha_{\mathfrak{R},\mathrm{r}} = 1 - \exp\left\{-\frac{1}{2} \cdot \frac{R}{f_\mathrm{s} \cdot \tau_{\mathfrak{R},\mathrm{r}}}\right\}. \tag{5.40}$$

The RMS attack time constant $\tau_{\mathfrak{R},\mathrm{a}}$ is usually lower than the RMS release time constant $\tau_{\mathfrak{R},\mathrm{r}}$ to allow a fast reaction on increasing power levels.

In the subband limiter, $\tau_{\mathfrak{R},\mathrm{a}}$ is as small as $1\,\mathrm{ms}$ to protect from excessive excursions. $\tau_{\mathfrak{R},\mathrm{r}}$ should be as large as the largest wavelength in the corresponding subband to avoid changes of gain within one period. For the fullband limiter, $\tau_{\mathfrak{R},\mathrm{a}}$ is determined by the time the transducer can stand a high power level and $\tau_{\mathfrak{R},\mathrm{r}}$ by the time the transducer needs to cool down again.

The limiter gain $G(\kappa)$ is derived from the smoothed RMS $\overline{\mathfrak{R}}_x(\kappa)$ to limit the electric power to $\mathfrak{P}_x^{\max}$:

$$G(\kappa) = \min\left\{\frac{\mathfrak{P}_x^{\max}}{g_{\mathrm{ls}}^2 \cdot \overline{\mathfrak{R}}_x^2(\kappa)},\, 1\right\} \tag{5.41}$$

with g_{ls}^2 denoting the proportionality factor between digital audio signal power and electric power at the loudspeaker, which has unit W and depends on the digital-analog conversion and the amplifier setting.

In order to prevent distortions due to jumps at the frame boundaries and too fast gain fluctuations, the limiter gain $G(\kappa)$ is smoothed

$$\overline{G}(\kappa) = \begin{cases} \alpha_{G,\mathrm{a}} \cdot G(\kappa) + (1-\alpha_{G,\mathrm{a}}) \cdot \overline{G}(\kappa-1) & \text{if } G(\kappa) \leq \overline{G}(\kappa-1) \\ \alpha_{G,\mathrm{r}} \cdot G(\kappa) + (1-\alpha_{G,\mathrm{r}}) \cdot \overline{G}(\kappa-1) & \text{otherwise.} \end{cases} \tag{5.42}$$

With descending gain, the limitation tightens and the gain is smoothed with the attack time $\tau_{G,\mathrm{a}}$, whereas a rising gain, accompanied with a relaxing limitation, is smoothed with the release time $\tau_{G,\mathrm{r}}$, resulting in the smoothing coefficients

$$\alpha_{G,\mathrm{r}} = 1 - \exp\left\{-\frac{1}{2} \cdot \frac{R}{f_\mathrm{s} \cdot \tau_{G,\mathrm{r}}}\right\}, \tag{5.43}$$

$$\alpha_{G,\mathrm{a}} = 1 - \exp\left\{-\frac{1}{2} \cdot \frac{R}{f_\mathrm{s} \cdot \tau_{G,\mathrm{a}}}\right\}. \tag{5.44}$$

To prevent a slow limiter reaction on quickly increasing signal levels, $\tau_{G,\mathrm{a}}$ is usually lower than $\tau_{G,\mathrm{r}}$ (see also Tables 5.5 and 5.6).

Finally, the input samples of the corresponding frame are multiplied with the smoothed output gain resulting in the limited signal $x^{\mathrm{lim}}(k)$ with

$$x^{\mathrm{lim}}(k) = x(k) \cdot \overline{G}(\kappa)\,. \tag{5.45}$$

5.3.4 Verification

In order to experimentally verify the effectiveness of the LOPRO system, a second measurement campaign with the *NXP 13.6x9.6x2.9 speaker* was conducted[11] in the hemi-anechoic chamber at the Institute of Technical Acoustics at the RWTH Aachen University.

Measurement Setup

The general setup is depicted in Figure 5.19. A device called *FireRobo*, which is built by the Institute of Technical Acoustics, is used as sound card and output amplifier. For recording the sound pressure, a *Brüel & Kjaer free-field 1/2" microphone type 4190* with a *Brüel & Kjaer microphone preamplifier type 2669* is used, which is connected to a *Brüel & Kjaer Nexus Conditioning Amplifier 2690.* The microphone is placed perpendicular to the speaker in a distance of 20 cm.

To assess the excursion of the membrane, an *LMS Laser Vibrometer* by *Polytec* with an *OFV-055 optical scanning head* is used. This device measures the velocity of the membrane up to 1.25 m/s, which is later integrated in software to get the excursion. As the membrane of the speaker was quite matt, a small piece (about 1 mm^2) of reflective tape was placed on the membrane. A pilot measurement indicated only a negligible impact of the tape on the behavior of the speaker.

Since the voice coil is encapsulated inside the transducer, its temperature cannot easily be measured without destroying the transducer. Therefore, as an approximation of the voice coil temperature, the temperature of the back side of

[11] With kind support of Markus Müller-Trapet from the Institute of Technical Acoustics.

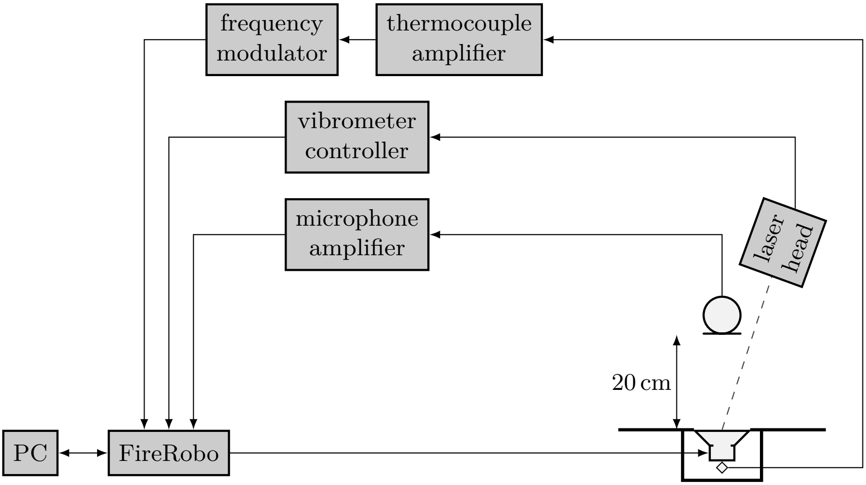


Figure 5.19: Setup for measurement of membrane excursion and transducer temperature. ◇ symbolizes the insulated thermocouple.

the transducer is acquired using an insulated thermocouple of type "T", which is attached to the speaker (inside the back volume) using some thermal grease and a piece of tape. The thermocouple amplifier *MCR-T-UI-E* by *Phoenix Contact* maps the temperature between 0 °C and 100 °C linearly to an output voltage between 0 V and 10 V. Since the input of the *FireRobo* is unable to capture direct current (DC) signals, the temperature voltage is frequency modulated using a *TOELLNER TOE 7401* sine generator with voltage-controlled oscillator input and, after recording, frequency demodulated by software.

The housing of the *NXP 13.6x9.6x2.9 speaker* is manufactured from acrylic glass using a CNC milling machine based on the drawings of the housing used in the first measurement campaign of Section 5.2.1. A back cavity of 1.5 cm^3 was chosen for these measurements.

For measurements of transfer functions, the speaker is excited with an exponential sweep. This allows not only to examine the whole continuous frequency range in one (rather short) measurement, but also to separate the linear response of the system from each harmonic response (Müller & Massarani 2001). This technique, however, requires a better SNR than the measurement with single sine waves as used in the first measurement campaign.

In contrast to the first measurement campaign of Section 5.2, no excitation signal was equalized to compensate for the speaker response.

Measurements are performed with and without LOPRO to verify its effectiveness. In either case, the stated power level refers to the "demanded" level *before* LOPRO to facilitate an easy mapping, i. e., the actual power level *after* LOPRO may be lower than the stated level.

Configuration

Three LOPRO configurations are evaluated in this campaign:

1. fullband limitation only, to verify temperature control,
2. subband limitation only, to verify limitation of membrane excursion, and
3. combined subband and fullband limitation, to evaluate the interaction and show the effect on THD.

The sampling rate $f_\mathrm{s} = 48\,\mathrm{kHz}$ is used for all processing and all measurements. The frame length is $R = 1\,\mathrm{ms} \cdot f_\mathrm{s} = 48\,\mathrm{samples}$ in all configurations.

In both configurations with fullband limitation (1 and 3), the fullband signal power is limited to $\mathfrak{P}_x^\mathrm{max} = 300\,\mathrm{mW}$, which is the maximum continuous power $\mathfrak{P}_x^\mathrm{cont}$ specified in (Knowles 2011). The RMS attack time constant is $\tau_{\mathfrak{R},\mathrm{a}} = 1\,\mathrm{s}$, which is the time the speaker can stand the maximum short-term power. The RMS release time constant is chosen arbitrarily to $\tau_{\mathfrak{R},\mathrm{r}} = 10\,\mathrm{s}$, both gain time constants to $\tau_{G,\mathrm{a}} = \tau_{G,\mathrm{r}} = 0.1\,\mathrm{s}$.

In both configurations with subband limitation (2 and 3), $I = 6$ real-valued subbands of an SDFT filterbank with DFT size $M = 240$, evenly-stacked frequency bands, and Hann windowing are used. Accordingly, the 5 "small" subbands have center frequencies 0 Hz, 200 Hz, 400 Hz, 600 Hz, and 800 Hz as depicted in Figure 5.20.

Table 5.5 shows the time constants for all time-domain limiters. The purpose of the second configuration "subband limitation only" is to limit the membrane excursion to about 0.7 mm peak-to-peak, which is just above the maximum linear excursion of 0.6 mm peak-to-peak. As the excursion is highest around resonance frequency and rather low beneath, the power limits are rather strict in the third

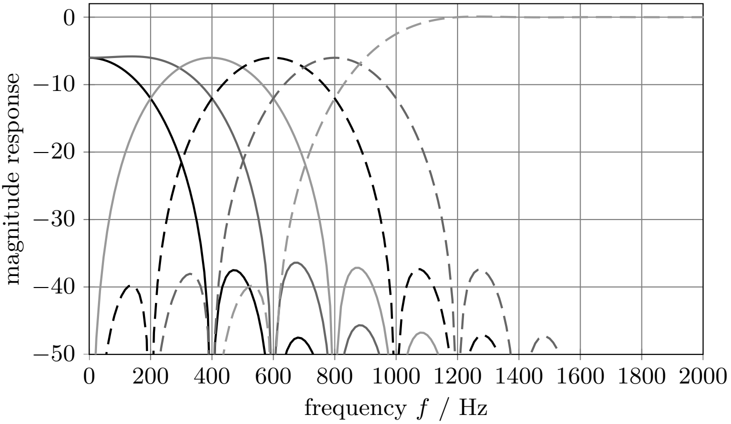


Figure 5.20: Magnitude responses of SDFT analysis filterbank with $I = 6$ real-valued subbands.

	RMS	RMS	gain	gain	Config. 1:	Config. 2:	Config. 3:
	attack	release	attack	release	power	power	power
	time	time	time	time	limit	limit	limit
frequencies	$\tau_{\mathfrak{R},\mathrm{a}}$	$\tau_{\mathfrak{R},\mathrm{r}}$	$\tau_{G,\mathrm{a}}$	$\tau_{G,\mathrm{r}}$	$\mathfrak{P}_x^{\max}$	$\mathfrak{P}_x^{\max}$	$\mathfrak{P}_x^{\max}$
0 Hz to 100 Hz	1 ms	20 ms	5 ms	10 ms	—	400 mW	2.5 mW
100 Hz to 300 Hz	1 ms	8 ms	5 ms	10 ms	—	200 mW	2.5 mW
300 Hz to 500 Hz	1 ms	4 ms	5 ms	10 ms	—	100 mW	2.5 mW
500 Hz to 700 Hz	1 ms	2 ms	5 ms	10 ms	—	25 mW	5 mW
700 Hz to 900 Hz	1 ms	2 ms	5 ms	10 ms	—	25 mW	25 mW
>900 Hz	—	—	—	—	—	—	—
fullband	1 s	10 s	0.1 s	0.1 s	300 mW	—	300 mW

Table 5.5: Parameters of limiters for *NXP 13.6x9.6x2.9 speaker.*

and fourth subband and more relaxed below. The third configuration “combined subband and fullband limitation” is supposed to also reduce harmonic distortions, which almost only occur at frequencies below resonance frequency. Therefore, the power limits at low frequency bands are even stricter in this configuration.

Limitation of Temperature

Figure 5.21 shows the development of the temperature at the back side of the speaker when excited 60 s with the simulated programme noise of (IEC 60268-1 1985) at the maximum short-term power $\mathfrak{P}_x^{\text{short}} = 700\,\text{mW}$ with fullband limitation (Configuration 1) and without.

Starting at 22 °C, the temperature rises in 60 s without limitation to 54 °C and keeps rising, whereas it levels off at 27 °C with fullband limitation. The presented fullband time-domain limiter thus effectively limits the temperature of the transducer.

Limitation of Excursion

The excursion at different excitation powers is plotted over frequency in Figure 5.22 with subband limitation (Configuration 2) and without.

As expected, the largest excursion occurs around the resonance frequency of about 700 Hz. Without limitation it goes up to 1.055 mm peak-to-peak at the maximum short-term power $\mathfrak{P}_x^{\text{short}} = 700\,\text{mW}$. With the presented subband limitation, the excursion is always below the target excursion of 0.7 mm peak-to-peak. Furthermore, it exceeds the maximum linear excursion of 0.6 mm peak-to-peak only with excitation power levels above the maximum continuous power $\mathfrak{P}_x^{\text{cont}} = 300\,\text{mW}$ and even then only for frequencies between 500 Hz and 800 Hz.

Figure 5.23 shows the spectrogram of the responses with and without subband limitation at 700 mW excitation power. It can be seen that the harmonic structure is strongly reduced around resonance frequency, which also indicates the effectiveness

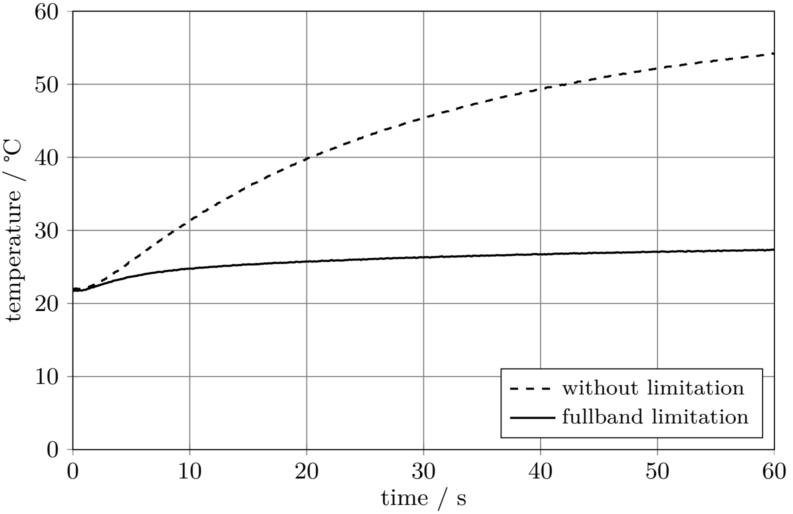


Figure 5.21: Temperature of *NXP 13.6x9.6x2.9 speaker* excited with IEC-60268 noise with and without fullband limitation (Configuration 1), power $\mathfrak{P}_x = \mathfrak{P}_x^{\text{short}} = 700\,\text{mW}$.

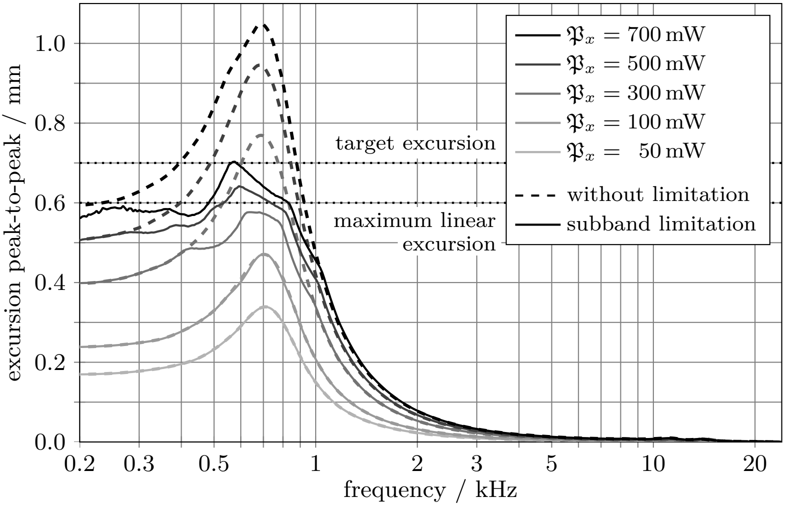
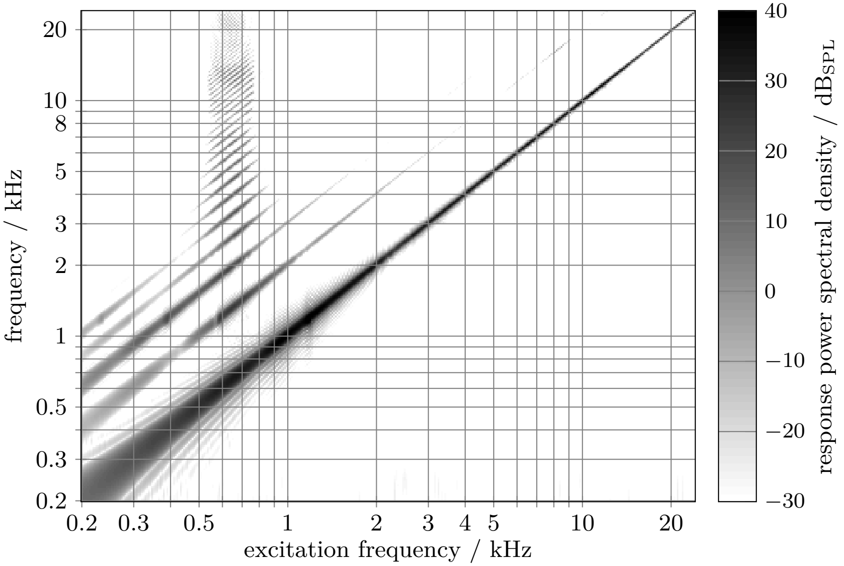
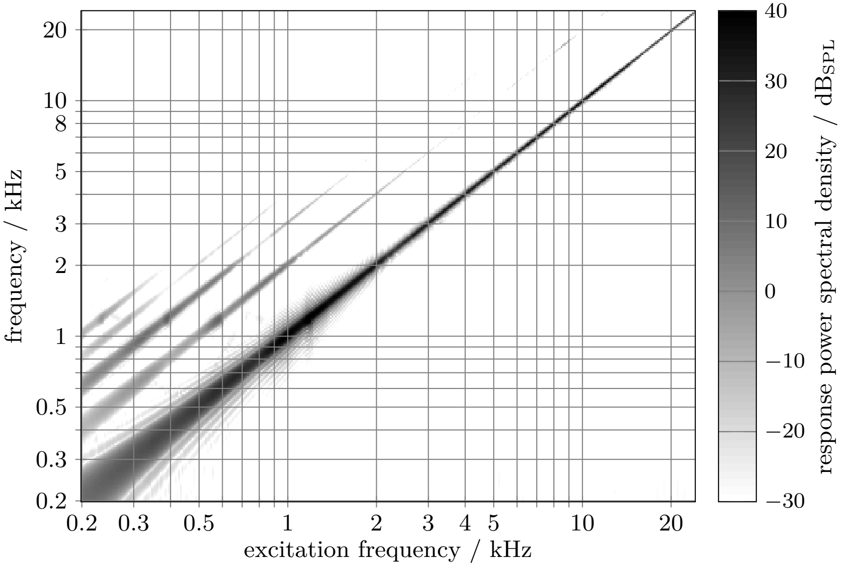


Figure 5.22: Excursion response of *NXP 13.6x9.6x2.9 speaker* at different excitation powers, with and without subband limitation (Configuration 2).

