Technische Universität Berlin
Fakultät I

Fachgebiet Audiokommunikation und -technologie

Masterarbeit

Simulation and analysis of measurement techniques for the fast acquisition of individual head-related transfer functions

Vorgelegt von: Mina Fallahi
[redacted]

Abgabe: 17.06.2014

Erstgutachter:
Prof. Dr. Stefan Weinzierl

Zweitgutachter:
Fabian Brinkmann

Eidesstattlische Erklärung

Hiermit versichere ich gegenüber der Fakultät I der Technischen Universität Berlin durch meine Unterschrift, die vorliegende Arbeit selbstständig ohne fremde Hilfe und nur unter Verwendung der angegebenen Quellen und Hilfsmittel verfasst zu haben. Alle Ausführungen, die anderen veröffentlichten oder nicht veröffentlichten Schriften wörtlich oder sinngemäß entnommen wurden, habe ich kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Fassung noch keiner anderen Prüfungsbehörde vorgelegen.

Berlin, den 14. Juli 2014

....

Mina Fallahi

Acknowledgement

Foremost, I would like to express my gratitude to Prof. Dr. Stefan Weinzierl and Fabian Brinkmann, for their kind motivation, technical support, advice and immense knowledge.

I am also thankful to Prof. Dr. Ing. Thomas Jürgensohn (Human-Factor-Consult GmbH), Alexander Fuß, Dr. Alexander Lindau and Marc Voigt.

In particular, I would like to thank all my family members, especially my parents and my sister, Dr. Delaram Fallahi, for their kind support and encouragement.

In this thesis, the ITA-Toolbox for MATLAB developed at the Institute of Technical Acoustics at RWTH Aachen has been used for data processing.

Abstract

Due to the need for individually acquired head-related transfer functions (HRTFs) and the long and tedious measurement process for high spatial resolution HRTFs, in this thesis, a system for the fast measurement of individual HRTFs was modeled and simulated. Two system identification methods for rapid HRTF measurement were implemented: the optimized Multiple Exponential Sweep Method (optimized MESM) and the Normalized Least Mean Square (NLMS) adaptive filter algorithm. The HRTF measurement was assumed to be done with a horizontally continuous rotation of the subject with respect to the sound source during the measurement. The HRTFs were acquired for different conditions regarding environmental noise, number of loudspeaker channels and nonlinear distortion. Both implemented methods were able to offer satisfactory results within measurement durations considerably shorter than that of conventionally used methods. Despite the robustness and stability of optimized MESM with respect to noise disturbances, this method showed limitations caused by the continuous rotation. It was shown, that, as long as the signal to noise ratio is held sufficiently high, the NLMS algorithm represents the better choice.

Zusammenfassung

Aufgrund der Notwendigkeit, kopfbezogene Übertragungsfunktionen (HRTFs) individuell zu messen, und Bezug nehmend auf das lange und mühsame Verfahren zur Messung der HRTFs mit hoher räumlicher Auflösung, wurde in dieser Arbeit ein System zur schnellen Messung von HRTFs modelliert und simuliert. Es wurden zwei Systemidentifikationsverfahren für die schnelle Messung der HRTFs implementiert: das optimierte Multiple Exponential Sweep Method (optimized MESM) und der Normalized Least Mean Square (NLMS) adaptive-Filter Algorithmus. Es wurde angenommen, dass die Messung mit einer horizontal kontinuierlichen Drehung der Versuchsperson in Bezug auf die Schallquelle während der Messung durchgeführt wird. Die HRTFs wurden für unterschiedliche Bedingungen bezüglich der Umgebungsgeräusche, Anzahl der Lautsprecherkanäle und nicht lineare Verzerrungen erworben. Beide implementierten Methoden konnten zufriedenstellende Ergebnisse innerhalb einer Messdauer bieten, die erheblich kürzer war, als die Messdauer von herkömmlich verwendeten Verfahren. Trotz der Stabilität von optimized MESM in Bezug auf Lärm und Rauschen, zeigte diese Methode Einschränkungen, die durch die kontinuierliche Umdrehung verursacht wurden. Es wurde gezeigt, dass, solange das Signal-Rausch Verhältnis hoch gehalten wird, der NLMS Algorithmus die bessere Wahl darstellt.

Contents

Glossary

BEM	boundary element method
BRIR	binaural room impulse response
BRTF	binaural room transfer function
DUT	device under test
ERB	equivalent rectangular bandwidth
ES	exponential sweep
ESA	error signal attenuation
FABIAN	**F**ast and **A**utomatic **B**inaural **I**mpulse response **A**cquisitio**N**
FEC	free air equivalent coupling
FFT	fast fourier transform
HIR	harmonic impulse response
HRIR	head-related impulse response
HRTF	head-related transfer function
ILD	interaural level difference
IR	impulse response
ITD	interaural time difference
LMS	least mean square
MLS	maximum length sequence
MESM	multiple exponential sweep method
NLMS	normalized least mean square
PSEQ	perfect sequence
SNR	signal to noise ratio
THD	total harmonic distortion

List of symbols

f_s	sampling frequency
f	frequency
θ	azimuth
φ	elevation
r	distance
T_{360}	the duration of one complete rotation of 360°
a_k	energy decay of k^{th} harmonic response (optimized MESM)
α	the percentage of the length of the useful impulse response (optimized MESM)
η - η_{opt}	number of the interleaved channels - optimal number of interleaved channels
K_{max}	maximum number of harmonic responses
M	number of loudspeaker channels
r_s	sweep rate
t_i	excitation time for each i^{th} system using MESM
T	duration of the excitation signal
T_{ES}	measurement duration using exponential sweep method
T_{INT}	measurement duration using interleaving strategy
T_{OV}	measurement duration using overlapping strategy
T_{MESM}	measurement duration using MESM
T_{OPT}	measurement duration using optimized MESM
τ_{gd}	group delay
τ_{IR}	length of the room impulse response
$\tau_{IR,k}$	length of the k^{th} nonlinear impulse response
τ_{IL}	minimum delay between two interleaved sweeps
τ_{OV}	minimum delay between two consequent sweeps using overlapping
τ_w	minimum delay between consequent excitations (optimized MESM)
τ_{st}	stop margin (the time to allow the system to decay after the sweep stops - optimized MESM)
τ_{sp}	saftey time (optimized MESM)
τ_{DUT}	the length of the usefull part of the impulse response (optimized MESM)
Δt_k	beginning time of the k^{th} nonlinear impulse response
ν	efficiency of MESM relative to conventional measurements
$\mathbf{d}(k)$	distance vector (mismatch between estimated and real impulse response)
$e(k)$	NLMS error signal
$\varepsilon(k)$	NLMS estimation error
$\mathcal{D}(k)$	mean square deviation
$\text{Error}_{\text{noise}}$	effect of environmental noise on NLMS inaccuracy
$\text{Error}_{\text{dynamic}}$	effect of variability of dynamic HRIR acquisition system on NLMS inaccuracy
k	iteration
λ_{min}	smallest eigenvalue of the input correlation matrix
λ_{max}	largest eigenvalue of the input correlation matrix
N	length of the impulse response or NLMS filter

$\mathbf{R}$	imput correlation matrix
R_{pp}	periodic autocorrelation function
${\sigma^2}_e$	variance of the error signal
${\sigma^2}_y$	variance of the signal captured by microphones

Chapter 1

Introduction

Binaural technology has the aim of supplying the listener with a reliable representation of the recorded sound. Due to its capacity to be used to perceive and localize the sound source, binaural technology can be employed in virtual reality to create a virtual sound source anywhere around the listener, for example it can be used for psychoacoustic experimentations [Nic 10]. It can also be used in applications such as entertainment products (games), or development of guidance systems for visually impaired people [Par 12a]. The free field sound propagation between the sound source and the listener's ears is described by Head-Related Transfer Functions (HRTFs). HRTFs include all spatial information which the listener uses to localize the sound source, and build therefore the basis of the binaural technology. For a high spatial resolution HRTF data set, acquiring the transfer function of the source-ear path for all possible source positions around the listener poses a tedious and time consuming task. In most practical applications, the measurement is carried out once, using recordings of one listener or more often, of an artificial head and the measured HRTFs are used to reproduce the binaural signals for other listeners. However, there have been studies [Møl 96, Wen 93], showing that listening to binaural signals, which originate from other subjects or from an artificial head, leads to localization and coloration errors. This is due to the fact that the HRTFs include the individual filtering information of reflections and diffractions from the subject's head, torso and pinna. Therefore, for applications with high demand on fidelity, individual HRTFs gain more importance. In order to acquire customized HRTFs, besides carrying out the measurements directly on individuals, there are also other methods which are based on simulation or modeling, each with its advantages and disadvantages. However, as long as the HRTF acquisition is to be done by real measurements with individual subjects, the long duration of the measurement imposes a serious constraint. On the one hand, the subject must keep still during the measurement to avoid the artifacts caused by head movements and it can get unpleasant for him or her to hold on throughout the measurement process. On the other hand, long measurement durations give rise to the appearance of time varying elements such as temperature changes or the subject's unwanted head movements. Therefore, reducing the measurement duration, without loss of quality in the results, represents the main motivation. In this thesis, two of the proposed methods for speeding up the HRTF measurement were studied and used to simulate a system for the fast measurement of head-related transfer functions. These two methods were the optimized Multiple Exponential Sweep Method (optimized MESM), suggested by Dietrich et al. [Die 13a] which is based on MESM, proposed by Majdak et al. [Maj 07], and the continuous azimuth HRTF measurement using Normalized Least Mean Square (NLMS) adaptive filters, which was introduced by Enzner [Enz 08, Enz 09]. This simulated measurement system consisted of a vertical arc of up to 39 loudspeaker channels with the subject's head positioned in the center of the arc. The subject was assumed to be rotated permanently during the measurement in the horizontal direction to accomplish the measurement within one complete rotation of 360°. This measurement setup is in accordance with the setup described by Enzner for the continuous azimuth HRTF measurement with NLMS adaptive filtering [Enz 09]. Although the other algorithm, the optimized MESM [Die 13a], is originally considered for a discrete azimuth HRTF measurement, this algorithm was also applied to the continuous rotation measurement setup in the simulations. After introducing an overview of the background studies carried out concerning HRTFs and human's spatial hearing in chapter 2, chapter 3 introduces the MESM and optimized MESM algorithms and deals with the review and derivation of the equations which define the performance of the algorithms. The NLMS adaptive filter and its application to HRTF measurement are introduced in chapter 4, with a discussion on the parameters that impact the performance of the system identification with this algorithm. The sound propagation path between loudspeakers and microphones within a HRTF measurement as well as the continuous rotation of the subject during the measurement are modeled in chapter 5. Finally, the results of the simulations are presented and discussed in chapter 6. These simulations had two aims. One aim was to study the performance of each algorithm for different measurement situations concerning number of loudspeaker channels, environmental noise or nonlinear distortions, of course in respect of the measurement duration. The other aim was to explore which implemented algorithm and under which conditions provides the more suitable method for the modeled measurement setup.