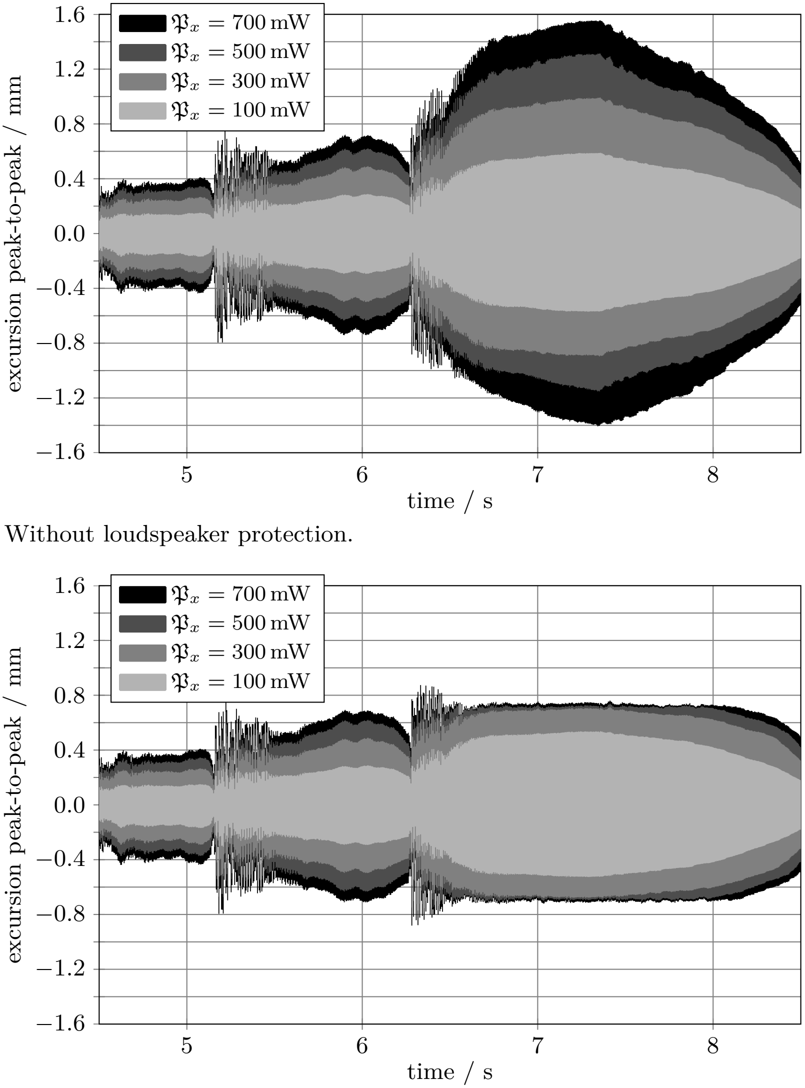


(a) Without limitation.



(b) With subband limitation (Configuration 2).

Figure 5.23: Spectrogram of response of *NXP 13.6x9.6x2.9 speaker* to exponential sweep excitation, power $\mathfrak{P}_x = \mathfrak{P}_x^{\mathrm{short}} = 700\,\mathrm{mW}$.



(a) Without loudspeaker protection.

(b) With subband limitation (Configuration 2).

Figure 5.24: Excursion of *NXP 13.6x9.6x2.9 speaker* with bass-clarinet music from EBU-SQAM compact disc as excitation with different powers $\mathfrak{P}_x$.

of the limitation. At lower frequencies, where the limitation is not as strict (see Table 5.5), the up to four harmonics remain.

Another example of the limitation of the excursion is given in Figure 5.24. In this experiment, the speaker is excited with the bass-clarinet music from the EBU-SQAM (EBU-SQAM-CD 2008, track 17; EBU-Tech 3253 2008) at different powers with and without subband limitation. It can be seen, that the subband limitation lets signals of low power or uncritical frequencies pass unchanged, but attenuates signals with excessive excursions.

Limitation of Distortion

The third configuration with combined subband and fullband limitation aims (additionally to loudspeaker protection) at reducing harmonic distortions and thus has tighter limits for the lower subbands. The excursion response at different excitation powers is shown in Figure 5.25. With this configuration, excursion is always safely below the maximum linear excursion of the *NXP 13.6x9.6x2.9 speaker* of 0.6 mm peak-to-peak.

The spectrogram as well as the THD response with and without combined subband and fullband limitation is depicted in Figure 5.26 for an excitation power of $\mathfrak{P}_x = \mathfrak{P}_x^{\text{short}} = 700\,\text{mW}$. It shows that the harmonic distortions are greatly reduced due to the combined limitation, especially at frequencies below 400 Hz to 500 Hz. Between 500 Hz and 1.5 kHz a weak second harmonic remains, which,

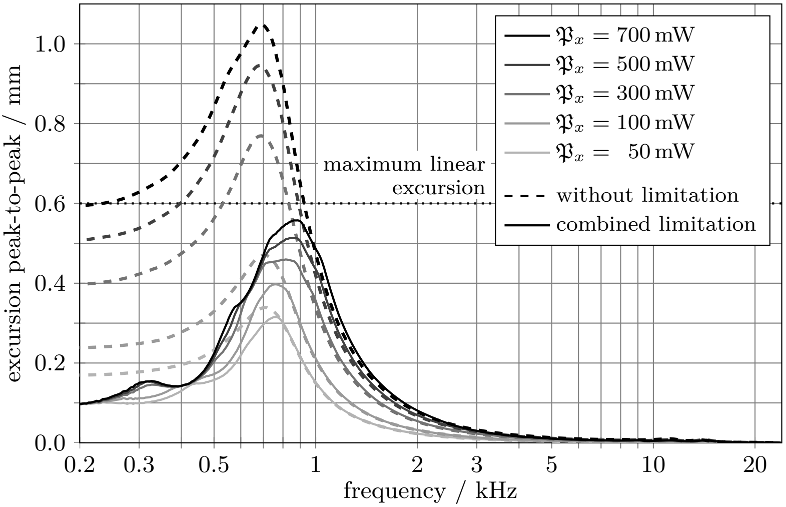
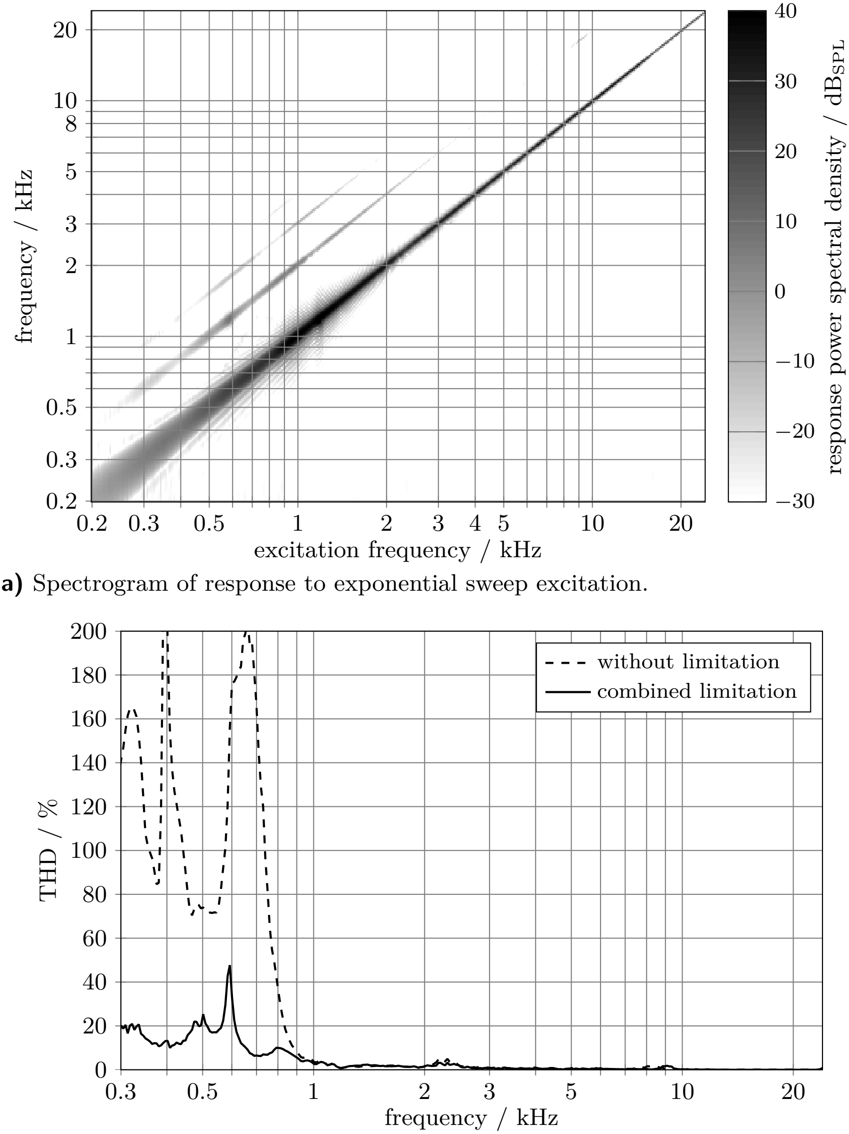


Figure 5.25: Excursion response of *NXP 13.6x9.6x2.9 speaker* at different excitation powers, with combined subband and fullband limitation (Configuration 3).



(a) Spectrogram of response to exponential sweep excitation.

(b) THD response with and without combined limitation.

Figure 5.26: Spectrogram and THD response of *NXP 13.6x9.6x2.9 speaker* with combined subband and fullband limitation (Configuration 3), power $\mathfrak{P}_x = \mathfrak{P}_x^{\mathrm{short}} = 700\,\mathrm{mW}$.

however, is only slightly audible. The reduction of harmonic distortions comes, of course, at the cost of a strong attenuation below 800 Hz at high excitation powers.

It should be noted, that the speaker housing exhibits a resonance at about 1.18 kHz which is not related to the transducer itself. This also causes the high THD values at about 400 Hz and 600 Hz, which are ascribed to the measurement setup.

5.3.5 Impact on Speech Intelligibility

In this section, the impact of the LOPRO algorithm on speech intelligibility is evaluated for the two transducers, the *NXP 13.6x9.6x2.9 speaker* and the *NXP 8x12x2 receiver*.

Configuration

In both cases, a combined subband and fullband limitation is used, which shall not only protect the transducer but also reduce harmonic distortions. This is Configuration 3 of Table 5.5 for the *NXP 13.6x9.6x2.9 speaker* and a similar configuration for the *NXP 8x12x2 receiver*, which is shown in Table 5.6. It has only $I = 5$ real-valued subbands and limits the fullband signal power to the maximum continuous power of $\mathfrak{P}_x^{\max} = 40\,\mathrm{mW}$ as specified in (NXP 2010b).

For all simulations, the model of signal flow of Figure 2.9a is used, including loudspeaker equalization and filtering with the transfer function of the transducer. The LOPRO algorithm is operating at the same sampling rate as the other parts of the simulation, which is 8 kHz in this case.

For loudspeaker equalization, a linear-phase FIR filter is used, which approximates the equalization transfer function $\mathcal{H}_i$ down to a certain cut-off frequency and uses a quadratic fit approximation according to (Hawksford 1999) below that frequency. The key parameters of the equalization filters for both transducers are given in Table 5.7.

A linear-phase FIR approximation with degree 500 of the inverse of the equalization transfer function $\mathcal{H}_i$ is used to "model" the real acoustic transfer function of the transducer in the simulation. As this filter is neither part of the LOPRO nor the NELE algorithm, this large filter degree was chosen to guarantee a low approximation error. In the simulation, the delay of this filter is compensated.

The combination of loudspeaker equalization filter and transducer transfer function filter has a quite flat response above the cut-off frequency with a magnitude within ±1 dB for the speaker and ±0.5 dB for the receiver.

NELE Parameter Settings for LOPRO

The LOPRO system applies a limitation of subband and fullband power of the loudspeaker signal to protect the transducer and reduce distortions. In theory, the NELE algorithm could consider this to improve speech intelligibility and, e. g., redistribute this power to other subbands beforehand. In practice, this is difficult as

frequencies	RMS attack time $\tau_{\mathfrak{R},\mathrm{a}}$	RMS release time $\tau_{\mathfrak{R},\mathrm{r}}$	compr. attack time $\tau_{G,\mathrm{a}}$	compr. release time $\tau_{G,\mathrm{r}}$	power limit $\mathfrak{P}_x^{\max}$
0 Hz to 100 Hz	1 ms	20 ms	5 ms	10 ms	0.2 mW
100 Hz to 300 Hz	1 ms	8 ms	5 ms	10 ms	0.1 mW
300 Hz to 500 Hz	1 ms	4 ms	5 ms	10 ms	0.1 mW
500 Hz to 700 Hz	1 ms	2 ms	5 ms	10 ms	1.9 mW
>700 Hz	—	—	—	—	—
fullband	1 s	10 s	0.1 s	0.1 s	40.0 mW

Table 5.6: Parameters of limiters for *NXP 8x12x2 receiver*.

	NXP 13.6x9.6x2.9 speaker	*NXP 8x12x2 receiver*
filter degree ($f_\mathrm{s} = 8\,\mathrm{kHz}$)	40	72
filter delay	5 ms	9 ms
cut-off frequency	400 Hz	200 Hz
magnitude at 0 Hz	25 dB	15 dB

Table 5.7: Key parameters of loudspeaker equalizer filters.

i	frequencies	$10\log\left\{\frac{P_s^{\max}}{P_0}\right\}$
1	50 Hz to 152 Hz	45 dB_SPL
2	152 Hz to 255 Hz	45 dB_SPL
3	255 Hz to 362 Hz	50 dB_SPL
4	362 Hz to 475 Hz	50 dB_SPL
5	475 Hz to 595 Hz	65 dB_SPL
6	595 Hz to 725 Hz	70 dB_SPL
7	725 Hz to 868 Hz	85 dB_SPL
≥ 8	$\geq$ 868 Hz	95 dB_SPL

Table 5.8: Adjusted maximum subband power $P_s^{\max}$ for *NXP 13.6x9.6x2.9 speaker* configuration using LOPRO.

the analysis filterbanks, the downsampling rates, and the time constants differ due to the different objectives of the algorithms. Especially, the time constants, which the LOPRO needs to effectively protect the transducer, are much smaller than the time constants, NELE can allow without impairing intelligibility by introducing fast fluctuations.

Therefore, the maximum subband powers should be adjusted only for the lower subbands, which are limited by LOPRO most of the time. The adjusted $P_s^{\max}$ for the *NXP 13.6x9.6x2.9 speaker* configuration are shown in Table 5.8.

The *NXP 8x12x2 receiver*, however, has a high sensitivity as can be seen in Figure 5.11. Accordingly, the maximum subband powers resulting in this sense from LOPRO are higher than the $95\,\mathrm{dB_{SPL}}$, which prevent hearing damage and were introduced in Section 2.2.5. Therefore, no special adjustment to LOPRO is necessary.

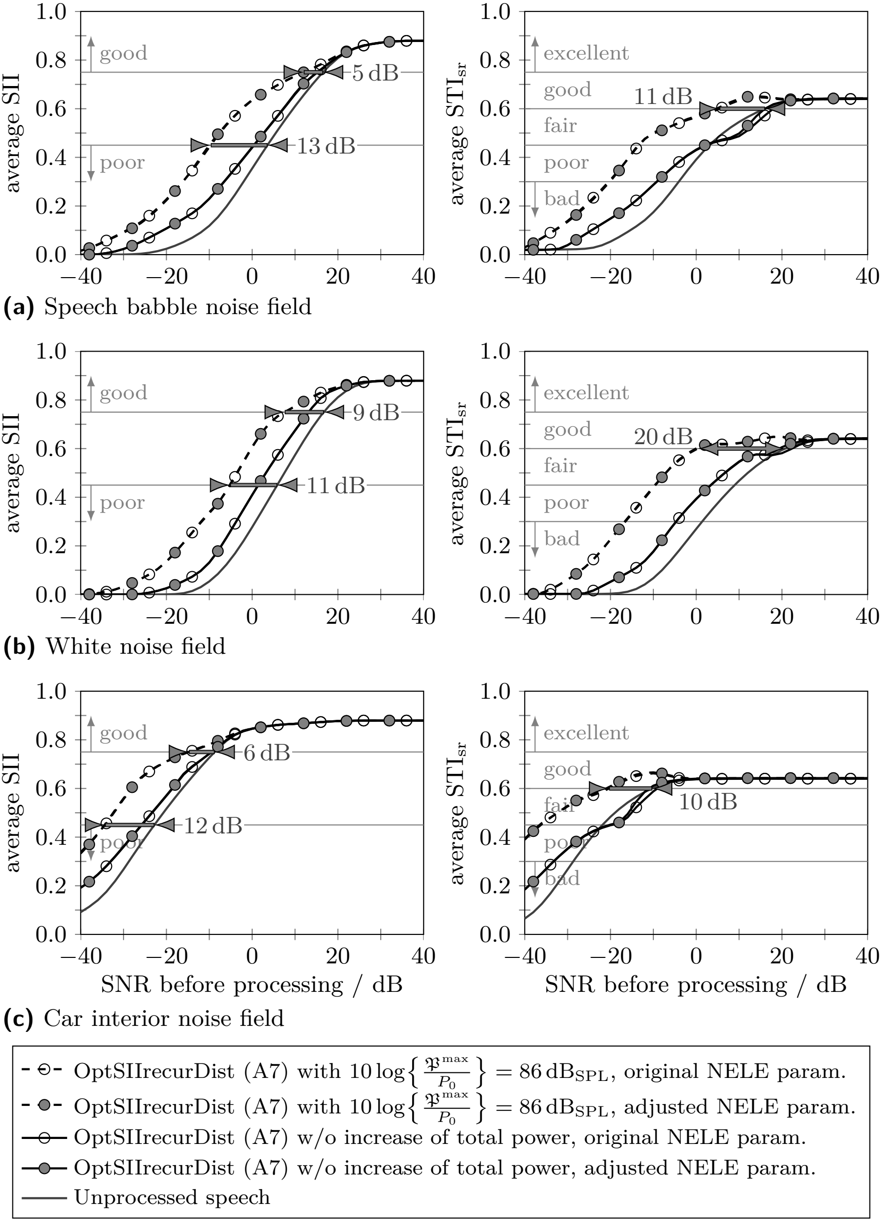
Simulation Results

Figure 5.27 shows the impact of LOPRO for the *NXP 13.6x9.6x2.9 speaker* on the average SII and $\mathrm{STI_{sr}}$ ratings. A comparison with Figures 4.9 and 4.10 reveals these main differences due to the LOPRO:

- The SII ratings at 40 dB SNR, i. e., in a virtually noise-free environment, are not affected by the limitation of the LOPRO, whereas the $\mathrm{STI_{sr}}$ ratings drop from “excellent” (0.90) to “good” (0.64). The $\mathrm{STI_{sr}}$ ratings for lower SNR are reduced accordingly.
- For the case of the increase of total power up to the thermal limit, both measures are reduced at low SNRs due to the dynamic attenuation below 800 Hz introduced by the LOPRO system. Accordingly, a speaker with lower cut-off frequency would yield better results.
- For speech babble as well as car interior noise, the SII ratings are lower at medium SNRs, i. e., the SII gain is only 5 dB to 6 dB while the benefit at lower SNR is 11 dB to 13 dB.

In general, the OptSIIrecurDist (A7) algorithm with original and with adjusted NELE parameter settings show almost identical performance.

Figure 5.28 depicts the results for the *NXP 8x12x2 receiver*. Comparing again with Figures 4.9 and 4.10 shows that the ratings are basically the same with and without the loudspeaker protection, which is to be expected due to the high sensitivity of the receiver. Only the decay due to the total power constraint occurs at lower SNRs since $\mathfrak{P}^{\max}$ can be chosen higher than $90\,\mathrm{dB_{SPL}}$.

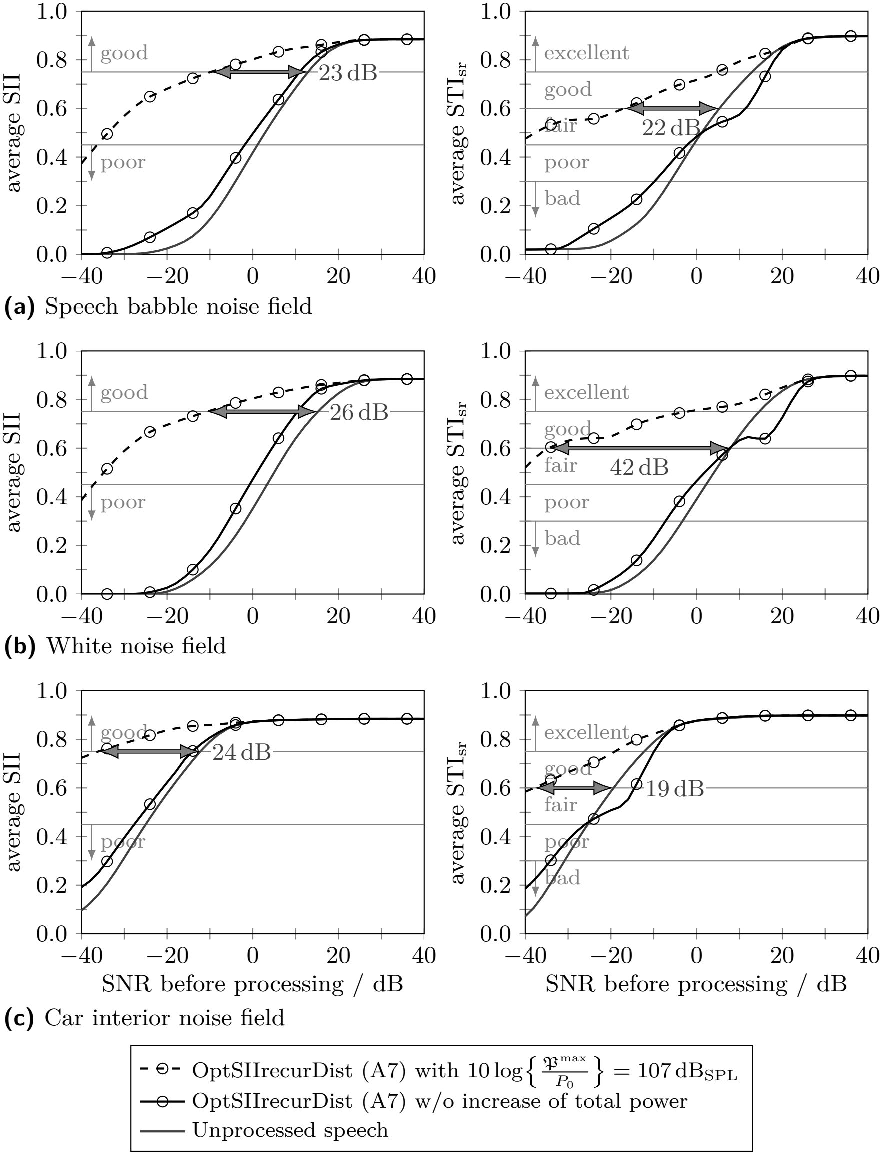


(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 5.27: Impact of loudspeaker protection for *NXP 13.6x9.6x2.9 speaker* on OptSIIrecurDist (A7). See Section 2.4 for simulation parameters. The arrows indicate the SII and STI gain.



(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure 5.28: Impact of loudspeaker protection for *NXP 8x12x2 receiver* on OptSIIrecurDist (A7). See Section 2.4 for simulation parameters. The arrows indicate the SII and STI gain.

Chapter 6

Other Applications for Near-End Listening Enhancement

Although the main focus of this thesis was on the application of NELE in mobile phones, the developed concepts can be applied in many devices, including headphones, hands-free conference terminals, car multimedia systems, public address systems, and hearing aids. While a detailed investigation of these applications scenarios is beyond the scope of this thesis, the similarities and differences are described and briefly discussed in the following.

Mobile Phone in Hands-Free Mode

In hands-free mode, the mobile phone is held and used in a variety of ways and positions. Nevertheless, the distance between mobile phone and near-end listener's head can be assumed to be about 50 cm. As a consequence, the transfer function $H_{\mathrm{ear}}(f)$ between loudspeaker and ears is in general unknown, but a *reasonable* estimate of its overall attenuation can be made. Therefore, decisions which rely on a *precise* estimate of the subband power at the ear should be avoided as in handset mode, cf. Section 2.1.2.

In contrast to the handset mode, the echo path $H_{\mathrm{echo}}(f)$ can not be neglected in hands-free mode, since the loudspeaker signal is directly fed back to the microphone of the mobile phone. Thus, the noise estimation has to be able to deal with a rather bad "noise-to-speech ratio". Some approaches to the solution are described below for car multimedia and public address systems.

Binaural Headset and ANC Headphones

The use of a binaural headset leads by itself to an increased intelligibility (Blauert 1997) and a decreased listening effort compared to the handset mode of a phone due to the binaural presentation of the speech signal. In addition, these devices enable an individual processing for each ear.

In headset mode, the mono far-end speech signal is usually presented diotically at both ears, which corresponds to a localization in the middle of the head (Licklider 1948). While this is good for diffuse noise, the speech signal could be rendered at a more beneficial position if a dominant noise direction can be determined. If the noise source is located more at one side, either left or right, shifting the speech signal to the opposite direction dramatically improves intelligibility. If the noise is "in phase", i. e., it comes from the front or the back, the speech signal should in contrast be played "out of phase", i. e., with reversed sign on one ear (Licklider 1948). This way, an improvement can be achieved simply by playing the monaural output of the NELE system just on one side respectively by inverting the sign in one channel.

For music signals, spectral shaping must be applied very carefully to avoid changing the tone color too much. Furthermore, both ears should receive a very similar weighting to retain the stereo effect. In this case, a sophisticated, noise adaptive but frequency and channel *in*dependent volume control might be the best option.

While ordinary headphones do not have a microphone installed to gather information about the ambient noise environment, closed-back active noise control (ANC) headphones are equipped with error microphones anyway. In these devices, the broadband feedback ANC works well for frequencies below 500 Hz to 700 Hz (Schumacher et al. 2011), while the closed-back can attenuate the noise above, both attenuating the noise by about 20 dB. In this case, a further improvement could be achieved by NELE considering the residual noise.

Car Multimedia System

Speed dependent volume control systems for car radios have been used for decades. Their performance can, however, be improved by NELE techniques which consider the actual noise level in the cabin and thus react on changing road surfaces and weather conditions.

Furthermore, a car multimedia system includes not only car radio, but also hands-free telephony and in-car communication. It is similar to the hands-free mode in mobile phones in the sense that speech and noise signals are perceived at both ears. But opposed to the generic hands-free mode, the position of the listener is usually quite fixed in a car, which allows better estimates of the transfer function $H_{\text{ear}}(f)$ and thus the subband power at the listener's ear.

As in all hands-free cases, the microphone signal also contains the loudspeaker signal, which complicates noise estimation. Especially with music signals, which are usually continuous and not very speech-like, the commonly used noise estimators will fail more or less completely. In this case, a basic echo canceller can remove the loudspeaker signal from the microphone input, followed by an ordinary noise estimator to disregard the speech of the passengers. Here, the echo canceller can be tuned very aggressively as its output is only used for noise estimation, where the temporal fine-structure of the subband signals is of minor interest.

Additionally, noise estimation can utilize side information from the speedometer, the revolution counter (Esch et al. 2012), the gears, the rain sensor, and other sensors.

A NELE system in car multimedia must differentiate between speech mode (hands-free telephony and in-car communication) and music mode (car radio). While in speech mode, changes of the tone color are acceptable or even beneficial, spectral shaping must be used very carefully in music mode. As described above, a noise adaptive, frequency *in*dependent volume control is advisable in this case.

Public Address System

Examples of public address systems can be found in railway stations, airports, stadiums, shopping malls, and other public buildings. The pre-recorded or live announcements, which must be broadcasted there, can be highly security relevant and thus should be as intelligible as possible. Unfortunately, these environments are often large, reverberant, and noisy. This is especially true for train platforms, where trains arrive and depart, brakes squeal, and engines run.

The loudspeakers of public address systems are usually placed every few meters in the ceiling above the listeners distributed over the whole building in different noise environments. This makes an adaptive NELE processing with local noise estimation and local reference microphones necessary. Furthermore, the distance between loudspeaker and listener is larger than in the hands-free cases discussed above and the radiated sound pressure must therefore be higher. On the one hand, this intensifies in turn the echo from the loudspeaker to the reference microphone. On the other hand, the loudspeakers can be much larger and of higher quality than in a mobile phone and thus can react more linearly at higher sound pressures.

Although double-talk is usually not a problem in public address systems, a good noise estimation is still needed for NELE as the acoustical environments are usually characterized by strong echo and reverberation. The direct echo path and the early reflections can be removed from the microphone signal before noise estimation by an echo canceller as described above. The late reverberations can be considered with an extension of the MMSE based noise PSD tracker of (Hendriks et al. 2010a) which was recently proposed in (Faraji & Hendriks 2012).

Hearing Aids

Nowadays, more and more hearing aid users can be supplied with an open-fit technique. In this case, only a small soft silicone dome is inserted into the ear channel, which leaves it as open as possible. Compared to the traditional fitted earmold, this technique reduces the so-called occlusion effect and is therefore attractive to the hearing aid user. However, the possibility of feedback is increased and the environmental background noise can pass almost unhindered to the ear drum, which reduces the SNR at the ear.

Practically all modern hearing aids contain a multiband automatic gain control which addresses the hearing loss of the user (Hamacher et al. 2005). This compressor basically performs by itself some kind of noise-independent NELE. Accordingly, it would be beneficial to integrate the noise-adaptive concepts presented in this thesis into the design of the multiband compressor instead of having a separate NELE block in serial or parallel.

Two different use cases exist for NELE in a digital hearing aid with open-fit technique:

1. enhancement of the pre-processed (i. e., noise reduced) *microphone signal*, e. g., a speech signal during a conversation in a noisy environment, and
2. enhancement of a clean audio/speech signal which is transmitted via an *audio-link* to the hearing device, e. g., from a mobile phone or a music player.

Especially in the second case, some environmental information recorded by the microphone should be added to the audio-link signal to avoid an acoustical isolation of the hearing aid user from the outside world.

Chapter 7

Summary

This thesis addresses the problem of *near-end listening enhancement* (NELE) in mobile telephony, i. e., the intelligibility improvement of a far-end speech signal which is perceived by the near-end user in local background noise environment.

Innovative solutions were developed in order to optimize intelligibility with respect to the *Speech Intelligibility Index* (SII). The SII was chosen as objective criterion due to its proven ability to predict the intelligibility of speech perceived in noise and its calculation rules, which are suitable for algorithm design. The developed methods consider for the first time the requirements and restrictions of realistic applications such as mobile phones. The noise spectrum is estimated blindly from the microphone signal, which is the only access to the acoustical environment in this case. At the same time, the utilized noise estimation algorithm disregards the voice of the near-end user in double-talk situations. A power bound in critical bands ensures that the ear of the near-end listener is protected from damage and pain.

The basic idea of the developed NELE algorithms consists of two steps: First, an optimum “speech spectrum level” in critical bands is determined which maximizes the SII under consideration of the current “disturbance spectrum level”, i. e., the spectral characteristics of the ambient noise. Then, the subband weights are calculated to achieve this optimum speech spectrum level with the far-end speech at the ear of the listener.

It is a core part of the concept to spectrally reallocate the audio power of the speech signal when necessary in order to improve intelligibility. This goes hand in hand with a moderate change of tone color of the speech through the influence of the noise spectrum. However, such a tone coloration is not perceived as distortion. If a (further) reallocation would *not* improve intelligibility, the tone color is preserved as much as possible.

Near-End Listening Enhancement ***without*** Total Power Constraint

The optimization with respect to the SII but without constraint on the total audio power led to the *bounded SII-based optimization* (OptSIIbound (A1)), which considers the power bound in each subband to prevent hearing damage. An investigation of the calculation rules of the SII revealed that a speech spectrum level of 15 dB above the disturbance spectrum level is optimal with regard to intelligibility. Accordingly, the general spectral shape of the output speech roughly follows that of the noise for low to medium SNRs, while the temporal and spectral fine-structure of speech still preserved. At high SNRs, the subband weights tend to 0 dB and thus no modification is applied in quiet environments. Simulations show that OptSIIbound (A1) yields a "good" instrumental rating for speech babble and white noise at a 23 dB to 47 dB lower input SNR than a system without processing.

In order to investigate the benefits of the frequency *dependent* enhancement, OptSIIbound (A1) was compared with a frequency *independent* amplification applying a single time-varying factor to yield the same output power as OptSIIbound (A1). As a result, the frequency *dependent* approach showed, especially for noise signals with non speech-like spectra, better instrumental intelligibility scores and a better subjective listening experience.

Near-End Listening Enhancement *with* Total Power Constraint

The derived weighting rule of OptSIIbound (A1) resembles for most mobile applications a benchmark which can only be reached with high-end loudspeakers. In mobile phones in contrast, the restrictions of the micro-loudspeakers and especially their maximum thermal load need to be considered.

In a *constrained* optimization of the SII, the total audio power was restricted to a (constant or time-adaptive) maximum power, which refers, e. g., to the maximum thermal load.

Two approaches with identical performance were presented to solve the resulting up to 21-dimensional non-linear equality constrained maximization problem: the *numerical power-constrained SII-based optimization* (OptSIInum (A3)) and the *recursive closed-form power-constrained SII-based optimization* (OptSIIrecur (A4)). For the latter, the non-linear optimization function is approximated by a linear function, which allows a closed-form solution using Lagrange multipliers. It features therefore a significantly lower computational complexity than OptSIInum (A3).

The analysis showed that for low SNRs the subband weights have a highpass characteristic up to 6 kHz. For medium SNRs, the spectral shape of the output speech roughly follows that of the noise and, at high SNRs, no modification is applied. If the total power constraint is "relaxed" enough to be inactive at medium to high SNRs, OptSIIrecur (A4) is identical to OptSIIbound (A1).

Even though the proposed algorithms maximize the SII of the output speech and demonstrably increase intelligibility in various noise environments, they might lead under certain conditions, e. g., for extreme noise types with a narrow bandpass spectrum and a tight power constraint, to a reduced subjective quality and lower speech intelligibility due to extreme "coloring effects". For this problem two solutions have been found in terms of the *a priori limitation of the disturbance spectrum level* (OptSIIrecurDist (A7)) and the *one-step closed-form power-constrained SII-based optimization* (OptSIIone (A8)), where the latter interpolates between the optimum highpass weights at low SNRs and 0 dB weights for high SNRs. OptSIIone (A8) has the slightly better performance than OptSIIrecurDist (A7), but is only applicable if the output power may not exceed the input power, whereas OptSIIrecurDist (A7) works for any power constraint.