This master thesis is part of a project to develop a system for the fast measurement of individual head-related transfer functions, in collaboration with Human Factors Consult [1]. The other part of the project, within another master thesis, concerns the construction of a real HRTF measurement system, for which the results of the present thesis could be used.

[1] www.human-factors-consult.de

Chapter 2

State of Research

This chapter introduces the main studies concerning Head-Related Transfer Functions (HRTFs) and their measurements, starting with the fundamentals of spatial hearing in section 2.1, followed by the role of HRTFs in binaural technology in section 2.2. The advantages of individually measured HRTFs are explained in section 2.3. The chapter concludes with introducing some main trends in HRTF individualization.

2.1 Spatial hearing and sound source localization

Human's ability of sound localization originates from binaural hearing. To define the position of a sound source with respect to the listener, the spherical coordinate system as shown in figure 2.1 can be considered. The origin of this coordinate system is located on the interaural axis at the point exactly between the two ear canal entrances [Bla 08]. The location of the source is defined with the azimuth (θ), elevation (φ) and distance (r) with respect to the origin. The spherical coordinate system of figure 2.1 includes the three orthogonal planes, namely the horizontal, the frontal and the median planes, which run through the origin.

For sound sources which are located away from the median plane, i.e. for lateral sound incidences, the emitted sound reaches the closer ear earlier and at a higher level. These two phenomena lead to Interaural Time Differences (ITD) and Interaural Level Differences (ILD) respectively. ILD and ITD are two important binaural features by which the auditory system localizes the sound source. For sound incidence from sources located on the median plane, the signals at both ears are very similar in level and arrival time. In the absence of binaural features the monoaural spectral cues play the dominant role. These spectral features are characterized by listener's individual morphology, especially by outer ear structure and the pinna. According to [Bla 97], the pinna acts like a filter which, depending on the sound source distance and direction, affects different parts of the spectrum of the sound signal. Studies have shown that

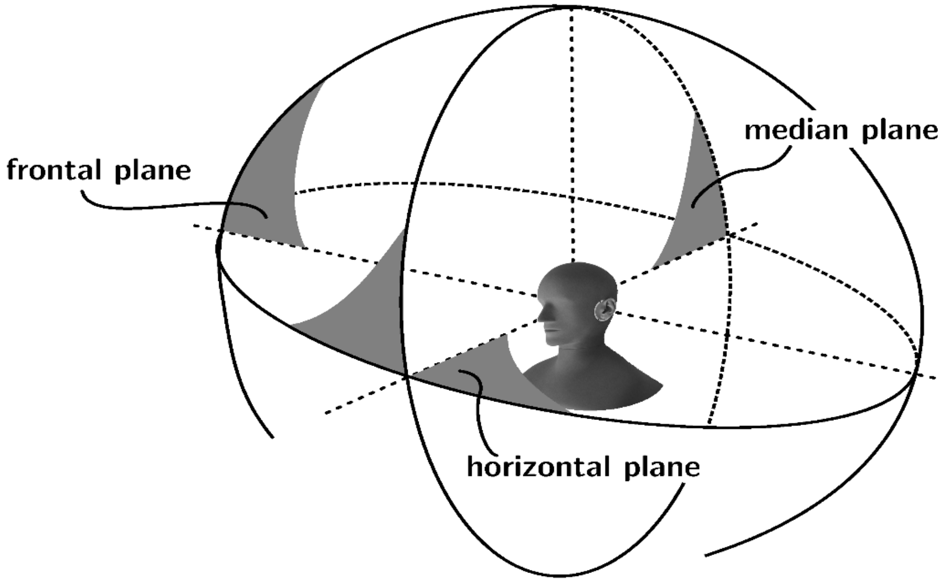


Figure 2.1: Head centered spherical coordinate system. [Vor 08]

familiarity with the signal improves the directional hearing in the median plane, especially for distinguishing between sources which are located axially symmetric with respect to the interaural axis [Bla 97]. Although for narrow-band signals these features might be absent, the listener wins further information by delicate head movement or posture against the sound source to promote the localization, as long as the signal is long enough to allow exploratory head movements (at least 200ms)[Bla 08]. A thorough overview of studies carried out on human's spatial hearing can be found in [Bla 97]. There have been several studies concerning the localization blur. Localization blur, as a criterion of human's localization accuracy, is the smallest change in the attributes of the sound source which leads to a just noticeable change in the location of the perceived auditory event. The results of these studies, which were done with different signals (clicks, sinusoids, narrow and broad band noise or speech) implied that the most precise spatial hearing lies close to the intersection of median and horizontal plane. The minimum value for localization blur on the horizontal plane, frontal direction, is about 0.75° for clicks signals [1]. For deviations from median plane to left and right on the horizontal plane, the localization blur increases to between three to ten times of the value found for the forward direction ([Bla 97], p. 40). For sound sources located on the median plane and with deviation in the elevation from the forward direction, the localization blur changes from 4° for white noise to approximately 17° for continuous speech of an unfamiliar person ([Bla 97], p.44). For free field sound propagation the interaural features cannot be investigated separately, but by dichotic representations via headphones these attributes can be studied individually. For

[1] See table 2.1 in [Bla 97], p. 39

these cases, instead of the exact sound source location, usually the lateral displacement of the sound source with respect to the median plane is detected. To determine the importance of ILD and ITD with respect to spatial hearing, the lateralization blur of these two parameters is considered which is defined as the smallest change in the ILD and ITD that leads to a lateral displacement of the auditory event. The lateralization blur for ILD increases generally with increasing signal levels and is signal and frequency dependent. The smallest lateralization blur of 0.6 dB was reported for a sinusoid of 2 kHz[2]. The lateralization blur for ITD also differs for different signals, signal levels and frequencies. The smallest lateralization blur for ITD was found in the range of 2 to 10 μs for click signals and between 6 to 12 μs for sinusoids and noise[3].

While the binaural cues ITD and ILD correlate with displacements in the lateral direction, the monoaural spectral cues serve the distinguishing among front or rear positions or the angle of the elevation. The combination of these features lead to specifying the azimuth, elevation and distance of the source in the free field. There are also other additional parameters which improve the spatial hearing such as head movements toward the sound source or visual cues [Bla 97]. Furthermore, in case of hearing in rooms, the listener's impression of the room depending on the room acoustical parameters can also contribute to spatial hearing.

2.2 Binaural technology and head-related transfer functions

The aim of binaural technology is to capture the sound signal arriving at both ears of the listener, and reproduce them exactly as they were so that all spatial and spectral aspects of the auditory experience are included [Møl 92]. The head-related transfer functions (HRTFs) as the basis of the binaural technology describe the free field sound propagation between the sound source and the listener's ears. According to [Møl 92, Møl 95] this propagation can be divided into two parts, a direction dependent part and a part independent of the direction. The direction dependent part is defined as the ratio between the pressure at the entrance to the (blocked) ear canal to that measured at the center position of the head with listener absent [Møl 92]. HRTFs include the filtering information due to reflecting and diffracting effects of the listener's head, pinnae, shoulders and torso. By taking the inverse Fourier transform of HRTFs, the equivalent head-related impulse responses (HRIRs) are obtained. For each source located at azimuth θ, elevation φ and distance r in the head centered spherical coordinate system the transfer path to the left and to the right ear can be described with a pair of HRTFs as a function of source position and frequency: HRTF_left(θ, φ, r, f) and HRTF_right(θ, φ, r, f). Naturally, HRTFs include the binaural cues ITD and ILD, and the

[2] See table 2.4 in [Bla 97], p. 161

[3] See table 2.3 in [Bla 97], p. 153

spectral cues can be presented as well. While HRTFs describe the free field sound transmission, in the case of listening in rooms, besides the direct sound arriving at the listener, there are also reflections from the room boundaries, which are added to the direct sound. In this case, sound transmission in rooms is described by BRIRs or BRTFs (Binaural Room Impulse Response - or Transfer Function). Because of simplicity and feasibility, HRTFs and BRTFs are often measured with an artificial head which has the shape and acoustical properties of an average human head. The binaural signals are usually played back with headphones to prevent unwanted disturbing cross-talk which exists between the left and right signals when representing via loudspeakers. Using headphones also makes a reproduction free from reflections in the room [Møl 92]. With the technique of auralization, by convolving the HRTF or BRTF with an anechoic signal, the binaural listening of this signal at a given position in the room can be represented via headphones within a simulated auditory scene [Kle 93]. One well known problem from which the binaural technology suffers, is the errors occuring in sound localization, mainly for sound sources in the frontal hemisphere and on the median plane: These sound sources often appear to be behind the listener or vice versa (front-back errors). These errors can be explained, among other reasons, by the absence of the head movements in binaural signals which improve the front-back localization in real life situations.

2.3 The need for individual HRTFs

Due to the fact that HRTFs include the filtering information caused by subject's head, pinnae, shoulders and torso, it is evident that the HRTFs of different subjects show individualities since subjects are morphologically different, especially with respect to the pinna structure. Experimenting with blind folded listeners with localization tasks Wenzel et al. [Wen 88] have already found individual differences among the subjects in the ability to localize the elevation of the sound source which could be predicted from acoustical characteristics of the subjects' outer ear. In a later study [Wen 93] Wenzel et al. showed that there was an increase in the rate of front-back errors when subjects listened to non-individual HRTF recordings. The listeners however, appeared to be still able to obtain useful directional cues from non-individual HRTFs for localizations in the horizontal plane. Similar results came out from studies of Møller et al. [Møl 96]. They compared the localization performance in real life to the case of listening to individual and non-individual binaural recordings and indicated that front-back errors occur commonly in the median plane where there is a lack of binaural cues. Consequently, they concluded that there are similar error rates for real life and simulations with individual recordings. However, when using non-individual HRTFs, there were significant distant errors and errors in the median plane. The increased rate of errors for median plane is expected since pinna is the most individual morphological element and provides important spectral cues for vertical localization as well as localization in median plane [But 77]. Algazi et al. [Alg 01a] acquired high spatial resolution HRTFs of 45 subjects at 1250 directions and included the anthropometric measurements for each subject as well. The study of the correlation between anthropometry and spectral and temporal features of HRTFs showed that the maximum ITD is strongly correlated with the head size. This, together with the fact that the dependence of the ILD on frequency actually varies with the head size show that, when listening to non-individual HRTFs, although the spectral cues are the main source of errors, they are not the only features which cause localization errors. Algazi et al. [Alg 01a] also found that because of small correlations between anatomical features and accurate pinna dimensions, it is not easy to estimate the pinna dimensions from head and torso measurements. There have also been attempts to find an idealized artificial head with HRTFs which minimize the localization errors. However, the study of Møller et al with 8 different artificial heads showed that the localization performance with HRTFs measured on artificial heads and with randomly chosen human HRTFs are of the same order [Møl 97]. According to experiments with speech stimuli carried out by Begault et al. [Beg 00], adding synthesized early and diffuse reflections to the HRTFs reduces the in-head localization errors. They also showed that front-back errors can be reduced by supplying head motion cues to listeners with head tracking. Völk et al. [Vol 08] repeated a similar study with MLS signal stimuli and measured BRTFs instead of using synthesized ones and concluded that the presence of reverberations in the impulse response can only improve the externalization in case of acquisition with a human head, whereas the use of artificial heads doesn't serve significant improvements. Finally, it should be pointed out, that the human is capable of adapting to non-individual HRTFs through training. Shinn-Cunningham et al. [Shi 98] used feedbacks to train the subjects. After the training phase, subjects showed smaller errors although the error-pattern of localization remained the same. The authors supposed that there might be some limitations on the plasticity of subjects and concluded that the adaptation also depends on the presented stimuli. Blum et al. [Blu 04] reported that it might be possible to adapt the auditory system to the non-individual HRTF by letting the subject participate interactively to explore his own entire auditory sphere. Zahorik et al. [Zah 06] trained their subjects to remediate the front-back reversals with improvements that lasted at least 4 months after training. The other study by Parseihian et al. [Par 12b] showed the ability of people to adapt in localizing virtual sound sources when listening to non-individual HRTFs with improvements in elevation localization, without necessarily needing a visual feedback. Nevertheless, the process of training the subjects to non-individual HRTFs poses a complicated and time-consuming task.