Instrumental objective evaluations showed a distinct intelligibility improvement of the proposed algorithms. In two large scale subjective listening tests with natural and synthetic speech, the word recognition rate was enhanced without increasing signal power in low and mid SNR conditions by up to 22 percentage points and improved from 85.8 % to 90.5 % for a high SNR.

Loudspeaker Protection

The micro-loudspeakers of modern mobile phones are often driven at their limits to satisfy the need for a loud sound reproduction. Accordingly, a *loudspeaker protection* (LOPRO) system is necessary to prevent accidental overheating of the loudspeaker and damage due to excessive excursions of the membrane.

Acoustic distortions, membrane excursion, and progress of temperature were experimentally studied in measurement campaigns with two commonly used micro-loudspeakers, the *NXP 13.6x9.6x2.9 speaker* and the *NXP 8x12x2 receiver*. The target of these experiments was to derive a simple model suitable for the LOPRO algorithm design. In addition to the well-known measure *total harmonic distortion* (THD), which is only defined for single sine waves as input, two other measures were used in this thesis: the *total intermodulation distortion* (TID) for mixtures of two sines and the *total non-linear distortion* (TND) for bandpass signals.

A LOPRO scheme for mobile phones was derived, which consists of an individual subband time-domain limitation in three to five subbands below 1 kHz. This considers the frequency dependent excursion of the loudspeaker membrane as well as the frequency dependent amplification during loudspeaker equalization. After re-synthesizing the fullband signal, a final fullband time-domain limiter takes care of the maximum total power that the voice coil of the loudspeaker can stand.

Based on the *generalized sliding DFT*, a highly efficient and perfect reconstructing filterbank was developed, which took the very short time constants into account that are essential for LOPRO.

Three configurations for temperature control, limitation of membrane excursion, and distortion reduction were tested with the *NXP 13.6x9.6x2.9 speaker*. These experiments at maximum short-term power finally verified the effectiveness of the proposed LOPRO concept. In simulations, the impact of this scheme on speech intelligibility was studied, also yielding that NELE and LOPRO interact properly.

In this thesis, algorithms for near-end listening enhancement (NELE) have been presented which improve the intelligibility of the far-end speech signal perceived in near-end acoustical background noise. In contrast to state-of-the-art algorithms from literature, the developed approaches estimate all noise information blindly, can cope with double-talk situations, behave transparently in noise-free environments, and prevent hearing damage of the listener.

It was shown, that the new concepts can also be applied in many different devices such as mobile phones, headphones, hands-free conference terminals, car multimedia systems, public address systems, and hearing aids.

Although the presented algorithms for near-end listening enhancement were driven by the application perspective, this thesis also includes the derivation of theoretical bounds, instrumental measures, and auditory evaluations. As a result, significant improvements of speech intelligibility under adverse acoustical conditions can be achieved with the proposed techniques. Their practical significance is also reflected by the fact that some of the presented concepts already became part of the product platform of well-known mobile phone manufacturers.

Appendix A

Choice of Algorithmic Parameters for Near-End Listening Enhancement

In this appendix, the influence of various parameters of the framework for NELE, which is described in Section 2.2, on the objective performance of the SII-based optimization is evaluated.

OptSIIbound (A1) (Section 3.2.1) is chosen as algorithm without total audio power constraint and OptSIIrecurDist (A7) (Section 4.2.4) as representative of NELE with the strict power constraint to the total input power.

It should be noted, that each parameter is examined individually even though there might be some cross-dependency between the parameters, especially for small numbers of subbands.

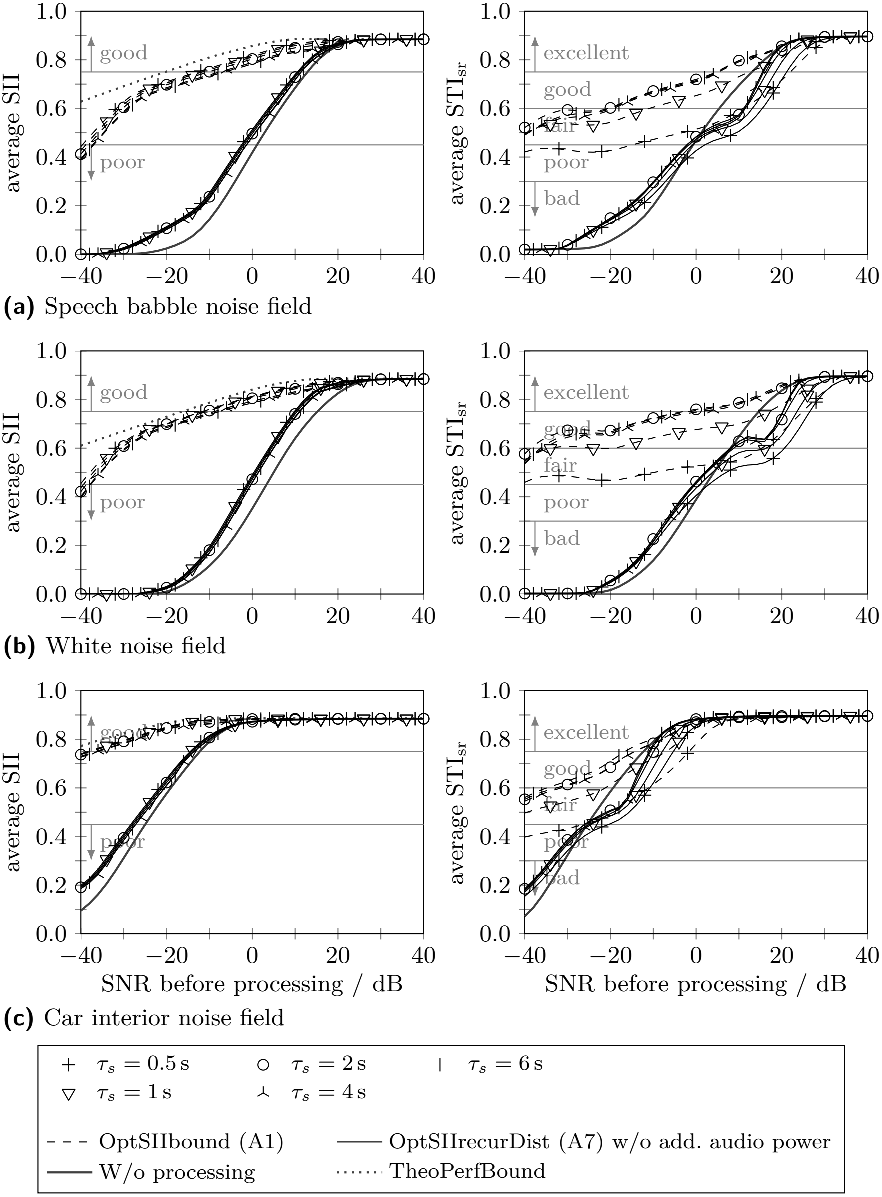
Influence of Speech Subband Power Estimator Buffer Length

The short-term subband power estimate of the far-end speech signal is calculated as the arithmetic mean of the squared, normalized magnitudes of the subband signals during the preceding $\tau_s \cdot \frac{f_\mathrm{s}}{R}$ update intervals of length R with voice activity, as described in Section 2.2.3.

Figure A.1 shows the dependency of the performance of the SII-based optimizations on the duration τ_s, which determines the memory of the speech subband power estimator. It can be seen, that all examined durations between 0.5 s and 4 s have a comparable performance in terms of SII. Differences occur mainly due to a slightly different average amplification. However, short memories of 1 s and less have a progressively worse STI_sr. For memories longer than 2 s, the system adapts increasingly slower to changes in intensity and spectral envelope of the far-end signal, which is, however, not tested in the simulation. Accordingly, $\tau_s = 2\,\mathrm{s}$ seems to be a reasonable setting.

Influence of Noise Subband Power Estimation Algorithm

Figure A.2 shows the dependency of the noise subband power estimation algorithm on the performance of the SII-based optimizations, cf. Section 2.2.4. The Minimum Statistics algorithm (Martin 2001, 2006), an MMSE based noise PSD tracking algorithm (Hendriks et al. 2010a), and a simple moving average algorithm with $\tau_n = 0.5\,\mathrm{s}$ are compared.

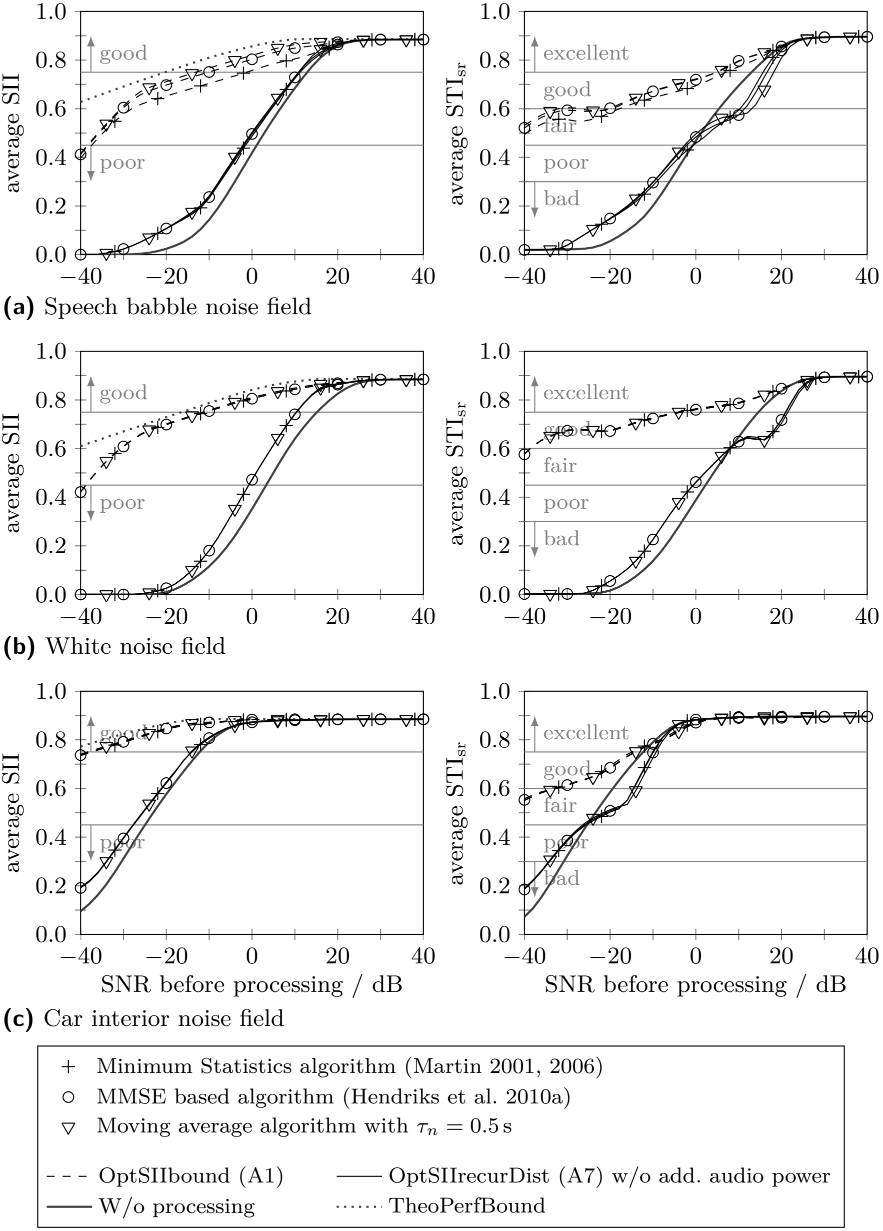


(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure A.1: Influence of speech subband power estimator buffer length τ_s on SII-based optimizations. See Section 2.4 for other simulation parameters.



(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure A.2: Influence of noise subband power estimation on SII-based optimizations. See Section 2.4 for other simulation parameters.

For quasi-stationary noise signals and especially for white noise, all tested noise subband power estimation algorithms show basically the same performance. In contrast, the performance of the Minimum Statistics approach degrades for speech babble and mid-range SNRs.

In general, the MMSE based algorithm tends to track non-stationary noise and speech babble noise better and faster than the Minimum Statistics algorithm. Furthermore, it seems to cope better with interfering near-end speech. Therefore, the MMSE based algorithm is used in this thesis.

The simple moving average algorithm shows a performance comparable to the MMSE base algorithm for all noise signals, including babble noise. However, it interprets in double-talk situations the interfering speech signal of the near-end user as noise, cf. Section 2.1. Thus, it is not suitable for most real-world applications.

Influence of Downsampling Rate

As shown in Figure A.3, the influence of the downsampling rate R, i. e., the update interval of the subband weights, on the SII performance is negligible in the evaluated range. The STI_{sr} scores are slightly better for an update interval corresponding to 40 ms, since the signal is modified more smoothly in this case and less modulation is introduced.

Higher downsampling rates, however, lead to a slower reaction of the system and, therefore, an update interval corresponding to 10 ms, i. e., $R = 80$ at 8 kHz sampling rate is used throughout this thesis.

Influence of Number of Subbands

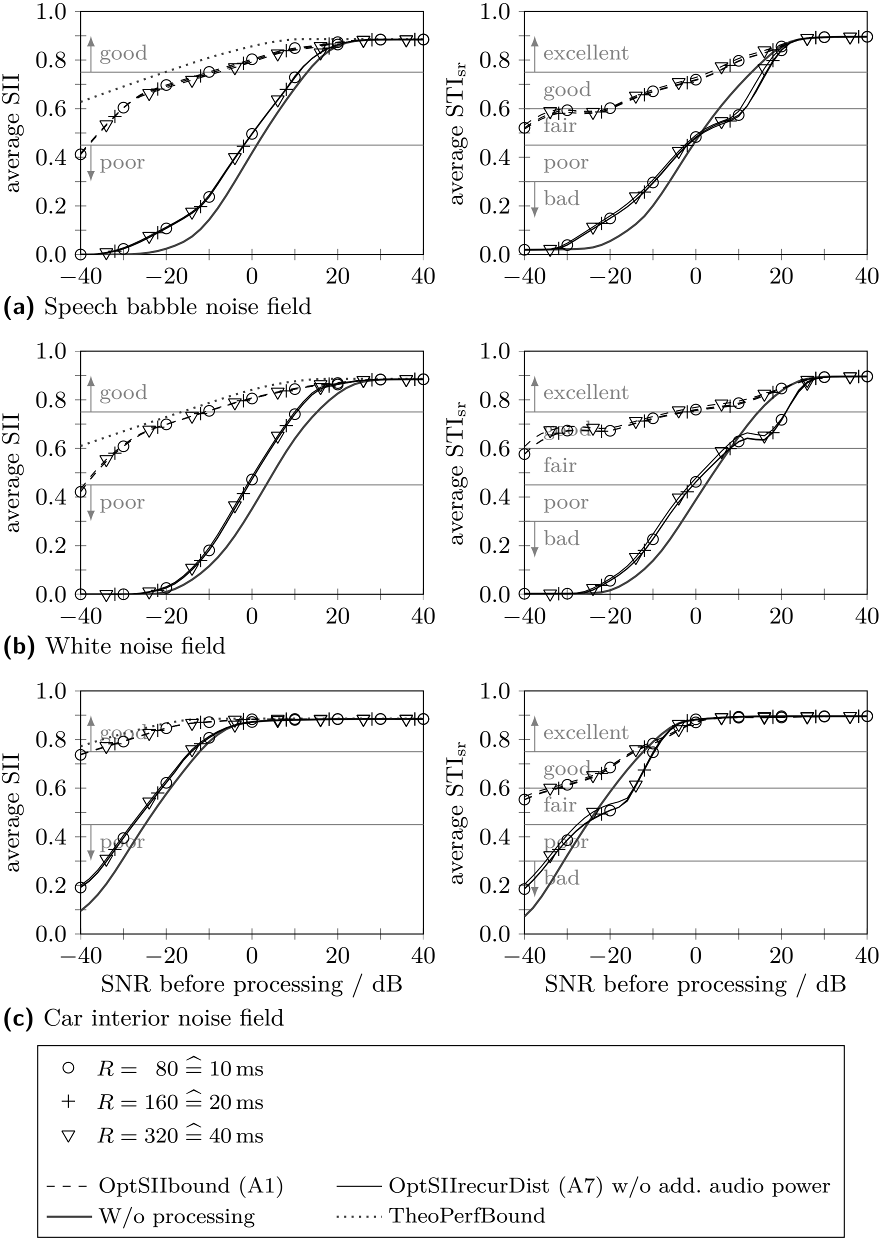
A good approximation of the Bark frequency scale is yielded for the sampling rate $f_{\text{s}} = 8\,\text{kHz}$ with the non-uniform filterbank equalizer (FBE), $M = 34$ subbands, and an allpass pole of $a = 0.4$. However, an optimized calculation of the DFT, the radix-2 FFT, can be used if the number of subbands is a power of two. Therefore, Figure A.4 shows the influence of number of subbands on the performance of the SII-based optimizations.

In general, the average SII increases with *increasing* number of subbands, but saturates for $M > 32$. In contrast, the STI_{sr} generally increases with *decreasing* number of subbands and levels off for $M < 64$. Both objective measures show identical ratings for $M = 32$ and $M = 34$, while the former allows a computational more efficient implementation.

Accordingly, $M = 32$ is a reasonable compromise for implementations in real devices. Nevertheless, $M = 34$ is used in this thesis as it provides a better approximation of the Bark frequency scale.

Influence of Length of Prototype Filter

The analysis prototype filter length of the FBE is denoted by L as described in Section 2.2.1. Reasonable choices for L are multiples of the DFT size M.