2.4 Trends in HRTF customization

The fact that individual HRTFs can contribute to large extends to the improvement of localization while listening to binaural recordings have prompted many researchers to the challenge of individualization. The major problem regarding doing the measurements for real persons individually is the very long and time consuming measurement process. In the most straight forward acquisition method, HRTFs are measured for every source position separately by moving the source after one measurement point to the next in the form of a stop & go measurement. The subject keeps still during the whole measurement to avoid artifacts due to head movements. In most practical cases, an array with loudspeakers at fixed elevations is used which is rotated with respect to the subject to different azimuths. This has the advantage that, as long as the measurement is done for a given azimuth, no extra time will be spent to set the new position of the source. However the situation can still get difficult for the subject if high resolution HRTF grids are aimed. Zotkin et al. [Zot 06] proposed the idea of applying the acoustical principle of reciprocity, which implies that the measured impulse response of the acoustical path between source and microphone will be the same if the source and microphone positions are exchanged. Therefore, by inserting miniature loudspeakers in the subject's ears and capturing the sound simultaneously by a microphone array, the HRTF measurement can be accelerated. The use of a microphone array instead of a loudspeaker array also reduces the inter-equipment reflections. Zotkin et al. tested the directly and reciprocally measured HRTFs and concluded that the two results agreed to good extends. One problem with reciprocally measured HRTFs however is the resulted weak Signal to Noise Ratio (SNR), since for subject's comfort and for physiological safety there are limitations in the amplitude of the signal emitted by in-ear loudspeakers. Although acoustical isolations can be used between the eardrum and the speaker, another problem arises from the size of the miniature sound source. Small loudspeakers have weak performance at lower frequencies. Poor results at lower frequencies might be a problem which concerns HRTF measurements generally as the HRTFs are often windowed to remove the reflections in the measurement room. However, the validity of the HRTF data set acquired reciprocally by Zotkin et al. was for frequencies above 1.5 kHz. Therefore, analytical methods had to be performed to compensate for the lack of content at lower frequencies. According to the study by Algazi et al. [Alg 01b] the origin of the low frequency localization features for different elevations is the reflections and diffractions from head and torso and the effect of pinna structure appears only for frequencies above 3 kHz. Therefore, for low frequencies HRTFs can be well approximated by simple geometric models of head and torso [Alg 02].

Another approach to the customization of HRTFs leaves the area of directly acoustic measurement and takes advantage of the knowledge of the anthropometry of the subjects. Katz

[Kat 01] used models of head and pinna to calculate the HRTFs numerically with Boundary Element Method (BEM). This method is able to change the geometry of the subject during the experiment, which is not possible in measurements with real persons. In addition, the BEM method is a good solution for measuring the HRTFs of small children due to difficulties involved during direct measurement with very young subjects [Fel 04]. The major constraint of BEM method is that the upper frequency, for which the calculation is valid, depends on the size of the elements. If the fine structure of the subject's outer ear and pinna should be considered, smaller elements are required. On the other hand the size and number of the elements determine the speed of the calculations and the memory requirements. For this reason and in order to avoid enormous data sets usually only the head, neck and pinna are modeled and shoulders and the torso are neglected. Calculations can get faster by applying the theory of reciprocity to the numerical calculations [Fel 04] or by use of multipole accelerated BEM and its spherical harmonic representation [Gum 10]. Otani et al. [Ota 06] mentioned that applying reciprocity to BEM might lead to different results in comparison to normal BEM and suggested the possibility of performing some parts of calculations, which are independent of the source position, in advance. Despite these attempts to speed up the calculations, another problem is the need for special laser scanners to acquire the exact model of subjects head and pinna, which poses extra financial burden. There are also attempts to construct models between HRTF features and anthropometry, as done by Jin et al. [Jin 00] or Rothbucher et al. [Rot 10a]. Given an existing HRTF data set, which also includes the anthropometric data of the subjects, this set can be used to train for example a linear regression model. However, for this end, a collection of HRTF of various subjects with corresponding anthropometric data is definitely required [Rot 10b].

Another field of study concerns the simultaneous or semi-parallel measurement of sound sources to reduce the measurement time. González et al. [Gon 04] proposed a general method for multichannel simultaneous linear impulse response measurement which is based on frequency-multiplexing. Using multi-tone signals, different groups of interleaved frequencies are allocated to different channels and due to orthogonality of signals the information corresponding to each channel can be separated later. Majdak et al. [Maj 07] developed the Multiple Exponential Sweep Method (MESM), a method for system identification of weakly non-linear systems excited with exponential sweeps. This method benefits from the fact that by exciting the system with logarithmic sweeps, linear and non-linear parts of the response can be separated. In this method the excitation signals for different channels are interleaved in time. By starting each sweep one after another, the linear impulse responses are located between the linear and the first harmonic impulse response corresponding to the next sweep. Dividing the existing channels into groups of interleaved systems, these can furthermore be overlapped by starting the next group before the last sweep of the present group has reached its end. Majdak et al. tested the MESM with an array of loudspeakers to measure 22 elevations within ca 7 seconds, reducing the measurement time by a factor of five, and showed that MESM performs excellently robust against non-linear distortions, as long as the systems retain their weakly nonlinear performance. Weinzierl et al. [Wei 09] introduced a more general multiple sweep measurement with sweeps which were spectrally colored according to the present background noise to improve the SNR and concluded that this method is advantageous against MESM for measurements with room impulse responses longer than 2 seconds. Masiero et al. [Mas 11] used also interleaved sweeps as excitation signal with the subject standing on a turntable inside a vertical arc of up to 40 loudspeakers and discussed the electro acoustical and mechanical aspects to be considered for an errorless acquisition. Going further than interleaving and overlapping, to yield even shorter measurement times, Dietrich et al. [Die 13a] introduced the optimized MESM which takes advantage of temporal structure of the linear impulse response to place the single harmonics among arbitrary fundamentals.

One reason for the long duration of sequential point to point measurement is the time which is needed between two subsequent measurements to set the new source location. To overcome this limitation, the HRTFs can be acquired by rotating the subject continuously for all azimuthal directions during the measurement. Ajdler et al. [Ajd 07] suggested a system for dynamic measurements with a rotating microphone and a fixed source position to measure all azimuthal angles within one rotation of only 1 second duration. After capturing the excitation signal, different impulse responses corresponding to different angular positions are then reconstructed using the 2D-Fourier representation and by taking into consideration the Doppler effect and compensating for it. In order to have this reconstruction successfully, the rotation speed should be adapted to a revolution time corresponding to a multiple of the period of the excitation signal and the emitted signal should be designed very carefully. Fukudome et al. [Fuk 07] also proposed a measurement system consisting of a rotating chair with a constant angular speed to rotate the subject continuously during the measurement with Maximum Length Sequences (MLS) excitation. The system identification is based on the cross-correlation technique which is commonly used for MLS excitation of LTI systems. However, modifications should be carried out to the period of the excitation signal to adapt the system identification to the rotating time variant system. The system of Fukudome et al. is capable of acquiring the impulse responses for all azimuthal directions within about 1 minute of rotation. Enzner [Enz 08] introduced the continuous azimuth measurement of HRTFs with system identification based on Normalized Least Mean Square (NLMS) adaptive filtering. Similar to the before mentioned methods, the subject of interest with in ear microphones, is rotated continuously with respect to a single sound source fixed at a given elevation. The impulse response corresponding to each angular azimuthal point can be identified using adaptive filtering algorithms [Hay 02] within a rotation time of around 20 seconds. The excitation signal plays an important role in the behavior of the system identification with NLMS adaptive filtering [Hay 02]. Antweiler et al. [Ant 95] showed that perfect sequences (sequences with an impulse-like periodic autocorrelation) lead to the best performance of adaptive filters. On the other hand, sweeps have shown to come off well as excitation signals, having high energy and dealing with the limitations due to non-linear distortions [Mul 08]. Telle et al. [Tel 10] introduced the perfect sweep as excitation signal which offers the properties of sweeps and perfect sequences at the same time. Antweiler et al. [Ant 12] confirmed the better performance of perfect sweeps in comparison to white noise. Enzner developed the system of continuous azimuth acquisition from single channel to the multi-channel case by letting the subject rotate around the vertical axis at the center of a vertically located loudspeaker array. During a single rotation of 360° the loudspeakers produce simultaneously the excitation signal. For the case of optimal multi-channel excitation, Antweiler [Ant 08] introduced the adequate generation of perfect sequences which leads to the perfect multi-channel system identification using NLMS adaptive filters.

2.5 Chapter Summary

In this chapter, the spatial hearing and the involved interaural and spectral features in sound source localization were introduced. Furthermore, head-related transfer functions and their role in the binaural technology were discussed. It was shown that the binaural technology has to deal with artifacts if the HRTFs acquisition and the binaural representation are not done for the same person and it is of interest to acquire HRTFs individually. Due to the long duration of HRTF measurement, especially for high resolution measurement grids, individually HRTF measurement with real persons poses a challenging problem. Different attempts on HRTF customizations were briefly introduced. Besides trends to acquire HRTFs without engaging the subjects directly in the measurements, other methods have been suggested to overcome the difficulty of individually measurements by shortening the measurement duration. The present thesis concerns two of the proposed approaches which aim at reducing the measurement time, namely the multiple exponential sweep method (MESM) and the optimized MESM, proposed by Majdak et al. [Maj 07] and Dietrich et al. [Die 13a] respectively, and the continuous HRTF acquisition based on system identification with NLMS adaptive filtering as proposed by Enzner [Enz 08]. The advantages and limitations of these methods are discussed in the following chapters.