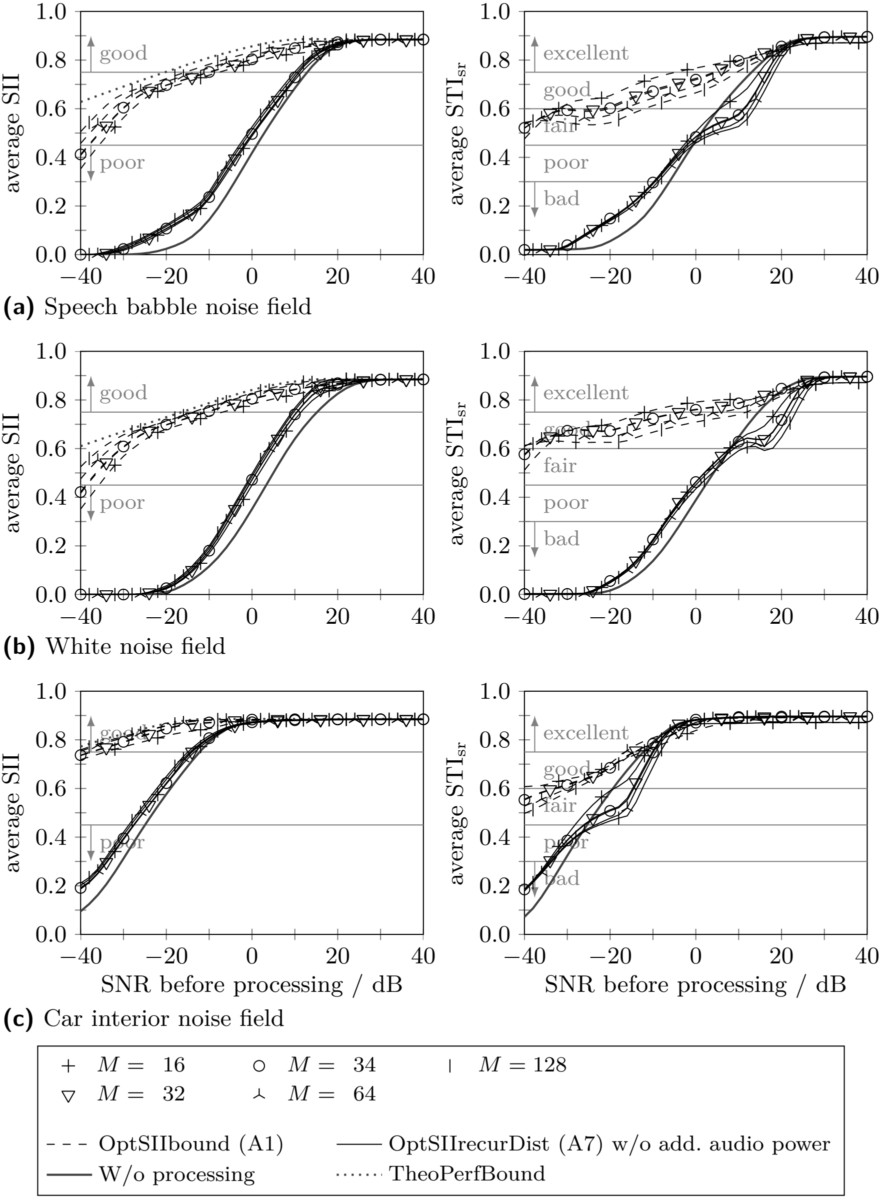


(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure A.3: Influence of downsampling rate R on SII-based optimizations. See Section 2.4 for other simulation parameters.

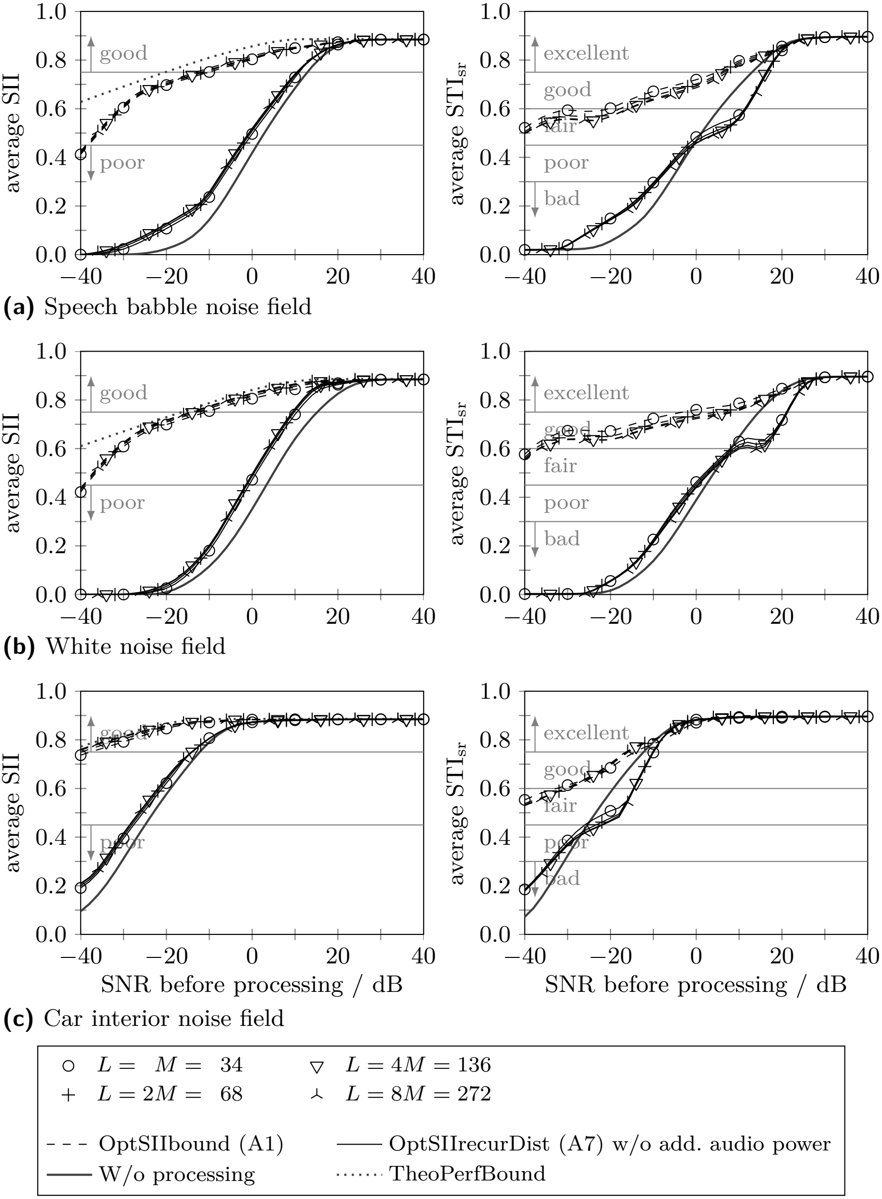


(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure A.4: Influence of number of subbands M on SII-based optimizations. See Section 2.4 for other simulation parameters.



(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure A.5: Influence of length of prototype filter L on SII-based optimizations. See Section 2.4 for other simulation parameters.

On the one hand, higher prototype filter degrees result in steeper filters and thus a higher frequency selectivity. On the other hand, a higher degree also entails a higher algorithmic delay of the time-domain filter unless a low delay filter variant is used.

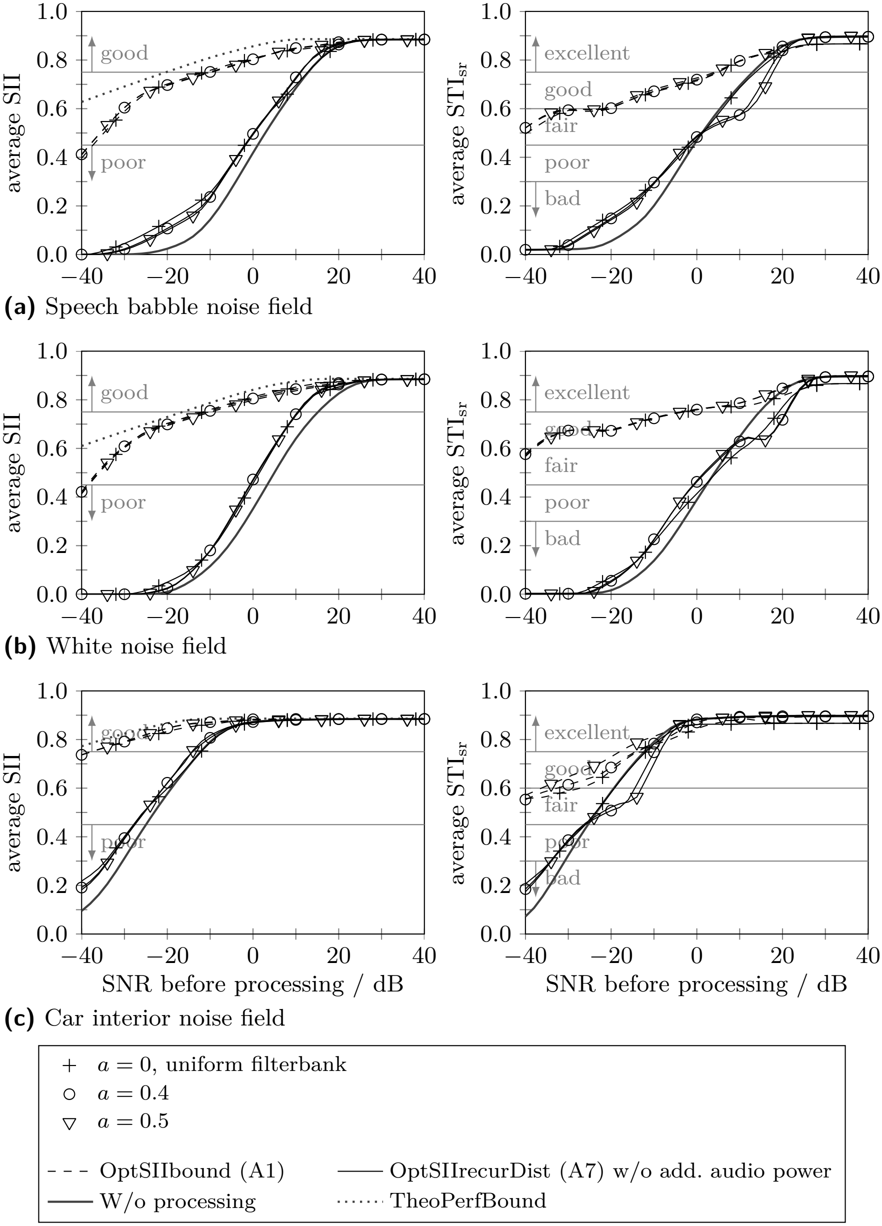
It can be seen in Figure A.5, that the length of the prototype filter has only minor influence on the performance of the SII-based optimizations. For speech babble and white noise, the average $\mathrm{STI}_{\mathrm{sr}}$ is slightly better for smaller prototype filter degrees, probably due to the weaker frequency selectivity. Thus, $L = M$ is used in this thesis.

Influence of Allpass Coefficient

According to (Smith & Abel 1999), an allpass pole of $a \approx 0.4$ yields the best approximation of the Bark frequency scale for the sampling rate of $f_{\mathrm{s}} = 8\,\mathrm{kHz}$. However, for $a = 0.5$, the multiplication in the denominator of (2.23) becomes a more efficient shifting operation. For $a = 0$, i. e., for a uniform filterbank, the allpass chain simplifies to a buffering, which is even more efficient.

Figure A.6 shows the performance of the SII-based optimizations for these three allpass coefficients. Both warped filterbanks have basically the same performance. While the uniform filterbank results in about the same average SII, the average $\mathrm{STI}_{\mathrm{sr}}$ is worse than for the warped case at medium and high SNRs. This last finding is supported by informal listening tests, which showed a preference of the warped FBE over the uniform AS FB (Sauert et al. 2008).

In this thesis, the optimum allpass poles according to (Smith & Abel 1999) are chosen, which are $a \approx 0.40$ at $f_{\mathrm{s}} = 8\,\mathrm{kHz}$ and $a \approx 0.58$ at $f_{\mathrm{s}} = 16\,\mathrm{kHz}$.



(a) Speech babble noise field

(b) White noise field

(c) Car interior noise field

Figure A.6: Influence of allpass coefficient a on SII-based optimizations. See Section 2.4 for other simulation parameters.

Mathematical Notation & Abbreviations

Mathematical Notation

In this thesis, the following conventions are used to denote quantities: vectors are underlined, e. g., $\underline{x}$, scalar values are not, e. g., x. Estimated or approximated variables are marked with a hat, e. g., $\hat{x}$, and averaged or smoothed values are denoted with a bar, e. g., $\bar{x}$.

Time-domain signals are written in lower-case letters, e. g., $x(k)$ with the sample index k. The complex-valued DFT coefficients are labeled with the calligraphic upper-case letters, e. g., $\mathcal{X}_\mu(\kappa)$ with DFT bin index $\mu \in \{0, 1, \ldots, M-1\}$, even DFT size M, and update index κ. The corresponding real-valued subband signals with subband index $i \in \{0, 1, \ldots, \frac{M}{2}\}$ are denoted by lower-case letters with the subband index as subscript, e. g., $x_i(k)$. The short-term subband power is written with the (normal) upper-case letter “P”, e. g., $P_{x,i}(\kappa)$, whereas the upper-case Fraktur letter, e. g., $\mathfrak{P}_x$, represents the total power of the fullband signal.

The vector of the spectrum levels E_i in all contributing subbands $i_{\mathrm{f}} \leq i \leq i_{\mathrm{l}}$ is denoted by $\underline{E} = (E_{i_{\mathrm{f}}}, E_{i_{\mathrm{f}}+1}, \ldots, E_{i_{\mathrm{l}}})$. In contrast, $\underline{E}_{\setminus i_{\mathrm{f}}} = (E_{i_{\mathrm{f}}+1}, E_{i_{\mathrm{f}}+2}, \ldots, E_{i_{\mathrm{l}}})$ represents the sliced vector of the spectrum levels E_i in all contributing subbands but i_{f}.

Mathematical Operators

$\approx$	approximately equal to
$\widehat{=}$	equivalent to (usually a unit conversion)
$\overset{!}{=}$ / $\overset{!}{\leq}$	shall be equal to / shall be less than or equal to
$\wedge$ / $\vee$	logical and / or
$\in$	element of
$\forall$	for all
x^*	complex conjugate of x
$\lvert x \rvert$	absolute value of x
$\lfloor x \rfloor$	floor function, i. e., largest integer which is not greater than x
$\lceil x \rceil$	ceiling function, i. e., smallest integer which is not less than x

$\mathrm{E}\{x(k)\}$	expectation value of $x(k)$
$\mathrm{Re}\{x\}$	real part of x
$\mathrm{Im}\{x\}$	imaginary part of x
$\exp\{x\}$	exponential function e^x
$\log\{x\}$	logarithm of x to base 10
$\max\limits_x\{f(x)\}$	maximum of $f(x)$ over x
$\arg\max\limits_x\{f(x)\}$	argument x of maximum of $f(x)$ over x
$\underset{x}{\mathrm{mean}}\{f(x)\}$	average of $f(x)$ over all x of a finite set

Principal Symbols

$\alpha_{G,\mathrm{a}}$	smoothing factor for increasing limitation (def. on p. 124)
$\alpha_{G,\mathrm{r}}$	smoothing factor for decreasing limitation (def. on p. 124)
α_i	octave-weighting factor for STI$_{\mathrm{sr}}$ calculation (def. on p. 32)
$\alpha_{\mathfrak{R},\mathrm{a}}$	smoothing factor for ascending RMS (def. on p. 123)
$\alpha_{\mathfrak{R},\mathrm{r}}$	smoothing factor for descending RMS (def. on p. 123)
β_i	redundancy correction factor for STI$_{\mathrm{sr}}$ calculation (def. on p. 32)
$\breve{\beta}_i$	normalization term for STI$_{\mathrm{sr}}$ calculation (def. on p. 31)
$\gamma(\kappa)$	parameter of OptSIIone (A8) (def. on p. 72)
$\mathit{\Gamma}_i$	factor in OptSIIrecur (A4) (def. on p. 58)
$\updelta(k)$	unit impulse sequence (def. on p. 19)
ε	unified attenuation for reduction of tone color change (def. on p. 60)
ζ	summation index
η	index of harmonic (def. on p. 98)
θ	direction of arrival (def. on p. 7)
κ	time index in subsampled domain (def. on p. 20)
λ	Lagrange multiplier (def. on p. 58)
$\lambda_{\breve{x},i}$	variance of probe envelope signal in STI$_{\mathrm{sr}}$ calculation (def. on p. 31)
$\lambda_{\breve{x}\breve{y},i}$	covariance between probe & response envelope signal (def. on p. 31)
μ	DFT bin index (def. on p. 20)
μ_0	frequency band shift for oddly-stacked DFT (def. on p. 117)
μ_{f}	first DFT bin index of excitation band (def. on p. 100)
μ_{l}	last DFT bin index of excitation band (def. on p. 100)
$\mu_{\breve{x},i}$	mean of probe envelope signal for STI$_{\mathrm{sr}}$ calculation (def. on p. 31)
$\mu_{\breve{y},i}$	mean of response envelope signal for STI$_{\mathrm{sr}}$ calculation (def. on p. 31)
$\mu_{\breve{z},i}$	mean of noise envelope signal for STI$_{\mathrm{sr}}$ calculation (def. on p. 31)
$\xi_{\mathrm{b},i}$	begin of the quadratic segment of $A_i(E_i, D_i)$ (def. on p. 45)