Chapter 3

Multiple Exponential Sweep Method (MESM) and optimized MESM

Majdak et al. [Maj 07] introduced the multiple exponential sweep method (MESM), which can be applied for system identification of multiple weakly nonlinear systems, where only the linear part of the impulse response is of interest. The main idea behind MESM consists in letting the excitation of subsequent systems overlap in time, in order to speed up the measurements. Based on the fundamental principles of MESM, Dietrich et al. [Die 13a] proposed an optimization on MESM, which, under certain conditions can lead to even shorter measurement times than MESM. This chapter treats these two approaches. Since MESM is based on system identification with exponential sweeps, the latter is introduced first shortly in section 3.1. Further on, in sections 3.2 to 3.4, MESM and its strategies as well as optimal choice for its parameters are reviwed. The discussion continues to optimized MESM in section 3.5. According to [Maj 07] and [Wei 09], MESM can also be used to improve the SNR, but the attention of the chapter will remain on MESM's ability to improve the measurement speed. The main focus of this chapter is dedicated to the review, derivation and discussion of the formula and equations envolved in the two methods, as originally presented in the main refrences [Maj 07], [Wei 09] and [Die 13a].

3.1 Exponential Sweep Method (ES)

Using exponential sweeps as excitation signal for system identification was proposed by Farina [Far 00]. The main idea of this method is the possibility of measuring the impulse response of a weakly-nonlinear almost time-invariant system without the need for repeating an extra measurement for nonlinearities, hence the possibility of the simultaneous measurement of impulse response and nonlinear distrotion, which can be applied for room acoustics and audio measurements as well. The straight forward method of measuring the impulse response is to excite the system with a deterministic wideband signal such as random noise sequences or sweeps and acquire the response of the system. The impulse response can then be achieved by spectral division of these two signals and taking the inverse Fourier transform of the result. The time aliasing problems caused by circular deconvolution can be avoided in excitations with sweeps by adding some silent segments to the end of the signal [Far 00]. But the main property of sweeps emphasized by [Far 00] is that in the case of a logarithmic sweep, that is a sweep with the frequency increasing exponentially over time, the higher order nonlinear distortions in the impulse response appear as harmonic impulse responses distinctly separated from the system's linear response. A logarithmic sweep starting at frequency f_1, ending at frequency f_2 and with a total duration of T can be synthesized in time domain as a sinus signal with exponentially varying frequency [Far 00]:

$$x(t) = \sin\left(\frac{2\pi f_1 T}{\ln\frac{f_2}{f_1}}\left(e^{\frac{t}{T}\ln\frac{f_2}{f_1}} - 1\right)\right) \tag{3.1}$$

For a logarithmic sweep, the time with respect to the linear impulse response, at which the k^{th} nonlinear response(k^{th} harmonic response) appears, can be exactly calculated as [Far 00]:

$$\Delta t_k = \frac{T\ln(k)}{c} = \frac{\ln(k)}{r_s} \tag{3.2}$$

with $c = \ln\left(\frac{f_2}{f_1}\right)$. $r_s = \frac{\ln(\frac{f_2}{f_1})}{T}$ is the sweep rate which represents the frequency range of the sweep in octaves normalized to the length of the sweep in seconds [Die 13a]. Figure 3.1 shows the result of the deconvolution of the output of a simulated nonlinear system to excitation with a logarithmic sweep.

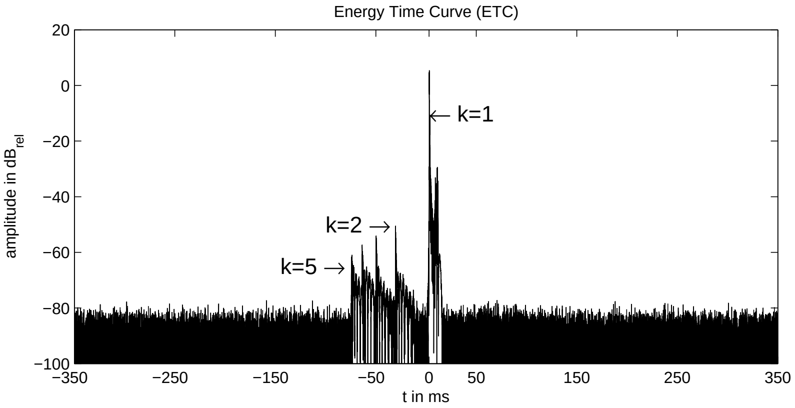


Figure 3.1: Linear impulse response ($k = 1$), and harmonic impulse responses ($k = 2$ to $k = 5$) of a simulated nonlinear system excited with an exponential sweep.

As depicted in Figure 3.1, the nonlinearities appear as similar copies separated from the linear response. Knowing the exact beginning time according to equation 3.2, the linear impulse response and the high order nonlinear responses can be separated using proper window functions. It should be noticed, that since the deconvolution with the excitation signal corresponds to a convolution with the time reversed sweep, the harmonic impulse responses are folded back to negative times [Wei 09].

The logarithmic sweep can also be constructed in frequency domain by designing the amplitude spectrum to increas by 3 dB/octave with a group delay which grows exponentially. It is of interest to maintain a constant temporal envelope for the logarithmic sweep to guarantee for the optimal crest factor of 3 dB for the excitation. This can be achieved by setting the group delay growth proportionally to the power of the logarithmic sweep $|H(f)|^2$ as [Mul 08]:

$$\tau_{gd}(f) = \tau_{gd}(f - df) + C|H(f)|^2 \tag{3.3}$$

with τ_{gd} as group delay, df as the width of each frequency bin and C defined as:

$$C = \frac{\tau_{gd}(f_{end}) - \tau_{gd}(f_{start})}{\sum_{f=0}^{\frac{f_s}{2}} |H(f)|^2} \tag{3.4}$$

In addition to the above mentioned properties, there are also other qualities which make sweeps in general more attractive against pseudo random sequences for acoustical measurement purposes. Sweeps can be constructed for any length and any measurement frequency range. Pseudo random sequences might theoretically have a high energy due to very low crest factor, as they have, in contrast to white noise, only two possible amplitudes of 0 and 1 (binary signals), but in practice, they do not perform as expected. The anti-aliasing filters used in audio analog to digital converters lead to drastic changes in the rectangular wave form of the binary sequences in case of high amplitude excitations. In addition, periodic pseudo random sequences are extremely vulnerable to time variations which limit the use of averages to compensate for low SNRs. In contrast, a single logarithmic sweep of proper length enables achieving sufficient SNR even with a single measurement. In addition, the impact of transient noise appears only within a narrow frequency band of the signal. Sweeps can also be filled up with zeros to a power of 2-length, 2^n $(n \in \mathbb{N})$, suitable to be analyzed by the Fast Fourier Transform (FFT) method. Although FFT analysis performs actually for sequences which are repeated periodically, the single sweep and the periodically repeated one do not show considerable spectral differances [Mul 08].

3.2 Interleaving and overlapping: MESM

If the exponential sweep method is used to identify M systems, the simplest way is to do the identification system by system. In case of an acoustical system identification, besides the length of excitation signal, T, an additional length of τ_{IR} for each system should also be considered in the measurement duration due to the reverberations in the room (the length of the room impulse response). As a result, the measurement duration for M systems using exponential sweep (ES) method is given as:

$$T_{ES} = (T + \tau_{IR})\, M \tag{3.5}$$

Considering M weakly nonlinear systems excited with exponential sweeps, the measurement result of each system will also include a set of separable harmonic responses. If the length of the 2^{nd} order harmonic response, $\tau_{IR,2}$, and the time of its occurrence, Δt_2, (which can be calculated by equation 3.2) are so, that the 2^{nd} order harmonic response fades away before the linear impulse response begins, one can use the remaining time distance to send the excitation signal for the next system. It means, sweeps can be sent semi-parallel in time. This is one important idea of MESM and is named as interleaving. By applying interleaving, after deconvolution, the linear impulse response (IR) of the first system in placed between the IR and 2^{nd} harmonic impulse response (2^{nd} HIR) of the second system. This process can be generalized to as many systems as necessary, as long as there is sufficient distance between the end of the 2^{nd} HIR and the beginning of the IR of the last system. Figure 3.2 shows the response of a nonlinear system to four interleaved sweeps as well as the result of the deconvolution.

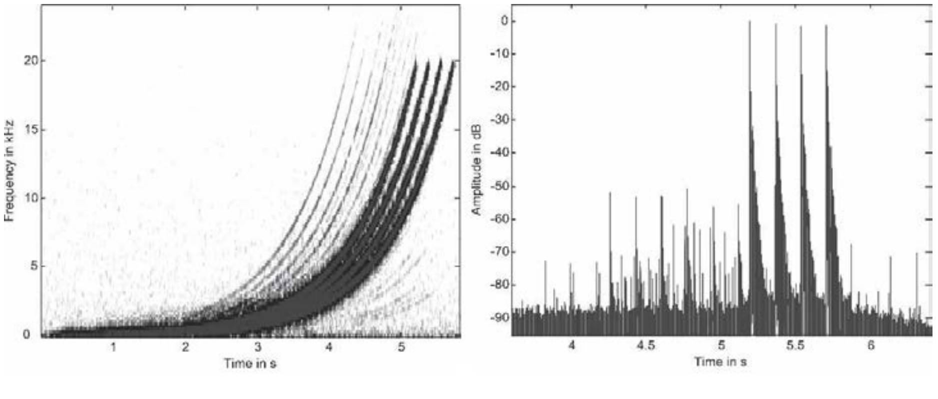


Figure 3.2: Response signal spectrogram of excitation with four interleaved sweeps (left), and the result of deconvolution with IRs (higher peaks) and nonlinear HIRs (lower peaks)(right). [Maj 07]

For a group of η interleaved sweeps, the following condition should be met [Maj 07]:

$$\Delta t_2 - \tau_{IR,2} > \tau_{IR}(\eta - 1) \Rightarrow \Delta t_2 > \tau_{IR,2} + \tau_{IR}(\eta - 1) \tag{3.6}$$

It means, a minimum delay of τ_{IR} should exist between the interleaved sweeps and considering equation 3.2, there is now a minimum duration for the sweep necessary which reads:

$$T' = [\tau_{IR}(\eta - 1) + \tau_{IR,2}] \frac{c}{\ln 2} \tag{3.7}$$

If this minimum length is shorter than the original sweep duration, the sweep length is set to the original length to meet the necessary SNR conditions. If the condition in equation 3.7 holds, each i^{th} sweep is to be played at time $(i-1)\,\tau_{IR}$ and η systems are interleaved. The last sweep begins at $(\eta - 1)\,\tau_{IR}$ and lasts T' seconds. After the last sweep is finished, the measurement should be further extended by another τ_{IR} to capture the reverberations for the last system. As a result, the measurement duration of η interleaved systems is given by:

$$T_{grp} = T' + \eta\tau_{IR} \tag{3.8}$$

If M systems are divided into groups of η interleaved members, the measurement should be repeated $\frac{M}{\eta}$ times and the whole measurement duration will be:

$$T_{INT} = \frac{M}{\eta} T_{grp} = \frac{M}{\eta} T' + M\tau_{IR} \tag{3.9}$$

Comparison between equations 3.5 and 3.9 shows that the interleaving results in a reduction of the measurement duration if the ratio $\frac{T'}{\eta}$ is smaller than T.