$\xi_{\mathrm{e},i}$	end of the quadratic segment of $A_i(E_i, D_i)$ (def. on p. 45)
$\tau_{G,\mathrm{a}}$	attack time constant for smoothed limiter gain (def. on p. 124)
$\tau_{G,\mathrm{r}}$	release time constant for smoothed limiter gain (def. on p. 124)
τ_n	length of the buffer for estimation of $\hat{P}_{n,i}(\kappa)$ (def. on p. 25)
$\tau_{\mathfrak{R},\mathrm{a}}$	attack time constant for smoothed RMS (def. on p. 123)
$\tau_{\mathfrak{R},\mathrm{r}}$	release time constant for smoothed RMS (def. on p. 123)
τ_s	length of the buffer for estimation of $\hat{P}_{s,i}^{\mathrm{in}}(\kappa)$ (def. on p. 24)
υ	recursion step index (def. on p. 58)
$\mathit{\Upsilon}$	number of recursion steps (def. on p. 59)
$\hat{\mathit{\Phi}}_{yy,\mu}$	estimate of PSD of noise-free microphone signal (def. on p. 100)
$\mathit{\Psi}$	number of subbands contributing to $\overline{\psi}(\kappa)$ (def. on p. 72)
$\overline{\psi}(\kappa)$	average signal-to-disturbance ratio (def. on p. 72)
ψ_{b}	begin of transition range of OptSIIone (A8) (def. on p. 74)
ψ_{e}	end of transition range of OptSIIone (A8) (def. on p. 74)
$\psi_i(\kappa)$	signal-to-disturbance ratio (SDR) (def. on p. 72)
$\mathit{\Omega}$	normalized frequency (def. on p. 16)
a	allpass coefficient (def. on p. 23)
$A_i(E_i, D_i)$	band audibility function (def. on p. 29)
$\hat{A}_i(E_i, D_i)$	approximation of band audibility function (def. on p. 56)
$aSNR_i$	apparent SNR for $\mathrm{STI}_{\mathrm{sr}}$ calculation (def. on p. 32)
$C_i(N_i)$	slope per octave of masking spread for SII calculation (def. on p. 28)
D_Δ	distance between $\overline{D}^{(\upsilon)}$ and threshold (def. on p. 69)
D_i	disturbance spectrum level (def. on p. 28)
D_i'	limited disturbance spectrum level (def. on p. 69)
$\overline{D}^{(\upsilon)}$	average limited disturbance spectrum level (def. on p. 69)
e	Euler's number
E_i	speech spectrum level (def. on p. 26)
$E_i^{(\upsilon)}$	speech spectrum level after υ-th step (def. on p. 58)
E_i^{adm}	upper limit of *admissible range* (def. on p. 55)
E_i^{in}	input speech spectrum level (def. on p. 42)
E_i^{max}	maximum allowed output speech spectrum level (def. on p. 46)
E_i^{opt}	optimum speech spectrum level (def. on p. 41)
E_i^{out}	output speech spectrum level (def. on p. 42)
f	continuous frequency (def. on p. 7)
$f_{\Delta,i}$	frequency bandwidth of the i-th subband (def. on p. 26)
$f_{\mathrm{c},i}$	center frequency of the i-th subband (def. on p. 26)
$f_{\mathrm{h},i}$	upper limiting frequency of the i-th subband (def. on p. 24)
$f_{\mathrm{l},i}$	lower limiting frequency of the i-th subband (def. on p. 24)

f_s	sampling rate (def. on p. 16)
$G(\kappa)$	gain of time-domain limiter (def. on p. 123)
$\overline{G}(\kappa)$	smoothed gain of time-domain limiter (def. on p. 124)
g_fb	normalization factor of analysis filterbank (def. on p. 19)
g_ls	proportionality factor in loudspeaker path (def. on p. 123)
g_mic	proportionality factor in microphone path (def. on p. 18)
$g_{\mathrm{sym},i}$	symmetry factor of the DFT (def. on pp. 20, 118)
$h(l)$	window function (def. on p. 19)
$\underline{H}$	window coefficients (def. on p. 120)
$H_\mathrm{A}(z)$	frequency response of allpass filter (def. on p. 22)
$H_\mathrm{ear}(f)$	transfer function from loudspeaker to ear (def. on p. 7)
$H_\mathrm{echo}(f)$	echo path from loudspeaker to microphone (def. on p. 7)
$\mathcal{H}_i$	frequency response of loudspeaker equalization (def. on p. 101)
$\underline{H}'_i$	modified window coefficients for the i-th subband (def. on p. 121)
$\overline{H}_\mathrm{leak}(f)$	average acoustic leakage from noise source to ear (def. on p. 12)
$H_{\mathrm{leak},\theta}(f)$	acoustic leakage from noise source with DOA θ to ear (def. on p. 7)
$H_\mathrm{ls}(f)$	transfer function of loudspeaker (def. on p. 7)
$\overline{H}_\mathrm{match}(f)$	average magnitude response from microphone to ear (def. on p. 8)
$\hat{H}_\mathrm{match}(\mathit{\Omega})$	estimate of magnitude response from mic. to ear (def. on p. 16)
$H_{\mathrm{match},\theta}(f)$	magnitude response from microphone to ear (def. on p. 8)
$H_\mathrm{mic}(f)$	transfer function of microphone and A/D conversion (def. on p. 7)
$\overline{H}_\mathrm{noise}(f)$	average magnitude response from noise to mic. (def. on p. 12)
$H_{\mathrm{noise},\theta}(f)$	transfer function from noise with DOA θ to mic. (def. on p. 7)
$H_\mathrm{nr}(f)$	freq. response of noise reduction in MaxTransfer (A6) (def. on p. 64)
$h_\mathrm{s}(l,\kappa)$	coefficients of single time-domain filter of FBE (def. on p. 22)
$H_\mathrm{speech}(f)$	transfer function from mouth of near-end user to mic. (def. on p. 7)
i	subband index (def. on p. 20)
I	number of subbands (def. on p. 116)
$I^{(\upsilon)}$	number of subbands with $D_i > \overline{D}^{(\upsilon)} + D_\Delta$ (def. on p. 69)
$I_{\mathrm{am},i}$	intensity of auditory masking for STI$_\mathrm{sr}$ calculation (def. on p. 32)
i_f	first contributing subband (def. on p. 23)
I_i	band importance function for SII calculation (def. on p. 30)
i_l	last contributing subband (def. on p. 23)
$I_{\mathrm{rs},i}$	intensity at absolute threshold for STI$_\mathrm{sr}$ calculation (def. on p. 32)
$I_{\breve{y},i}$	intensity of response signal for STI$_\mathrm{sr}$ calculation (def. on p. 32)
j	imaginary unit
k	sample index (def. on p. 14)
K_1	constant for MaxTransfer (A6) (def. on p. 65)

K_2	constant for MaxTransfer (A6) (def. on p. 65)
$K_i(E_i, D_i)$	auxiliary variable for SII calculation (def. on p. 29)
$\mathbb{K}_s(\kappa)$	set of time indices for estimation of $\hat{P}^{\text{in}}_{s,i}(\kappa)$ (def. on p. 24)
$\mathbb{K}_n(\kappa)$	set of time indices for estimation of $\hat{P}_{n,i}(\kappa)$ (def. on p. 25)
l	filter tap (def. on p. 19)
L	size of prototype filter of analysis filterbank (def. on p. 19)
l_0	delay of time-domain filter of generalized DFT (def. on p. 22)
$L_i(E_i)$	speech level distortion factor for SII calculation (def. on p. 28)
m	frame index of frames with voice activity (def. on p. 27)
M	number of DFT bins (def. on p. 10)
M_i	modulation metric for STI_{sr} calculation (def. on p. 31)
M_i'	corrected modulation metric for STI_{sr} calculation (def. on p. 32)
$\mathbb{M}_{f_1,f_2}$	set of intermodulation DFT indices (def. on p. 99)
$\mathbb{M}_i$	set of DFT indices for calculation of subband power (def. on p. 27)
$n(k)$	near-end noise signal at the listener's ear (def. on p. 18)
$\mathcal{N}_\mu(\kappa)$	DFT coefficients of near-end noise signal (def. on p. 25)
$\hat{\mathcal{N}}_\mu$	estimated DFT coefficients of noise signal (def. on p. 100)
N_i	noise spectrum level (def. on p. 26)
$\mathbb{N}$	set of positive integers
$\mathbb{N}_0$	set of non-negative integers
p_0	reference sound pressure of 20 µPa (def. on p. 18)
P_0	reference power of 20 µPa (def. on p. 18)
p_i	root-time-mean-square sound pressure (def. on p. 18)
$\hat{P}_{n,i}(\kappa)$	estimate of noise subband power (def. on p. 20)
$\hat{P}_n^{\min}(\kappa)$	noise floor for estimate of noise subband power (def. on p. 65)
$\hat{\overline{P}}_s(\kappa)$	average estimate of speech subband power (def. on p. 65)
$\hat{P}_{s,i}(\kappa)$	estimate of speech subband power (def. on p. 27)
$\hat{P}^{\text{in}}_{s,i}(\kappa)$	estimate of input speech subband power (def. on p. 20)
$P^{\text{opt}}_{s,i}(\kappa)$	optimum output speech subband power (def. on p. 61)
$\hat{P}^{\text{out}}_{s,i}(\kappa)$	estimate of output speech subband power (def. on p. 42)
$P_s^{\max}$	maximum allowed output speech subband power (def. on p. 26)
$P_{y,\eta}$	power of η-th harmonic of microphone signal (def. on p. 98)
$\mathfrak{P}^{\max}(\kappa)$	maximum allowed total audio power (def. on p. 53)
$\mathfrak{P}^{\max,(\upsilon)}(\kappa)$	remaining power budget in OptSIIrecur (A4) (def. on p. 58)
$\mathfrak{P}_x(\kappa)$	electric power of loudspeaker signal (def. on p. 98)
$\mathfrak{P}_x^{\text{cont}}$	maximum continuous power of transducer (def. on p. 93)
$\mathfrak{P}_x^{\text{long}}$	maximum long-term power of transducer (def. on p. 93)

$\mathfrak{P}_x^{\mathrm{max}}$	maximum electric power of loudspeaker signal (def. on p. 123)
$\mathfrak{P}_x^{\mathrm{short}}$	maximum short-term power of transducer (def. on p. 93)
R	downsampling rate (def. on p. 20)
$\mathfrak{R}_x(\kappa)$	RMS of loudspeaker signal (def. on p. 123)
$\overline{\mathfrak{R}}_x(\kappa)$	smoothed RMS of loudspeaker signal (def. on p. 123)
$s(k)$	speech signal (def. on p. 27)
$S(\underline{E}, \underline{D})$	Speech Intelligibility Index (def. on p. 29)
$\hat{S}(\underline{E}, \underline{D}, \lambda)$	optimization function in OptSIIrecur (A4) (def. on p. 57)
$\tilde{S}(\underline{D})$	theoretical bound for Speech Intelligibility Index (def. on p. 46)
$s^{\mathrm{in}}(k)$	input speech signal (def. on p. 14)
$s^{\mathrm{out}}(k)$	output speech signal (def. on p. 14)
$\tilde{s}^{\mathrm{out}}(k)$	noise reducted speech signal in MaxTransfer (A6) (def. on p. 64)
$\mathcal{S}_\mu(\kappa)$	DFT coefficients of speech signal (def. on p. 27)
$\mathcal{S}_\mu^{\mathrm{in}}(\kappa)$	DFT coefficients of input speech signal (def. on p. 20)
$s_i^{\mathrm{in}}(k)$	subband input speech signal (def. on p. 20)
$s_i^{\mathrm{out}}(k)$	subband output speech signal (def. on p. 42)
STI_{sr}	speech-based revised Speech Transmission Index (def. on p. 32)
$THD_{\mathfrak{P}_x}$	total harmonic distortion (def. on p. 98)
$TID_{\mathcal{H}_{f_1}^2\mathfrak{P}_x+\mathcal{H}_{f_2}^2\mathfrak{P}_x}$	total intermodulation distortion (def. on p. 99)
TI_i	transmission index for STI$_{\mathrm{sr}}$ calculation (def. on p. 32)
$TND_{\mathcal{H}^2\mathfrak{P}_x}$	total non-linear distortion (def. on p. 101)
U_i	standard speech spectrum level at normal voice effort (def. on p. 28)
$V_i(E_i)$	self-speech masking spectrum level for SII calculation (def. on p. 28)
$W_i(\kappa)$	subband weight (def. on p. 20)
$W_i'(\kappa)$	limited subband weight to prevent hearing damage (def. on p. 26)
$W_i''(\kappa)$	limited subband weight for LimOptSIIbound (A5) (def. on p. 63)
$W_i^{\mathrm{max}}(\kappa)$	maximum subband weight (def. on p. 26)
$x(k)$	loudspeaker signal (def. on p. 16)
$x^{\mathrm{lim}}(k)$	limited loudspeaker signal (def. on p. 16)
$\mathcal{X}_\mu(k)$	DFT coefficients of loudspeaker signal (def. on p. 117)
$\tilde{\mathcal{X}}_\mu(k)$	DFT coefficients of windowed loudspeaker signal (def. on p. 120)
$\mathcal{X}_\mu'(k)$	modified DFT coefficients of loudspeaker signal (def. on p. 121)
$x_i(k)$	subband loudspeaker signal (def. on p. 116)
$\breve{x}_i(k)$	probe intensity envelope for STI$_{\mathrm{sr}}$ calculation (def. on p. 31)
$y(k)$	near-end microphone signal (def. on p. 14)
$\mathcal{Y}_\mu(\kappa)$	DFT coefficients of near-end microphone signal (def. on p. 20)
$\breve{y}_i(k)$	response intensity envelope for STI$_{\mathrm{sr}}$ calculation (def. on p. 31)

z	z-transform
$\breve{z}_i(k)$	noise intensity envelope for $\mathrm{STI}_{\mathrm{sr}}$ calculation (def. on p. 31)
$Z_i(N_i)$	masking spectrum level for SII calculation (def. on p. 28)
$\mathbb{Z}$	set of integers

Acronyms

AI	Articulation Index
ANC	active noise control
AS FB	analysis-synthesis filterbank
BASIE	Bayesian adaptive speech intelligibility estimation
BMLD	binaural masking level difference
DC	direct current
DFT	discrete Fourier transform
DOA	direction of arrival
DRC	dynamic range compression
EIC	equivalent intensity change
ERB	equivalent rectangular bandwidth
ERP	ear reference point
FBE	filterbank equalizer
FBSM	filterbank summation method
FFT	fast Fourier transform
FIR	finite impulse response
FP7	Seventh Framework Programme for Research of the European Union
GDFT	generalized DFT
GSDFT	generalized sliding DFT
LISTA	The Listening Talker
LOPRO	loudspeaker protection
MMSE	minimum mean-square error
NELE	near-end listening enhancement
PSD	power spectral density
RMS	root mean square
SDFT	sliding DFT
SDR	signal-to-disturbance ratio
SII	Speech Intelligibility Index
SNR	signal-to-noise ratio
SPL	sound pressure level
SRT	speech recognition threshold

STI	Speech Transmission Index
STI_r	revised Speech Transmission Index
STI_sr	speech-based revised Speech Transmission Index
THD	total harmonic distortion
TID	total intermodulation distortion
TND	total non-linear distortion
TTS	text-to-speech
VAD	voice activity detector

Presented NELE Algorithms

TheoPerfBound	theoretical performance bound (cf. Section 3.1.4)
OptSIIbound (A1)	bounded SII-based optimization (cf. Section 3.2.1)
SNRrecov (A2)	SNR recovery algorithm (cf. Section 3.2.3)
OptSIInum (A3)	numerical power-constrained SII-based optimization (cf. Section 4.1.1)
OptSIIrecur (A4)	recursive closed-form power-constrained SII-based optimization (cf. Section 4.1.2)
LimOptSIIbound (A5)	limited bounded SII-based optimization (cf. Section 4.1.5)
MaxTransfer (A6)	maximal power transfer (cf. Section 4.2.1)
OptSIIrecurDist (A7)	recursive closed-form power-constrained SII-based optimization with a priori limitation of disturbance spectrum level (cf. Section 4.2.4)
OptSIIone (A8)	one-step closed-form power-constrained SII-based optimization (cf. Section 4.2.5)

Bibliography

Publications by the author are marked with an asterisk (∗).

3GPP TS 26.090 (Dec. **2009**). *Adaptive Multi-Rate (AMR) speech codec. Transcoding functions.* Technical Specification. Version 9.0.0. 3rd Generation Partnership Project (3GPP) (cit. on p. 23).

3GPP TS 26.131 (Apr. **2011**). *Terminal acoustic characteristics for telephony. Requirements.* Technical Specification. Version 9.4.0. 3rd Generation Partnership Project (3GPP) (cit. on pp. 8, 17, 103, 110).

3GPP TS 26.132 (Apr. **2011**). *Speech and video telephony terminal acoustic test specification.* Technical Specification. Version 10.0.0. 3rd Generation Partnership Project (3GPP) (cit. on p. 17).

3GPP TS 26.190 (Dec. **2009**). *Adaptive Multi-Rate - Wideband (AMR-WB) speech codec. Transcoding functions.* Technical Specification. Version 9.0.0. 3rd Generation Partnership Project (3GPP) (cit. on p. 23).

ANSI S3.5-1969 (**1969**). *Methods for the Calculation of the Articulation Index.* American National Standards Institute (cit. on p. 37).

ANSI S3.5-1997 (**1997**). *Methods for the Calculation of the Speech Intelligibility Index.* American National Standards Institute (cit. on pp. 26–30, 33, 34, 43, 56, 61, 90).

Behler, Gottfried K. (Sept. 6, **2010**). Personal communication. RWTH Aachen University, Institute of Technical Acoustics (cit. on p. 98).

Blauert, Jens (**1997**). *Spatial Hearing. The Psychophysics of Human Sound Localization.* Trans. from the German by John S. Allen. Revised edition. Cambridge, Massachusetts; London, England: The MIT Press. ISBN: 978-0-262-02413-6 (cit. on p. 139).

Brookes, Mike (Nov. 2, **2012**). *VOICEBOX. Speech Processing Toolbox for MATLAB.* Department of Electrical & Electronic Engineering, Imperial College. URL: `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html` (visited on 11/15/2012) (cit. on pp. 25, 90).

Brouckxon, Henk; **Verhelst**, Werner; **Schuymer**, Bart De (Sept. **2008**). “Time and Frequency Dependent Amplification for Speech Intelligibility Enhancement in Noisy Environments”. In: *Proc. of Conf. of International Speech Communication Association (INTERSPEECH).* (Brisbane, Australia, Sept. 22–26, 2008). Vol. 9, pp. 557–560 (cit. on pp. 2, 38).

Chanda, Pinaki Shankar; **Park**, Sungjin (Apr. **2007**). “Speech Intelligibility Enhancement using Tunable Equalization Filter”. In: *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).* (Honolulu, Hawaii, USA, Apr. 15–20, 2007). Vol. 4, pp. 613–616. ISBN: 978-1-4244-0727-9. DOI: 10.1109/ICASSP.2007.366987 (cit. on pp. 2, 36, 79, 81, 89).

Choi, Jae-Hun; **Park**, Woo-Sang; **Chang**, Joon-Hyuk (Aug. **2009**). “Speech Reinforcement Based on Soft Decision Under Far-End Noise Environments”. In: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E92-A.8, pp. 2116–2119. ISSN: 0916-8508. DOI: 10.1587/transfun.E92.A.2116 (cit. on pp. 2, 38).

Cooke, Martin (Mar. **2006**). “A glimpsing model of speech perception in noise”. In: *Journal of the Acoustical Society of America* 119.3, pp. 1562–1573. ISSN: 0001-4966. DOI: 10.1121/1.2166600 (cit. on p. 39).