[Maj 07] also suggested the idea that, in case of weakly nonlinear systems, where the number of high order harmonic responses is small, it is not necessary to wait until the first sweep is finished to send the second sweep. As shown in figure 3.3, as long as the highest harmonic response of the next sweep does not disturb the reverberation caused by the previous sweep, these two sweeps can overlap in time.

Using this strategy of overlapping and applying deconvolution, the impulse response of individual systems do not interfere with each other, although their excitation overlaps. Overlapping of two consequent sweeps works as long as the beginning point of the highest harmonic response does not interfere with the information contained in the linear part of the system. For this end, the next sweep should keep a minimum distance from the previous one. Assuming that after deconvolution, a maximum number of K_{max} harmonic responses can be recognized,

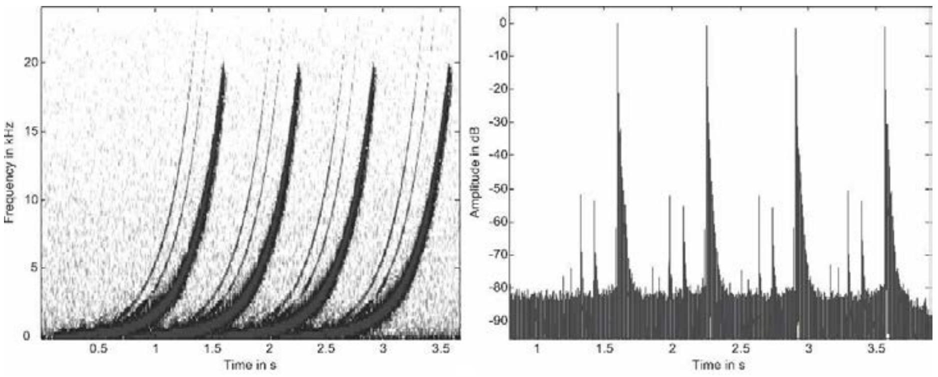


Figure 3.3: Response signal spectrogram for four overlapped sweeps (left) and the IR and HIRs after deconvolution (right). [Maj 07]

the minimum delay between two consequent sweeps is given by [Maj 07]:

$$\tau_{OV} = \Delta t_{K_{max}} + \tau_{IR} = \frac{T}{c} \ln K_{max} + \tau_{IR} \tag{3.10}$$

The measurement duration for M overlapped systems will be [Wei 09]:

$$T_{OV} = T + (M-1)(\Delta t_{K_{max}} + \tau_{IR}) + \tau_{IR} \tag{3.11}$$

One might think that using fast sweeps, the beginning time of the maximum order harmonic response, $\Delta t_{K_{max}}$, could get smaller since for increasing sweep rates, the harmonic responses move closer together. But that would at the same time result in a reduced obtained SNR due to the shorter excitation signal. In addition, it should be noticed that any interference between the linear impulse response and the second harmonic response should be avoided. This sets a constraint for the sweep rate which can be calculated according to equation 3.2 as [Die 13a]

$$r_s \leq \frac{\ln 2}{\tau_{IR,2}} \tag{3.12}$$

The overlapping strategy will result in a reduction in measurement duration if the maximum order of nonlinearity is low enough (weakly nonlinear systems).
Combination of the two strategies of interleaving and overlapping builds the basis of MESM. First, M systems are interleaved in $\frac{M}{\eta}$ groups and then, these groups are overlapped. The measurement duration will be equal to $\frac{M}{\eta}$ times interleaving for η systems, plus the delay for overlapping $\frac{M}{\eta}$ groups with a maximum order of distortion K_{max} [Wei 09]:

$$T_{MESM} = T' + \frac{M}{\eta}\,(\eta\tau_{IR}) + \Delta t'_{K_{max}} \lceil \frac{M}{\eta} - 1 \rceil = T' + \Delta t'_{K_{max}} \lceil \frac{M}{\eta} - 1 \rceil + \tau_{IR} M \tag{3.13}$$

$\lceil x \rceil$ denotes the next higher integer of x. Note that as a result of interleaving, the sweep duration and the begin of the k^{th} harmonic response should be modified by equation 3.7. Figure 3.4 shows the response signal and the impulse responses for overlapping two groups, each containing two interleaved sweeps.

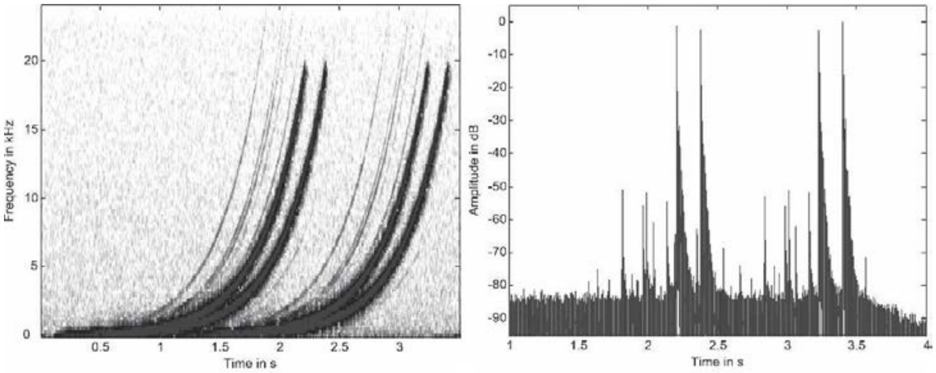


Figure 3.4: Response signal for overlapping two groups, each group containing two interleaved sweeps (left), and the IRs and HIRs (right) for system identification with MESM. [Maj 07]

For a given η, the excitation time for each i^{th} system is given by [Maj 07]:

$$t_i = \tau_{IR}\,(i-1) + \lfloor \frac{i-1}{\eta} \rfloor \Delta t'_{K_{max}} \tag{3.14}$$

$\lfloor x \rfloor$ denotes the next lower integer of x. Therefore, knowing the length of the linear impulse response, (τ_{IR}), these impulse responses can be separated by windowing. MESM can be applied if only the linear impulse response is of interest since the harmonic impulse responses of subsequent systems interfere with each other and cannot be completely separated as in case of ES method [Maj 07]. However, as shown by Torras-Rossel et al. [Tor 11], even the linear impulse response does not stay totally unaffected from nonlinearities and in particular, the odd orders of nonlinearities contaminate the linear response. It means that the ES method can reject a significant amount of distortions from the linear part but after applying a time window, not all nonlinearities can be suppressed from the measurement results.

3.3 Optimization of MESM parameters

Comparing equations 3.5 and 3.13, it can be seen that the improvement in measurement duration using MESM against ES can be evaluated by comparing the two terms TM and $T' + \Delta t'_{K_{max}} \lceil \frac{M}{\eta} - 1 \rceil$:

$$T_{ES} = (T + \tau_{IR})\,M \longleftrightarrow T_{MESM} = T' + \Delta t'_{K_{max}} \lceil \frac{M}{\eta} - 1 \rceil + \tau_{IR} M \tag{3.15}$$

For MESM, the measurement duration depends on K_{max}, τ_{IR}, $\tau_{IR,2}$ and T (or sweep rate r_s). In addition, the number of interleaved systems, η, is important. In order to find an optimal η, which minimizes the duration T_{MESM}, this duration can be rewritten as a function of η by substituting equations 3.2 and 3.7 in equation 3.13 which results in:

$$T_{MESM}(\eta) = \frac{1}{\ln 2}\Bigg[(\tau_{IR}(\eta-1)+\tau_{IR,2})\left(c+\ln K_{max}\lceil\frac{M}{\eta}-1\rceil\right) \\ + M\tau_{IR}\ln 2\Bigg] \tag{3.16}$$

The derivative of T_{MESM} with respect to η yields:

$$\frac{dT_{MESM}}{d\eta} = \frac{1}{\ln 2}\left(\tau_{IR}c + \tau_{IR}\ln K_{max}\lceil\frac{M}{\eta}-1\rceil)\right) \tag{3.17}$$

Since $\frac{M}{\eta}$ $\geq$1, the derivative of T_{MESM} is always positive (for all $\eta \neq 0$) which means that T_{MESM} increases with increasing η. Therefore the optimal number of interleaved systems, η_{opt}, is the minimum η which still meets the condition of minimum sweep length (equation 3.7):

$$\eta_{opt} = \lceil\frac{T\ln 2}{c\tau_{IR}} + \frac{\tau_{IR}-\tau_{IR,2}}{\tau_{IR}}\rceil \tag{3.18}$$

Weinzierl et al. [Wei 09] mentioned the issue that if small values are set for η_{opt}, it might happen that a non interleaved measurement (η=1) with sweep duration T yields a shorter measurement duration than MESM. Therefore, the efficiency of finding an optimal η for MESM relative to conventional measurement is considered as $\nu = \frac{T_{conventional}}{T_{MESM}}$ [Wei 09]. Figure 3.5 shows the efficiency of an optimal MESM measurement, also named as measurement acceleration, as a function of number of systems N and the length of linear impulse response τ_{IR}. It can be seen that finding an optimal η is not much beneficial if the linear impulse response contains long reverberations, but for large numbers of systems and small τ_{IR}, MESM results in a considerable improvement of measurement duration.