Cooke, Martin; **Mayo**, Catherine; **Valentini-Botinhao**, Cassia; **Kandia**, Varvara; **Petkov**, Petko; **Stylianou**, Yannis; **Tang**, Yan; **Villegas**, Julián; **Karasaikos**, Vasilis (Apr. 23, **2012**). *The Hurricane Challenge: An evaluation framework for the assessment of modified speech.* Tech. rep. LISTA D5.1. The Listening Talker Project (cit. on p. 83).

∗ **Cooke**, Martin; **Mayo**, Catherine; **Valentini-Botinhao**, Cassia; **Stylianou**, Yannis; **Sauert**, Bastian; **Tang**, Yan (May **2013**). “Evaluating the intelligibility benefit of speech modifications in known noise conditions”. In: *Speech Communication* 55.4, pp. 572–585. ISSN: 0167-6393. DOI: 10.1016/j.specom.2013.01.001 (cit. on pp. 3, 83–87, 89, 90).

Crochiere, Ronald E.; **Rabiner**, Lawrence R. (**1983**). *Multirate Digital Signal Processing.* Prentice-Hall Signal Processing Series. Englewood Cliffs, New Jersey: Prentice-Hall, Inc. ISBN: 978-0-13-605162-6 (cit. on p. 117).

Dolan, Terrence R.; **Robinson**, Donald E. (May **1967**). “Explanation of Masking-Level Differences That Result from Interaural Intensive Disparities of Noise”. In: *Journal of the Acoustical Society of America* 42.5, pp. 977–981. ISSN: 0001-4966. DOI: 10.1121/1.1910706 (cit. on p. 6).

EBU-SQAM-CD (Oct. 8, **2008**). *Sound Quality Assessment Material. Recordings for subjective tests.* European Broadcasting Union. URL: http://tech.ebu.ch/publications/sqamcd (visited on 04/08/2013) (cit. on p. 131).

EBU-Tech 3253 (Sept. 19, **2008**). *Sound Quality Assessment Material. Recordings for subjective tests. Users’ handbook for the EBU SQAM CD.* Tech. rep. Geneva: European Broadcasting Union (cit. on p. 131).

Erro, Daniel; **Stylianou**, Yannis; **Navas**, Eva; **Hernaez**, Inma (Sept. **2012**). “Implementation of Simple Spectral Techniques to Enhance the Intelligibility of Speech using a Harmonic Model”. In: *Proc. of Conf. of International Speech Communication Association (INTERSPEECH).* (Portland, OR, USA, Sept. 9–13, 2012). Vol. 13 (cit. on pp. 2, 38, 84).

Esch, Thomas; **Rüngeler**, Matthias; **Heese**, Florian; **Vary**, Peter (Oct. **2012**). “Estimation of Rapidly Time-Varying Harmonic Noise for Speech Enhancement”. In: *IEEE Signal Processing Letters* 19.10, pp. 659–662. ISSN: 1070-9908. DOI: 10.1109/LSP.2012.2211011 (cit. on p. 140).

Faraji, Neda; **Hendriks**, Richard C. (Sept. **2012**). “Noise Power Spectral Density Estimation for Public Address Systems in Noisy Reverberant Environments”. In: *Proc. of Intl. Workshop on Acoustic Echo and Noise Control (IWAENC).* (Aachen, Germany, Sept. 4–6, 2012). Vol. 13. ISBN: 978-3-8007-3451-1 (cit. on p. 141).

Gardner, William A. (Nov. **1992**). “A unifying view of coherence in signal processing”. In: *Signal Processing* 29.2, pp. 113–140. ISSN: 0165-1684. DOI: 10.1016/0165-1684(92)90015-O (cit. on p. 6).

Garofolo, John S.; **Lamel**, Lori F.; **Fisher**, William M.; **Fiscus**, Jonathan G.; **Pallett**, David S.; **Dahlgren**, Nancy L.; **Zue**, Victor (Oct. **1990**). *TIMIT – Acoustic-Phonetic Continuous Speech Corpus.* LDC93S1. Philadelphia: Linguistic Data Consortium. ISBN: 978-1-58563-019-6 (cit. on p. 33).

Gaubitch, Nikolay D.; **Brookes**, Mike; **Naylor**, Patrick A.; **Sharma**, Dushyant (June **2010**). “Bayesian Adaptive Method for Estimating Speech Intelligibility in Noise”. In: *Proc. of Intl. AES Conference. Audio Forensics: Practices and Challenges.* (Hillerød, Denmark, June 17–19, 2010). Vol. 39 (cit. on pp. 89, 90).

Goldsworthy, Ray L.; **Greenberg**, Julie E. (Dec. **2004**). “Analysis of speech-based speech transmission index methods with implications for nonlinear operations”. In: *Journal of the Acoustical Society of America* 116.6, pp. 3679–3689. ISSN: 0001-4966. DOI: 10.1121/1.1804628 (cit. on pp. 30, 31).

Hall, Joseph L.; **Flanagan**, James L. (**2010**). “Intelligibility and listener preference of telephone speech in the presence of babble noise”. In: *Journal of the Acoustical Society of America* 127.1, pp. 280–285. ISSN: 0001-4966. DOI: 10.1121/1.3263603 (cit. on pp. 2, 37).

Hamacher, Volkmar; **Chalupper**, Josef; **Eggers**, Joachim J.; **Fischer**, Eghart; **Kornagel**, Ulrich; **Puder**, Henning; **Rass**, Uwe (**2005**). “Signal Processing in High-End Hearing Aids: State of the Art, Challenges, and Future Trends”. In: *EURASIP Journal on Applied Signal Processing* 2005.18, pp. 2915–2929. ISSN: 1110-8657. DOI: 10.1155/ASP.2005.2915 (cit. on p. 141).

Harris, John G.; **Skowronski**, Mark D. (**2002**). “Energy redistribution speech intelligibility enhancement, vocalic and transitional cues”. In: *Journal of the Acoustical Society of America* 112.5, pp. 2305–2305. ISSN: 0001-4966. DOI: 10.1121/1.1506689 (cit. on pp. 1, 36).

Hawksford, Malcolm O. J. (Sept. **1999**). "MATLAB Program for Loudspeaker Equalization and Crossover Design". In: *Journal of the Audio Engineering Society* 47.9, pp. 706–719. ISSN: 0004-7554 (cit. on p. 133).

Hendriks, Richard C.; **Heusdens**, Richard; **Jensen**, Jesper (Mar. **2010a**). "MMSE based noise PSD tracking with low complexity". In: *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).* (Dallas, Texas, USA, Mar. 14–19, 2010), pp. 4266–4269. ISBN: 978-1-4244-4295-9. DOI: `10.1109/ICASSP.2010.5495680` (cit. on pp. 25, 141, 147, 149).

— (Apr. 15, **2010b**). *MMSE based noise PSD tracking algorithm.* Version V1. Signal & Information Processing Lab, Delft University of Technology. URL: `http://siplab.tudelft.nl/content/mmse-based-noise-psd-tracking-algorithm` (visited on 02/03/2011) (cit. on p. 25).

Houtgast, Tammo; **Steeneken**, Herman J. M. (**1971**). "Evaluation of Speech Transmission Channels by Using Artificial Signals". In: *Acustica* 25, pp. 355–367 (cit. on p. 30).

Houtgast, Tammo; **Steeneken**, Herman J. M.; **Ahnert**, Wolfgang; **Braida**, Louis; **Drullman**, Rob; **Festen**, Joost; **Jacob**, Kenneth; **Mapp**, Peter; **McManus**, Steve; **Payton**, Karen; **Plomp**, Reinier; **Verhave**, Jan; **Wijngaarden**, Sander van (**2002**). *Past, present and future of the Speech Transmission Index.* Ed. by Sander J. van Wijngaarden. PO Box 23, 3769 ZG Soesterberg, The Netherlands: TNO Human Factors. ISBN: 978-90-76702-02-5 (cit. on pp. 30, 32).

Hsu, T. S.; **Poornima**, K. A. (**2000**). "Temperature prediction of the voice coil of a moving coil loudspeaker by computer simulation". In: *Journal of the Acoustical Society of Japan* 21.2, pp. 57–62. ISSN: 0388-2861 (cit. on p. 93).

— (June **2001**). "Loudspeaker failure modes and error correction techniques". In: *Applied Acoustics* 62.6, pp. 717–734. ISSN: 0003-682X. DOI: `10.1016/S0003-682X(00)00064-5` (cit. on p. 93).

Huang, Dong-Yan; **Rahardja**, Susanto; **Ong**, Ee Ping (**2010**). "Lombard Effect Mimicking". In: *Proc. of ISCA Workshop on Speech Synthesis.* (Kyoto, Japan). Vol. 7, pp. 258–263 (cit. on pp. 2, 37).

IEC 60268-1:1985 (Jan. **1985**). *Sound system equipment – Part 1: General.* International Electrotechnical Commission (cit. on p. 127).

IEC 60268-16:2003 (May **2003**). *Sound system equipment – Part 16: Objective rating of speech intelligibility by speech transmission index.* International Electrotechnical Commission (cit. on pp. 30, 32).

IEEE (Sept. **1969**). "IEEE Recommended Practice for Speech Quality Measurements". In: *IEEE Transactions on Audio and Electroacoustics* 17.3, pp. 225–246. ISSN: 0018-9278. DOI: `10.1109/TAU.1969.1162058` (cit. on p. 85).

ITU-T Recommendation G.712 (Nov. **2001**). *Transmission performance characteristics of pulse code modulation channels.* Version 11/2001. International Telecommunication Union (cit. on p. 103).

ITU-T Recommendation G.729 (Jan. **2007**). *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP).* Version 01/2007. International Telecommunication Union (cit. on pp. 24, 27).

ITU-T Recommendation P.310 (June **2009**). *Transmission characteristics for narrow-band digital handset and headset telephones.* Version 06/2009. International Telecommunication Union (cit. on pp. 23, 24).

ITU-T Recommendation P.311 (June **2005**). *Transmission characteristics for wideband (150-7000 Hz) digital handset telephones.* Version 06/2005. International Telecommunication Union (cit. on pp. 23, 24).

ITU-T Recommendation P.56 (Mar. **1993**). *Objective Measurement of Active Speech Level.* Version 03/93. International Telecommunication Union (cit. on pp. 33, 90).

ITU-T Recommendation P.57 (Apr. **2009**). *Artificial ears.* Version 04/2009. International Telecommunication Union (cit. on pp. 9, 97).

Jacobsen, Eric; **Lyons**, Richard (Mar. **2003**). “The sliding DFT”. In: *IEEE Signal Processing Magazine* 20.2, pp. 74–80. ISSN: 1053-5888. DOI: 10.1109/MSP.2003.1184347 (cit. on pp. 119, 120).

— (Jan. **2004**). “An update to the sliding DFT”. In: *IEEE Signal Processing Magazine* 21.1, pp. 110–111. ISSN: 1053-5888. DOI: 10.1109/MSP.2004.1516381 (cit. on p. 119).

Jeub, Marco; **Dörbecker**, Matthias; **Vary**, Peter (Mar. **2011**). “A Semi-Analytical Model for the Binaural Coherence of Noise Fields”. In: *IEEE Signal Processing Letters* 18.3, pp. 197–200. ISSN: 1070-9908. DOI: 10.1109/LSP.2011.2108284 (cit. on p. 6).

Jokinen, Emma; **Alku**, Paavo; **Vainio**, Martti (Aug. **2012**). “Comparison of Post-Filtering Methods for Intelligibility Enhancement of Telephone Speech”. In: *Proc. of European Signal Processing Conf. (EUSIPCO).* (Bucharest, Romania, Aug. 27–31, 2012). European Association for Signal Processing (EURASIP), pp. 2333–2337. ISBN: 978-1-4673-1068-0 (cit. on pp. 2, 37).

Kammeyer, Karl-Dirk (**2004**). *Nachrichtenübertragung.* German. Ed. by Norbert Fliege and Martin Bossert. 3rd ed. Wiesbaden, Germany: B. G. Teubner Verlag. ISBN: 978-3-519-26142-1. DOI: 10.1007/978-3-322-94062-9 (cit. on p. 100).

Knowles (Sept. 12, **2011**). *13.6 X 9.6 X 2.9 MM Speaker.* Specification. Version E. Product No. 2403 260 00044 (cit. on pp. 23, 93, 102, 126).

Krebber, Winfried (**1995**). “Sprachübertragungsqualität von Fernsprech-Handapparaten”. German. PhD thesis. RWTH Aachen University. ISBN: 978-3-18-335710-9 (cit. on pp. 7–9).

Kretsinger, Elwood A.; **Young**, Norton B. (**1960**). “The Use of Fast Limiting to Improve the Intelligibility of Speech in Noise”. In: *Speech Monographs* 27, pp. 63–69 (cit. on pp. 1, 35).

Langner, Brian; **Black**, Allen W. (Mar. **2005**). “Improving the Understandability of Speech Synthesis by Modeling Speech in Noise”. In: *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).* (Philadelphia, PA, USA, Mar. 18–23, 2005). Vol. 1, pp. 265–268. ISBN: 978-0-7803-8874-1. DOI: 10.1109/ICASSP.2005.1415101 (cit. on p. 35).

Leonard, R. Gary (Mar. **1984**). “A database for speaker-independent digit recognition”. In: *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).* (San Diego, CA, USA, Mar. 19–21, 1984). Vol. 9, pp. 328–331. DOI: 10.1109/ICASSP.1984.1172716 (cit. on p. 90).

Leonard, R. Gary; **Doddington**, George (**1993**). *TIDIGITS*. LDC93S10. Philadelphia: Linguistic Data Consortium. ISBN: 978-1-58563-018-9 (cit. on p. 90).

Licklider, J. C. R. (Mar. **1948**). “The Influence of Interaural Phase Relations upon the Masking of Speech by White Noise”. In: *Journal of the Acoustical Society of America* 20.2, pp. 150–159. ISSN: 0001-4966. DOI: 10.1121/1.1906358 (cit. on pp. 139, 140).

Löllmann, Heinrich W. (Sept. **2011**). “Allpass-Based Analysis-Synthesis Filter-Banks: Design and Application”. PhD thesis. RWTH Aachen University. ISBN: 978-3-86130-308-4 (cit. on pp. 22, 23).

Löllmann, Heinrich W.; **Vary**, Peter (July **2007**). “Uniform and Warped Low Delay Filter-Banks for Speech Enhancement”. In: *Speech Communication* 49 (7-8): *Special Issue on Speech Enhancement*, pp. 574–587. ISSN: 0167-6393. DOI: 10.1016/j.specom.2007.04.009 (cit. on pp. 21–23, 35).

Lombard, Étienne (**1911**). “Le signe d’élévation de la voix [The sign of the elevation of the voice]”. French. In: *Annales des maladies de l’oreille et du larynx* 37, pp. 101–119 (cit. on pp. 1, 37).

Martin, Rainer (July **2001**). “Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics”. In: *IEEE Transactions on Speech and Audio Processing* 9.5, pp. 504–512. ISSN: 1063-6676. DOI: 10.1109/89.928915 (cit. on pp. 25, 147, 149).

— (June **2006**). “Bias compensation methods for minimum statistics noise power spectral density estimation”. In: *Signal Processing* 86.6: *Special Issue on Applied Speech and Audio Processing (dedicated to Prof. Hänsler).* Ed. by Henning Puder and Gerhard Schmidt, pp. 1215–1229. ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2005.07.037 (cit. on pp. 25, 147, 149).

McLoughlin, Ian Vince; **Chance**, R. J. (July **1997**). “LSP-based speech modification for intelligibility enhancement”. In: *Proc. of Intl. Conf. on Digital Signal Processing.* (Santorini, Greece, July 2–4, 1997). Vol. 2. IEEE, pp. 591–594. ISBN: 978-0-7803-4137-1. DOI: 10.1109/ICDSP.1997.628419 (cit. on pp. 2, 37).

Moore, Brian C. J.; **Glasberg**, Brian R. (Mar. **2007**). “Modeling binaural loudness”. In: *Journal of the Acoustical Society of America* 121.3, pp. 1604–1612. ISSN: 0001-4966. DOI: 10.1121/1.2431331 (cit. on p. 39).

Moore, Brian C. J.; **Glasberg**, Brian R.; **Baer**, Thomas (Apr. **1997**). “A Model for the Prediction of Thresholds, Loudness, and Partial Loudness”. In: *Journal of the Audio Engineering Society* 45.4, pp. 224–240. ISSN: 0004-7554 (cit. on p. 39).

Müller, Swen; **Massarani**, Paulo (June **2001**). “Transfer-function measurement with sweeps”. In: *Journal of the Audio Engineering Society* 49.6, pp. 443–471. ISSN: 0004-7554 (cit. on p. 125).

Niederjohn, Russell J.; **Grotelueschen**, James H. (Aug. **1976**). “The Enhancement of Speech Intelligibility in High Noise Levels by High-Pass Filtering Followed by Rapid Amplitude Compression”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.4, pp. 277–282. ISSN: 0096-3518. DOI: `10.1109/TASSP.1976.1162824` (cit. on pp. 1, 36, 79, 80).

— (Aug. **1978**). “Speech Intelligibility Enhancement in a Power Generating Noise Environment”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.4, pp. 378–380. ISSN: 0096-3518. DOI: `10.1109/TASSP.1978.1163100` (cit. on p. 36).

NXP Semiconductors (June 8, **2010a**). *15x11x3.5 MM RA Speaker*. Specification. Version J. Order No. 2403 260 00001 (cit. on pp. 23, 54, 93).

— (Mar. 30, **2010b**). *RA 8x12x2 Receiver*. Specification. Version G. Order No. 2403 260 00031 (cit. on pp. 23, 93, 110, 133).

Park, Hochong; **Yoon**, Jae-Yul; **Kim**, Jung-Hoe; **Oh**, Eunmi (May **2010**). “Improving Perceptual Quality of Speech in a Noisy Environment by Enhancing Temporal Envelope and Pitch”. In: *IEEE Signal Processing Letters* 17.5, pp. 489–492. ISSN: 1070-9908. DOI: `10.1109/LSP.2010.2044937` (cit. on pp. 2, 37, 79, 82).

Plomp, Reinier; **Mimpen**, A. M. (**1979**). “Improving the Reliability of Testing the Speech Reception Threshold for Sentences”. In: *International Journal of Audiology* 18.1, pp. 43–52. ISSN: 1499-2027. DOI: `10.3109/00206097909072618` (cit. on p. 89).