3.4 Calibration measurement for MESM

According to equations 3.7 and 3.14, the length of linear and second harmonic responses, (τ_{IR} and $\tau_{IR,2}$), as well as the maximum order of nonlinearities, K_{max}, are required to calculate the MESM parameters, which are the sweep duration, T', and the excitation time for each sweep, t_i. This necessitates measurements which should be done prior to MESM measurement. This calibration measurement can be done using an un-optimized method such as ES method, in order to read the length of linear and second harmonic responses as well as the maximum number of harmonics, which can be recognized in measurement results. Since the length of

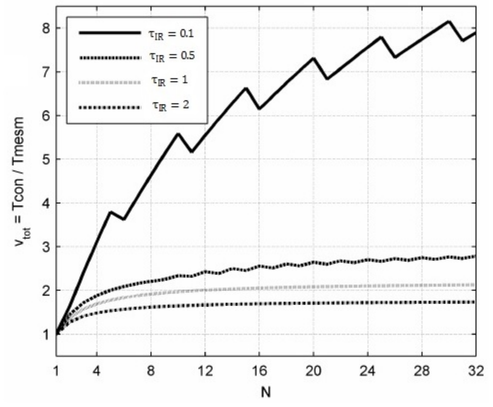


Figure 3.5: Measurement acceleration: MESM measurement duration for different values of τ_{IR} ($\tau_{IR,2} = 0.1\tau_{IR}$) , $K_{max} = 5$, $\frac{f_2}{f_1} = 400$) relative to the duration of a conventional ES measurement with sweep length T=3s as a function of N(number of systems)[Wei 09](Parameter τ_{IR} is named L_1 in [Wei 09]. The renaming was done for uniformity).

the impulse response depends on sound absorption of walls, floors and ceiling of the room, the lengths τ_{IR} and $\tau_{IR,2}$ do not make considerable differences for different systems. Parameter K_{max} depends on the degree of nonlinearities of each system. However, the variability of K_{max} will be reduced if the measurement equipment is of the same type for all channels. So, after repeating the un-optimized measurement for each system separately, the maximum values for τ_{IR}, $\tau_{IR,2}$ and K_{max} among all values are used to calculate the MESM parameters. In the case of HRTF measurement, the calibration measurement is done in the absence of the subject with the microphone placed in the center position of his head whereas the MESM measurement is performed with microphones in the ear canal. Switching from calibration to MESM measurement however does not change the length of the impulse response since the length of HRTF is short in comparison to room impulse response [Maj 07]. But one aspect to be considered is that, when sweeps are played in a semi-parallel manner, as in MESM, the microphones capture the summation result of the existing sound in the room. Since the amplitude of this summation is higher than the amplitude in single separate measurements, there might be the risk of further clipping in measurement equipment and K_{max} could differ from the value obtained from calibration measurements. Therefore, the amplitude of excitation signals should be modified for MESM. Majdak et al. [Maj 07] suggested to perform the calibration measurement with a sweep duration and amplitude which fulfills the SNR requirements and lets enough headroom for each of the systems and achieve the MESM parameters, and in the next step, to repeat the measurement again, this time with MESM to adapt the amplitude to have enough headroom for the case of semi-parallel excitation. This MESM measurement might reduce the SNR since the amplitudes are lowered to avoid clipping but an extended duration for sweep will then be used to increase the SNR again to the desired value. Finally, this eventually extended sweep length from the very recent measurement is used, together with the values τ_{IR}, $\tau_{IR,2}$ and K_{max} from the first calibration measurement, to calculate the MESM parameters.

It should be mentioned again that the results of MESM measurement can be compared temporally and spectrally to conventional measurement if the system is weakly nonlinear. Dietrich et al. [Die 13a] also pointed out that the level must be kept constant between actual and calibration measurements. Once the loudspeakers cause the nonlinearities, other equipment (microphones and amplifiers) should work in linear range only.

Majdak et al. [Maj 07] carried out a MESM measurement with $\tau_{IR} = 100$ ms, $\tau_{IR,2} = 10$ ms and K_{max}=5. η_{opt} and the extended sweep duration T' were calculated as 3 and 1.815 s respectively. A number of M= 22 channels were measured within 7.1 seconds, which, compared to a sequential measurement with a sweep length of T=1.5s, resulted in an improvement of measurement speed by a factor 5. By rotating to the next azimuths in 2.5° steps, a total number of 1550 points were measured within 10 minutes. If time variances are considered as a stochastic process, MESM shows, due to a shorter measurement time, less vulnerability to time variance artifacts and performs robust against nonlinear distortions [Maj 07].

3.5 Optimized MESM

As discussed in the previous sections, compared to conventional ES measurements, MESM can reduce the measurement duration for large numbers of systems and shorter reverberation times. Dietrich et al. [Die 13a] proposed a generalized overlapping strategy which leads under certain conditions to even better results than MESM. This strategy takes advantage of temporal structure of impulse responses as well as the length of harmonic responses, and is called optimized MESM by authors. The idea of optimized MESM corresponds generally to overlapping as for MESM but without any interleaving. Actually, the formula for measurement time as defined in equation 3.11 is also used for optimized MESM but with some modifications:

$$T_{OPT} = T + (M-1)\tau_w + \tau_{st} \tag{3.19}$$

τ_{st} is the time which the last system needs to decay after the sweep has stopped. This time is chosen for safety as long as one impulse response τ_{IR}.

The difference to equation 3.11 is that the minimum time delay between consequent excitations, τ_w, is not given by equation 3.10 anymore. The reason is the temporal structure of the impulse response. When measuring the transfer function of acoustical path as a Device Under Test (DUT), the resulted impulse response contains, besides the direct sound, also reflections from objects in the room and from the room itself. For the case of HRTF measurement, the

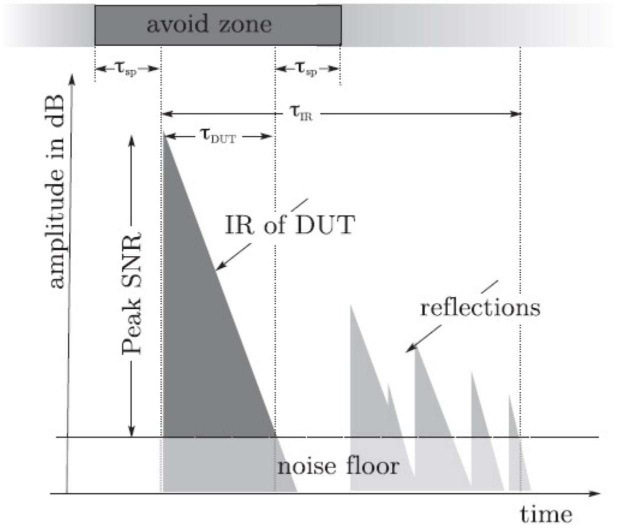


Figure 3.6: Temporal structure of an impulse response measured with exponential sweep. [Die 13a]

direct sound as well as early reflections are important and the rest of reflections, which might even occur in anechoic environments due to objects in the room, do not supply any useful information. It means that it is not the whole length of impulse response, τ_{IR}, which should be protected against harmonic responses or reflections. It is sufficient to protect the useful part of the impulse response, τ_{DUT}. As shown in figure 3.6, an avoid zone can be put around the useful impulse response, suggested as an optional safety time ,τ_{sp}, before and after the impulse response of DUT. The percentage of the length of the useful impulse response is defined as:

$$\alpha = \frac{\tau_{DUT} + 2\tau_{sp}}{\tau_{IR}} \leq 1 + \frac{2\tau_{sp}}{\tau_{IR}} \tag{3.20}$$

If $\tau_{sp} = 0$, it is clear that for α=1, $\tau_{DUT} = \tau_{IR}$ and no improvement in measurement time is achieved. Considering this strategy, the optimized MESM corresponds to a MESM measurement with only overlapping, named as adapted overlapping [Die 13a], for which the time delay between the sweeps is calculated by:

$$\tau_{w,adapted} = \Delta t_{K_{max}} + \tau_{DUT} + \tau_{sp} \tag{3.21}$$

The next strategy of optimized MESM is based on the idea that it is sufficient for harmonic impulse responses to be placed in a manner that they do not interfere with the avoid zones. [Die 13a] used two conditions for this end. First: for each harmonic response, the beginning time Δt_k should take place after the end of an avoid zone:

$$(-\Delta t_k \quad mod \quad \tau_w) \geq \tau_{DUT} + \tau_{sp} \tag{3.22}$$

Second: each harmonic response of length $\tau_{IR,k}$ should end, before the next avoid zone begins:

$$(-\Delta t_k \quad mod \quad \tau_w) + \tau_{IR,k} \leq \tau_w - \tau_{sp} \tag{3.23}$$

The combination if equations 3.22 and 3.23 implies that:

$$\tau_{DUT} + \tau_{sp} \leq \left(-\frac{\ln k}{rs} \quad mod \quad \tau_w\right) \leq \tau_w - \tau_{sp} - \tau_{IR,k} \tag{3.24}$$

To fulfill the constraint in equation 3.24 two parameters should be found: τ_w and r_s. At the same time, there exists a minimum limit for τ_w. For the case, where there are no distortions, the only issue which limits the time delay is the presence of reflections so that the time distance between two avoid zones should be at least as long as the length of impulse response: $\tau_w \geq \tau_{IR}$. In the presence of harmonic distortions, this limit should be extended to avoid the interference of the maximum harmonic response with the avoid zone. So, the other constraint for τ_w besides equation 3.24 is:

$$\tau_w \geq max\left(\tau_{DUT} + 2\tau_{sp} + max(\tau_{IR,k}), \tau_{IR}\right) \tag{3.25}$$

Another strategy of optimized MESM takes advantage of weakly nonlinearity of systems, which implies that the length of the harmonic responses, $\tau_{IR,k}$ decreases with increasing order. One way to obtain these lengths is to read them from calibration measurements. The values $\tau_{IR,k}$ depend on signal to noise ratio. If a harmonic response vanishes below the noise floor, it is not considered at all since in this case $\tau_{IR,k} < 0$. It is important that the SNR for main measurements should not exceed the SNR at calibration measurements. Besides direct temporal measurements, [Die 13a] also suggested considering the energy decay of each harmonic response instead of its duration in time domain. If the energy decay of k^{th} harmonic response is a_k and if the decay rate is considered as unchanged for all harmonic responses, then the length of k^{th} harmonic response will be:

$$\tau_{IR,k} = \frac{SNR - a_k}{SNR}\tau_{IR} \tag{3.26}$$

In the worst case, there is no decay in the energy content of harmonic responses, which means that $\tau_{IR,k} = \tau_{IR}$. However, if the systems are weakly nonlinear, the condition $\tau_{IR,k} < \tau_{IR}$ will contribute to shorter measurement times.