Raitio, Tuomo; **Suni**, Antti; **Vainio**, Martti; **Alku**, Paavo (**2011**). “Analysis of HMM-Based Lombard Speech Synthesis”. In: *Proc. of Conf. of International Speech Communication Association (INTERSPEECH)*. (Florence, Italy, Aug. 28–31, 2011), pp. 2781–2784 (cit. on p. 35).

Rankovic, Christine M. (Apr. **1991**). “An application of the articulation index to hearing aid fitting.” In: *Journal of Speech and Hearing Research* 34.2, pp. 391–402. ISSN: 1092-4388 (cit. on pp. 2, 37).

Rasetshwane, Daniel Motlotle; **Boston**, J. Robert; **Li**, Ching-Chung; **Durrant**, John D.; **Genna**, Greg (**2009**). “Enhancement of speech intelligibility using transients extracted by wavelet packets”. In: *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. (New Paltz, NY, Oct. 18–21, 2009), pp. 173–176. ISBN: 978-1-4244-3678-1. DOI: `10.1109/ASPAA.2009.5346465` (cit. on pp. 2, 36).

Rhebergen, Koenraad S.; **Versfeld**, Niek J. (Apr. **2005**). "A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners". In: *Journal of the Acoustical Society of America* 117.4, pp. 2181–2192. ISSN: 0001-4966. DOI: `10.1121/1.1861713` (cit. on p. 41).

∗ **Sauert**, Bastian; **Enzner**, Gerald; **Vary**, Peter (Sept. **2006**). "Near End Listening Enhancement with Strict Loudspeaker Output Power Constraining". In: *Proc. of Intl. Workshop on Acoustic Echo and Noise Control (IWAENC).* (Paris, France, Sept. 12–14, 2006) (cit. on pp. 3, 38, 64).

∗ **Sauert**, Bastian; **Löllmann**, Heinrich W.; **Vary**, Peter (Oct. **2008**). "Near End Listening Enhancement by Means of Warped Low Delay Filter-Banks". In: *Proc. of ITG-Fachtagung Sprachkommunikation.* (Aachen, Germany, Oct. 8–10, 2008). Vol. 8. Berlin [u.a.]: VDE-Verlag. ISBN: 978-3-8007-3120-6 (cit. on pp. 3, 21, 38, 48, 154).

∗ **Sauert**, Bastian; **Vary**, Peter (Apr. **2006a**). "Improving Speech Intelligibility in Noisy Environments by Near End Listening Enhancement". In: *Proc. of ITG-Fachtagung Sprachkommunikation.* (Kiel, Germany, Apr. 26–28, 2006). Vol. 7. Berlin [u.a.]: VDE-Verlag (cit. on pp. 3, 38, 48).

∗ — (May **2006b**). "Near End Listening Enhancement: Speech Intelligibility Improvement in Noisy Environments". In: *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).* (Toulouse, France, May 14–19, 2006). Vol. 1, pp. 493–496. ISBN: 978-1-4244-0469-8. DOI: `10.1109/ICASSP.2006.1660065` (cit. on pp. 2, 3, 38, 48).

∗ — (Aug. **2009**). "Near End Listening Enhancement Optimized with Respect to Speech Intelligibility Index". In: *Proc. of European Signal Processing Conf. (EUSIPCO).* (Glasgow, Scotland, Aug. 24–28, 2009). Vol. 17. European Association for Signal Processing (EURASIP). New York, NY: Hindawi Publ., pp. 1844–1848 (cit. on pp. 3, 39, 47).

∗ — (Aug. **2010a**). "Near End Listening Enhancement Optimized with Respect to Speech Intelligibility Index and Audio Power Limitations". In: *Proc. of European Signal Processing Conf. (EUSIPCO).* (Aalborg, Denmark, Aug. 23–27, 2010). Vol. 18. European Association for Signal Processing (EURASIP), pp. 1919–1923 (cit. on pp. 3, 39, 56).

∗ — (Oct. **2010b**). "Recursive Closed-Form Optimization of Spectral Audio Power Allocation for Near End Listening Enhancement". In: *Proc. of ITG-Fachtagung Sprachkommunikation.* (Bochum, Germany, Oct. 6–8, 2010). Vol. 9. Berlin [u.a.]: VDE-Verlag. ISBN: 978-3-8007-3300-2 (cit. on pp. 3, 39, 59).

∗ — (Sept. **2011**). "Near End Listening Enhancement Considering Thermal Limit of Mobile Phone Loudspeakers". In: *Proc. of Conf. on Elektronische Sprachsignalverarbeitung (ESSV).* (Aachen, Germany, Sept. 28–30, 2011). Vol. 61. ITG, DEGA. Dresden, Germany: TuDPress Verlag der Wissenschaften GmbH, pp. 333–340. ISBN: 978-3-94271-037-4 (cit. on pp. 3, 24, 25, 39, 63).

∗ — (May **2012a**). “Listening Enhancement for Mobile Phones – How to Improve the Intelligibility in a Noisy Environment”. In: *The Listening Talker Workshop (LISTA).* (Edinburgh, Scotland, May 2–3, 2012). Invited Talk. International Speech Communication Association (ISCA) (cit. on p. 3).

∗ — (Sept. **2012b**). “Near-End Listening Enhancement in the Presence of Bandpass Noises”. In: *Proc. of ITG-Fachtagung Sprachkommunikation.* (Braunschweig, Germany, Sept. 26–28, 2012). Vol. 10. Berlin [u.a.]: VDE-Verlag. ISBN: 978-3-8007-3455-9 (cit. on pp. 3, 39, 74).

Schäfer, Magnus (Aug. **2005**). “System Identification of an Artificial Head Measurement System with Telephone Handset”. German. Studienarbeit. Muffeter Weg 3a, 52074 Aachen, Germany: RWTH Aachen University, Institute of Communication Systems and Data Processing (cit. on pp. 8, 9).

∗ **Schäfer**, Magnus; **Jeub**, Marco; **Sauert**, Bastian; **Vary**, Peter (Aug. **2010**). “Reverberation-Based Post-Processing for Improving Speech Intelligibility”. In: *Intl. Congress on Acoustics (ICA).* (Sydney, Australia, Aug. 23–27, 2010). Australian Acoustical Society. ISBN: 978-0-64654-052-8 (cit. on p. 3).

∗ **Schönle**, Martin; **Beaugeant**, Christophe; **Steinert**, Kai; **Löllmann**, Heinrich W.; **Sauert**, Bastian; **Vary**, Peter (Sept. **2006**). “Hands-Free Audio and its Application to Telecommunication Terminals”. In: *Proc. of Intl. AES Conference. Audio for Mobile and Handheld Devices.* (Seoul, Korea, Sept. 2–4, 2006). Vol. 29 (cit. on p. 3).

Schumacher, Thomas; **Krüger**, Hauke; **Jeub**, Marco; **Vary**, Peter; **Beaugeant**, Christophe (May **2011**). “Active Noise Control in Headsets: A New Approach for Broadband Feedback ANC”. In: *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).* (Prague, Czech Republic, May 22–27, 2011), pp. 417–420. ISBN: 978-1-4577-0538-0. DOI: 10.1109/ICASSP.2011.5946429 (cit. on p. 140).

Shin, Ho Seon; **Choi**, Min-Seok; **Kim**, Taesu; **Kang**, Hong-Goo (Mar. **2010**). “Binaural Loudness Based Speech Reinforcement with a Closed-Form Solution”. In: *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).* (Dallas, Texas, USA, Mar. 14–19, 2010), pp. 4274–4277. ISBN: 978-1-4244-4295-9. DOI: 10.1109/ICASSP.2010.5495682 (cit. on pp. 2, 39).

Shin, Jong Won; **Jin**, Yu Gwang; **Park**, Seung Seop; **Kim**, Nam Soo (Apr. **2009**). “Speech Reinforcement Based on Partial Masking Effect”. In: *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).* (Taipei, Taiwan, Apr. 19–24, 2009), pp. 4401–4404. ISBN: 978-1-4244-2353-8. DOI: 10.1109/ICASSP.2009.4960605 (cit. on pp. 2, 39).

Shin, Jong Won; **Lim**, Woohyung; **Sung**, Junesig; **Kim**, Nam Soo (Aug. **2007**). “Speech Reinforcement based on Partial Specific Loudness”. In: *Proc. of Conf. of International Speech Communication Association (INTERSPEECH).* (Antwerp, Belgium, Aug. 27–31, 2007). Vol. 8, pp. 978–981 (cit. on p. 39).

Skowronski, Mark D.; **Harris**, John G. (May **2006**). “Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments”. In: *Speech Communication* 48.5, pp. 549–558. ISSN: 0167-6393. DOI: 10.1016/j.specom.2005.09.003 (cit. on pp. 36, 79, 81).

Smith III, Julius O.; **Abel**, Jonathan S. (Nov. **1999**). “Bark and ERB bilinear transforms”. In: *IEEE Transactions on Speech and Audio Processing* 7.6, pp. 697–708. ISSN: 1063-6676. DOI: 10.1109/89.799695 (cit. on pp. 23, 154).

SPIB (Sept. 19, **1995**). *Examples of the NOISEX-92 database.* Rice University. URL: http://spib.rice.edu/spib/select_noise.html (visited on 02/01/2010) (cit. on p. 33).

Steeneken, Herman J. M.; **Houtgast**, Tammo (Jan. **1980**). “A physical method for measuring speech-transmission quality”. In: *Journal of the Acoustical Society of America* 67.1, pp. 318–326. ISSN: 0001-4966. DOI: 10.1121/1.384464 (cit. on p. 30).

— (Sept. **1991**). “On the Mutual Dependency of Octave-Band-Specific Contributions to Speech Intelligibility”. In: *Proc. of European Conf. on Speech Communication and Technology (EUROSPEECH).* (Genova, Italy, Sept. 24–26, 1991). Vol. 2. International Speech Communication Association (ISCA), pp. 1133–1136 (cit. on p. 30).

— (June **1999**). “Mutual dependence of the octave-band weights in predicting speech intelligibility”. In: *Speech Communication* 28.2, pp. 109–123. ISSN: 0167-6393. DOI: 10.1016/S0167-6393(99)00007-2 (cit. on p. 30).

Summers, W. Van; **Pisoni**, David B.; **Bernacki**, Robert H.; **Pedlow**, Robert I.; **Stokes**, Michael A. (**1988**). “Effects of noise on speech production: Acoustic and perceptual analysis”. In: *Journal of the Acoustical Society of America* 84, pp. 917–928. ISSN: 0001-4966. DOI: 10.1121/1.396660 (cit. on pp. 1, 37).

Taal, Cees H.; **Hendriks**, Richard C.; **Heusdens**, Richard (Mar. **2012**). “A Speech Preprocessing Strategy for Intelligibility Improvement in Noise Based on a Perceptual Distortion Measure”. In: *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).* (Kyoto, Japan, Mar. 25–30, 2012), pp. 4061–4064. ISBN: 978-1-4673-0045-2. DOI: 10.1109/ICASSP.2012.6288810 (cit. on pp. 2, 39).

Tang, Yan; **Cooke**, Martin (Sept. **2010**). “Energy Reallocation Strategies for Speech Enhancement in Known Noise Conditions”. In: *Proc. of Conf. of International Speech Communication Association (INTERSPEECH).* (Makuhari, Chiba, Japan, Sept. 26–30, 2010). Vol. 11, pp. 1636–1639 (cit. on pp. 38, 84).

— (Aug. **2011**). “Subjective and Objective Evaluation of Speech Intelligibility Enhancement Under Constant Energy and Duration Constraints”. In: *Proc. of Conf. of International Speech Communication Association (INTERSPEECH).* (Florence, Italy, Aug. 27–31, 2011). Vol. 12, pp. 345–348 (cit. on pp. 2, 38).

— (Sept. **2012**). “Optimised spectral weightings for noise-dependent speech intelligibility enhancement”. In: *Proc. of Conf. of International Speech Communication Association (INTERSPEECH).* (Portland, OR, USA, Sept. 9–13, 2012). Vol. 13 (cit. on pp. 2, 39, 84).

Tantibundhit, Charturong; **Boston**, J. Robert; **Li**, Ching-Chung; **Durrant**, John D.; **Shaiman**, Susan; **Kovacyk**, Kristie; **El-Jaroudi**, Amro (Nov. **2007**). “New signal decomposition method based speech enhancement”. In: *Signal Processing* 87.11, pp. 2607–2628. ISSN: 0165-1684. DOI: 10.1016/j.sigpro.2007.04.014 (cit. on pp. 2, 36).

Thomas, Ian B. (Apr. **1968**). “The Influence of First and Second Formants on the Intelligibility of Clipped Speech”. In: *Journal of the Audio Engineering Society* 16.2, pp. 182–185. ISSN: 0004-7554 (cit. on p. 37).

Thomas, Ian B.; **Niederjohn**, Russell J. (Oct. **1968**). “Enhancement of Speech Intelligibility at High Noise Levels by Filtering and Clipping”. In: *Journal of the Audio Engineering Society* 16.4, pp. 412–415. ISSN: 0004-7554 (cit. on pp. 36, 79, 80).

— (June **1970**). “The Intelligibility of Filtered-Clipped Speech in Noise”. In: *Journal of the Audio Engineering Society* 18.3, pp. 299–303. ISSN: 0004-7554 (cit. on pp. 1, 36, 79, 80).

Thomas, Ian B.; **Ohley**, William J. (**1972**). “Intelligibility Enhancement Through Spectral Weighting”. In: *Proc. of Conf. on Speech Communication and Processing.* (Newton, Mass, USA, Apr. 24–26, 1972), pp. 360–363 (cit. on pp. 2, 36, 37, 79, 80).

∗ **Valentini-Botinhao**, Cassia; **Godoy**, Elizabeth; **Stylianou**, Yannis; **Sauert**, Bastian; **King**, Simon; **Yamagishi**, Junichi (May **2013**). “Improving Intelligibility in Noise of HMM-Generated Speech via Noise-Dependent and -Independent Methods”. In: *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).* (Vancouver, Canada, May 26–31, 2013), pp. 7854–7858. DOI: 10.1109/ICASSP.2013.6639193 (cit. on pp. 3, 87, 88).

Valentini-Botinhao, Cassia; **Maia**, Ranniery; **Yamagishi**, Junichi; **King**, Simon; **Zen**, Heiga (Mar. **2012**). “Cepstral Analysis Based on the Glimpse Proportion Measure for Improving the Intelligibility of HMM-Based Synthetic Speech in Noise”. In: *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP).* (Kyoto, Japan, Mar. 25–30, 2012), pp. 3997–4000. ISBN: 978-1-4673-0045-2. DOI: 10.1109/ICASSP.2012.6288794 (cit. on p. 35).

Valentini-Botinhao, Cassia; **Yamagishi**, Junichi; **King**, Simon (Sept. **2012**). “Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise”. In: *Proc. of Conf. of International Speech Communication Association (INTERSPEECH).* (Portland, OR, USA, Sept. 9–13, 2012). Vol. 13 (cit. on p. 84).

Varga, Andrew; **Steeneken**, Herman J. M. (July **1993**). “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems”. In: *Speech Communication* 12.3, pp. 247–251. ISSN: 0167-6393. DOI: `10.1016/0167-6393(93)90095-3` (cit. on p. 33).

Vary, Peter (**1980**). “Fast Digital Frequency Response Measurement with Multifrequency Signals”. In: *AEÜ. Intl. Journal of Electronics and Communications* 34.5, pp. 190–195. ISSN: 1434-8411 (cit. on p. 10).

— (June **2006**). “An Adaptive Filterbank Equalizer for Speech Enhancement”. In: *Signal Processing* 86.6: *Special Issue on Applied Speech and Audio Processing (dedicated to Prof. Hänsler)*. Ed. by Henning Puder and Gerhard Schmidt, pp. 1206–1214. ISSN: 0165-1684. DOI: `10.1016/j.sigpro.2005.06.020` (cit. on pp. 21, 22).

Vorländer, Michael (Apr. **2008**). *Technische Akustik II.* German. Tech. rep. Neustraße 50, 52066 Aachen, Germany: RWTH Aachen University, Institute of Technical Acoustics (cit. on p. 18).

Wilbanks, W. A.; **Whitmore**, John K. (Apr. **1968**). “Detection of Monaural Signals as a Function of Interaural Noise Correlation and Signal Frequency”. In: *Journal of the Acoustical Society of America* 43.4, pp. 785–797. ISSN: 0001-4966. DOI: `10.1121/1.1910897` (cit. on p. 6).

Yoo, Sungyub D.; **Boston**, J. Robert; **Durrant**, John D.; **Kovacyk**, Kristie; **Karn**, Stacey; **Shaiman**, Susan; **El-Jaroudi**, Amro; **Li**, Ching-Chung (Sept. **2004**). “Relative Energy and Intelligibility of Transient Speech Components”. In: *Proc. of European Signal Processing Conf. (EUSIPCO).* (Vienna, Austria, Sept. 6–10, 2004). Vol. 12. European Association for Signal Processing (EURASIP), pp. 1031–1034 (cit. on p. 36).

Yoo, Sungyub D.; **Boston**, J. Robert; **El-Jaroudi**, Amro; **Li**, Ching-Chung; **Durrant**, John D.; **Kovacyk**, Kristie; **Shaiman**, Susan (Aug. **2007**). “Speech signal modification to increase intelligibility in noisy environments”. In: *Journal of the Acoustical Society of America* 122.2, pp. 1138–1149. ISSN: 0001-4966. DOI: `10.1121/1.2751257` (cit. on pp. 2, 36).

Zorilă, Tudor-Catalin; **Kandia**, Varvara; **Stylianou**, Yannis (Sept. **2012**). “Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression”. In: *Proc. of Conf. of International Speech Communication Association (INTERSPEECH).* (Portland, OR, USA, Sept. 9–13, 2012). Vol. 13 (cit. on pp. 2, 38, 84).

Zwicker, Eberhard (**1961**). “Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)”. In: *Journal of the Acoustical Society of America* 33.2, p. 248. ISSN: 0001-4966. DOI: `10.1121/1.1908630` (cit. on p. 25).

Zwicker, Eberhard; **Fastl**, Hugo (**1999**). *Psychoacoustics. Facts and Models.* 2nd ed. Berlin, Heidelberg, New York: Springer. ISBN: 978-3-540-65063-8 (cit. on pp. 21, 23, 25, 26, 64).