As there is no analytic solution to find the optimal values for τ_w and r_s, which fulfill the constraints in equations 3.24 and 3.25, different possible values for τ_w and r_s are demonstrated in a normalized search space ($r_s\tau_{IR}$, $\frac{\tau_w}{\tau_{IR}}$) in order to be studied. The normalization with respect to the length of the room impulse response makes the analysis independent of this length

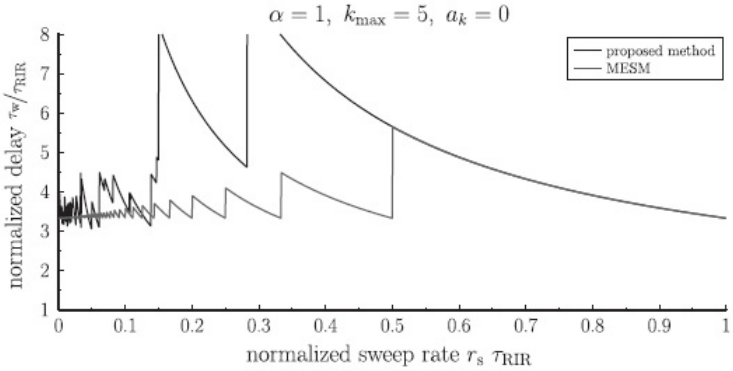


Figure 3.7: Minimum normalized delay for MESM (with $\tau_{IR} = \tau_{DUT}$) and optimized MESM for $\alpha = 1$ and $a_k = 0$. [Die 13a]

[Die 13a]. Figure 3.7 shows such a search space [1]. The goal is to find the minimum time delay, τ_w, between consequent sweeps, which leads to the minimum measurement time according to equation 3.19. Here, the minimum delay is compared between a MESM measurement (with interleaving, overlapping and η_{opt}, but considering $\tau_{IR} = \tau_{DUT}$), and an optimal MESM measurement, which in this case corresponds to a MESM measurement with only overlapping, since α=1 and a_k=0, it means that $\tau_{IR,k} = \tau_{IR}$ for all five harmonic responses. As can be seen, for $r_s\tau_{IR} > 0.5$, MESM and the proposed method have exactly the same manner, since for increasing sweep rates the harmonic responses move closer together so that η=1, which implies a MESM with only overlapping strategy. Using optimized MESM for α=1 and a_k=0 leads to only slightly smaller values of τ_w for a narrow range of sweep rates.

In the next step, the influence of the parameter α is considered in figure 3.8. The comparison is done between following measurement situations: original MESM, MESM, for which the overlapping is done considering equation 3.21, labeled as adapted MESM, and the proposed method, for which $a_k = 0$ but $\alpha < 1$. In this case the proposed method improves the measurement time for small values of α and low sweep rates.

And next, the effect of the length of harmonic responses is studied (figure 3.9). The attenuation, a_k, is chosen to be a_k=20dB and a_k=40dB for every order k. It is again assumed that $\alpha = 1$. The curve labeled as reference MESM corresponds to original MESM for which $a_2 = 0$. For the adapted MESM a_2=20dB or a_2=40dB is considered. In this case the proposed method leads to a shorter measurement duration only for large a_k and small sweep rates.

Finally, figure 3.10 shows the results for combinations using realistic values for α and a_k. The reference MESM refers to a normal MESM measurement without taking advantage of a shorter

[1] In figures 3.7 to 3.10, the normalization in both axes is labeled with τ_{RIR} which should actually be τ_{IR}, since the room impulse response and the useful part of the impulse response are named as τ_{IR} and τ_{DUT} within the paper's text in [Die 13a].

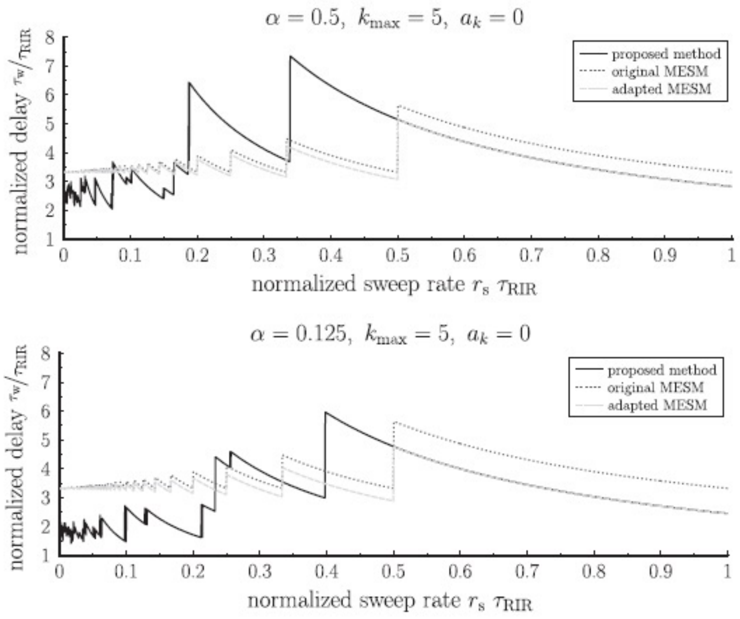


Figure 3.8: Minimum normalized delay for original MESM, MESM with overlapping considering equation 3.21 (adapted MESM) and optimized MESM with $a_k = 0$, $\alpha < 1$ for $\alpha = 0.5$(top) and $\alpha = 0.125$(bottom). [Die 13a]

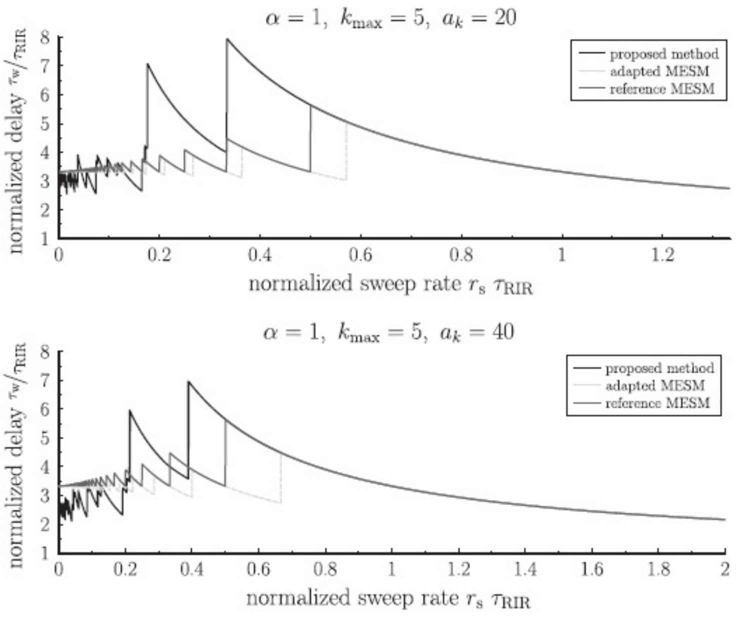


Figure 3.9: Minimum normalized delay for referecne MESM (MESM with $a_2 = 0$), adapted MESM (MESM with a_2=20dB or 40dB) and proposed method (optimized MESM) for a_k=20dB (top) and a_k=40dB (bottom). [Die 13a]

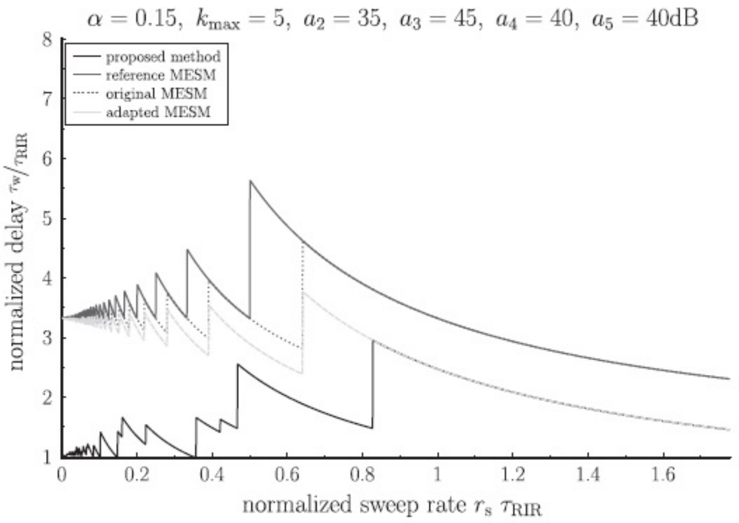


Figure 3.10: Minimum normalized delay for reference MESM, original MESM (MESM with $\tau_{IR,2} < \tau_{IR}$), adapted MESM (MESM with $\tau_{IR,2} < \tau_{IR}$ and $\alpha < 1$) and optimized MESM. [Die 13a]

second harmonic response. The original MESM uses only decreased second order harmonic lengths. The adapted MESM uses decreased second harmonic lengths as well as $\alpha < 1$ and the proposed method corresponds to optimized MESM. It is obvious that the proposed method outperforms all other methods for normalized sweep rates approximately below 0.8.
As a result, optimized MESM improves the measurement speed if the length of the measured impulse response is small compared to the length of the room impulse response (if α<1), and if the smaller length of the harmonic impulse responses are considered. In figure 3.10, the curve labeled as proposed method shows the results for a measurement done with a system consisting of M= 40 loudspeakers. τ_{DUT}= 4ms and τ_{IR} = 40ms were considered. Achieving the best combination of r_s and τ_w as 5.59 and 48.09 ms respectively, Dietrich et al. measured the impulse responses for all 40 loudspeaker locations within 3.25 seconds.

It should be pointed out here that for optimized MESM there is only one delay time, τ_w, which is needed to be found whereas for original MESM two delay times, namely the delay between the interleaved sweeps, (τ_{IL}), and the delay between the overlapped sweeps, (τ_{OV}), are required. Therefore, the comparison between the original and the optimized methods is done under the assumption that the number of systems, M, is large enough, because, if this number tends to infinity there can be one mean delay considered for original MESM which is given as [Die 13a]:

$$\bar{\tau}_{w,MESM} = \frac{(\eta - 1)\tau_{IL} + \tau_{OV}}{\eta} \tag{3.27}$$

The minimum value for τ_{IL} is the length of linear impulse response, τ_{IR}.
As long as the length of the measured impulse response is much smaller than the time delay between two subsequent sweeps, and as long as the number of harmonic impulse responses is small (weakly nonlinear systems), with the consideration that no further weak nonlinearities appear during the measurement, MESM and optimized MESM result in measured impulse responses with the same SNR and spectral and temporal structure as obtained by a sequential ES measurement [Die 13a].

3.6 Chapter summary

This chapter introduced and reviewed MESM by Majdak et al. [Maj 07] and optimized MESM proposed by [Die 13a]. According to [Maj 07] and [Wei 09], MESM leads to the improvement of measurement speed if a large number of channels are used to measure an impulse response of small length. The optimal choice of MESM parameters and the guidance to calibration measurements to set these parameters were reviewed. In addition, the even better performance of optimized MESM against original MESM was explained. Optimized MESM improves the measurement speed for the case that a small percentage of measured impulse response in comparison to the whole room impulse response is of interest and takes advantage of the shorter length of the harmonic responses in comparison to the linear response as well. Following the instructions and considering the constraints discussed in the chapter, MESM and optimized MESM result in impulse responses with the same spectral and temporal structure as measured with a sequential ES method.

Chapter 4

System identification with Normalized Least Mean Square (NLMS) adaptive filters

LMS and NLMS adaptive filters have gained popularity among other adaptive filtering methods due to their simplicity and ease of implementation. The NLMS is the most used adaptation algorithm in acoustic echo cancelation [Ben 01]. Combined with perfect sequences as excitation signals, which improve the adaptation speed, the NLMS approach can be applied to track the response of time varying systems and is used in single- and multichannel system identifications. Enzner [Enz 08], [Enz 09] proposed the system identification with NLMS adaptive filtering to measure HRTFs for all azimuths by rotating the subject continuously in the horizontal direction. This method can also be extended to the simultaneous measurements of more than one sound source. This chapter begins with a brief introduction in the LMS and NLMS algorithms in section 4.1. Section 4.2 deals with the conditions which guarantee the stability of the NLMS algorithm. In Sections 4.3 and 4.4 the conditions are discussed, which lead to the optimal excitation of NLMS algorithm. The HRTF measurement systems proposed by Enzner [Enz 08], [Enz 09] are described in sections 4.5 and 4.6. Finally, section 4.7 introduces a measureable criterion to judge the accuracy of system identification of time varying HRTF measurement systems with the NLMS algorithm.

4.1 LMS and NLMS adaptive filters

Least Mean Square algorithm (LMS) is a special implementation of the method of steepest descent. The method of steepest descent is recursive, it means that its formulation is based on a feedback system with a filter computation which proceeds iteratively step by step. According to [Hay 02], the basic idea of this method is to find an optimal solution $\mathbf{w}_0$ among some other unknown vectors $\mathbf{w}$, which satisfies the following condition:

$$J(\mathbf{w}_0) \leq J(\mathbf{w}) \qquad \text{for all } \mathbf{w} \tag{4.1}$$

The cost function $J(\mathbf{w})$ is a continuously differentiable function of $\mathbf{w}$. In case of the LMS algorithm, the cost function is the mean square of the deviation of the response of the filter $\mathbf{w}$ to a desirable response. For adaptive filtering, a well suited condition is to assume that the cost function is reduced at each iteration:

$$J(\mathbf{w}(k+1)) < J(\mathbf{w}(k)) \tag{4.2}$$

For the LMS algorithm, as a special case of the method of steepest descent, the adjustment applied to the vector $\mathbf{w}$ is in a direction opposite to the gradient vector of the cost function, $\bigtriangledown J$ [1], and is described as [Hay 02]:

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \frac{1}{2}\mu \bigtriangledown J \tag{4.3}$$

μ>0 is called the step size. Step size has an important role in the behavior of the LMS and NLMS adaptive filters, as will be discussed in the next sections. The difference between LMS and the method of steepest descent is that the latter uses exact measurements of the gradient vector at each iteration whereas the LMS relies on an estimation of the gradient vector. The LMS algorithm consists of three general steps [Hay 02]:

1. Computing the output of the filter ($\hat{y}(k)$) in response to the input signal ($\mathbf{p}(k)$): $\hat{y}(k) = \mathbf{h}^T(k)\mathbf{p}(k)$
2. Generating an estimation error ($e(k)$) by comparing this output with a desired response ($y(k)$): $e(k) = y(k) - \hat{y}(k)$
3. Adjustment of the parameters of the filter in accordance with the estimation error: $\mathbf{h}(k+1) = \mathbf{h}(k) + \mu\mathbf{p}(k)e(k)$

Because of the estimation used in LMS algorithm, vector $\mathbf{h}(k)$, which is not the same as $\mathbf{w}(k)$, performs a random motion around the minimum point of the error. As a result, the LMS suffers from a gradient noise. Since the adjustment of the filter depends directly on the input signal $\mathbf{p}$, this gradient noise gets worse for large inputs. The Normalized Least Mean Square Method (NLMS) overcomes this problem by normalizing the adjustment at iteration $k+1$

[1] For a little more detailed, but still summarized discussion on gradient vector as well as the derivation of equations 4.4 and 4.5, see appendix A

with respect to the squared Euclidean norm of the input signal at iteration k. Therefore, the iterative adapting process of NLMS will be:

$$\mathbf{h}(k+1) = \mathbf{h}(k) + \frac{\mu}{\|\mathbf{p}(k)\|^2}\mathbf{p}(k)e(k) \tag{4.4}$$

with:

$$e(k) = y(k) - \mathbf{h}^T(k)\mathbf{p}(k) \tag{4.5}$$

4.2 Stability of the NLMS algorithm

One way to interpret the adaptation process and stability of the NLMS algorithm is the use of geometric interpretation as in [Som 89]. To this end, consider the block diagram of a single channel system identification shown in figure 4.1.

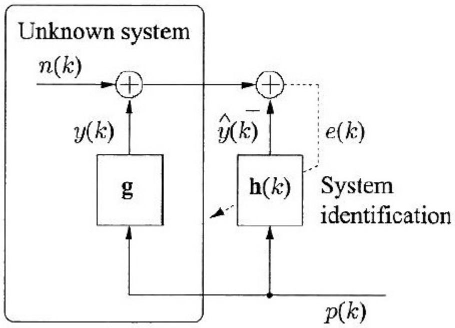


Figure 4.1: Single channel system identification. [Ant 08]

g depicts the impulse response of the unknown system to be identified as the estimation **h**(k), and **p**(k) is the input vector. n(k) represents the influence of the environmental noise on the adaptation process. Using the new notations of figure 4.1, the recursive updating process of the NLMS algorithm is rewritten as:

$$\mathbf{h}(k+1) = \mathbf{h}(k) + \frac{\mu}{\|\mathbf{p}(k)\|^2}\mathbf{p}(k)e(k) \qquad \textit{with} \quad \mu : \text{step size} \tag{4.6}$$

with:

$$e(k) = (\mathbf{g} - \mathbf{h}(k))^T\,\mathbf{p}(k) + n(k) = \varepsilon(k) + n(k) \qquad \text{(error signal)} \tag{4.7}$$

$\varepsilon(k)$ is the mismatch between the desired output $y(k)$ and its estimation $\hat{y}(k)$. The distance vector, which defines the mismatch between **g** and **h**(k), is defined as:

$$\mathbf{d}(k) = \mathbf{g} - \mathbf{h}(k) \tag{4.8}$$

Assuming that $n(k) \equiv 0$ (noiseless environment), the error signal in terms of the distance vector is reduced to:

$$e(k) = (\mathbf{d}(k))^T \mathbf{p}(k) \tag{4.9}$$

Considering equations 4.8 and 4.9, the NLMS algorithm of equation 4.6 changes to:

$$\mathbf{d}(k+1) = \mathbf{d}(k) - \mu \frac{(\mathbf{d}(k))^T \mathbf{p}(k)}{\|\mathbf{p}(k)\|^2} \mathbf{p}(k) \tag{4.10}$$

The geometric representation uses a vector space representation which decomposes the distance vector into two components [Ant 08]:

$$\mathbf{d}(k) = \mathbf{d}^{\perp}(k) + \mathbf{d}^{\|}(k) \tag{4.11}$$

As shown in figure 4.2, the parallel component $\mathbf{d}^{\|}(k)$ can be interpreted as the orthogonal projection of $\mathbf{d}$ on to the input signal $\mathbf{p}(k)$:

$$\mathbf{d}^{\|}(k) = \frac{(\mathbf{d}(k))^T \mathbf{p}(k)}{\|\mathbf{p}(k)\|} \frac{\mathbf{p}(k)}{\|\mathbf{p}(k)\|} = \frac{(\mathbf{d}(k))^T \mathbf{p}(k)}{\|\mathbf{p}(k)\|^2} \mathbf{p}(k) \tag{4.12}$$

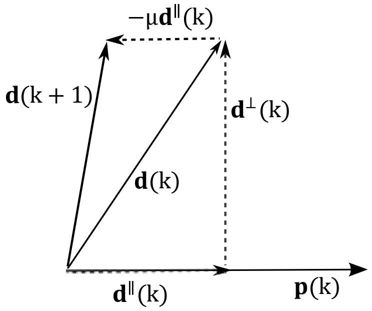

$-\mu\mathbf{d}^{\|}(\mathrm{k})$
$\mathbf{d}(\mathrm{k}+1)$
$\mathbf{d}^{\perp}(\mathrm{k})$
$\mathbf{d}(\mathrm{k})$
$\mathbf{d}^{\|}(\mathrm{k})$
$\mathbf{p}(\mathrm{k})$

Figure 4.2: Geometric interpretation of the NLMS algorithm (Idea of the figure was adopted from [Ant 08]).

which, in combination with equation 4.10 leads to:

$$\mathbf{d}(k+1) = \mathbf{d}(k) - \mu\mathbf{d}^{\|}(k) \tag{4.13}$$

It means, only the parallel component $\mathbf{d}^{\|}(k)$ contributes to a reduction of the length of the vector $\mathbf{d}(k+1)$. As can be observed from figure 4.2, this reduction is met only for:

$$0 < \mu < 2 \tag{4.14}$$

which represents the stability criterion of the NLMS algorithm. Also, the mean square deviation is defined as:

$$\mathcal{D}(k) = E\left[\|\mathbf{d}(k)\|^2\right] \tag{4.15}$$

In equation 4.15, E denotes the expected value, which is considered here the same as the mean value. If we take the mean of the square Euclidean norms of both sides of equation 4.10, we will have:

$$\mathcal{D}(k+1) - \mathcal{D}(k) = \mu^2 E\left[\frac{|e(k)|^2}{\|\mathbf{p}(k)\|^2}\right] - 2\mu E\left[\frac{e(k)\mathbf{d}(k)\mathbf{p}(k)}{\|\mathbf{p}(k)\|^2}\right] \tag{4.16}$$

With the assumption, that the input signal energy $\|\mathbf{p}(k)\|^2$, from one iteration to the next, can be approximated by a constant [2] and considering equation 4.9, we will have [3] :

$$\mathcal{D}(k+1) - \mathcal{D}(k) = \mu^2 \frac{E\left[|e(k)|^2\right]}{E\left[\|\mathbf{p}(k)\|^2\right]} - 2\mu \frac{E\left[e(k)e(k)\right]}{E\left[\|\mathbf{p}(k)\|^2\right]} \tag{4.17}$$

According to equation 4.17, the mean square deviation decreases with increasing iteration k and the NLMS filter is stable in the mean square error sense if the condition in equation 4.14 is met [Hay 02]. For the general case, in the presence of environmental noise $n(k)$, equation 4.17 changes to:

$$\mathcal{D}(k+1) - \mathcal{D}(k) = \mu^2 \frac{E\left[|e(k)|^2\right]}{E\left[\|\mathbf{p}(k)\|^2\right]} - 2\mu \frac{E\left[e(k)\varepsilon(k)\right]}{E\left[\|\mathbf{p}(k)\|^2\right]} \tag{4.18}$$

Since for the LMS and NLMS algorithms the exact optimal answer is never reached, the filter converges to an optimal answer about which the estimated impulse response $\mathbf{h}$ changes. The amount of this change depends on the step size value. In order to find the optimal step size for the NLMS method and by derivation of equations 4.17 and 4.18, it is shown that the largest decrease in the distance vector can be achieved for [Mad 00]:

$$\mu_{\text{opt}}(k) = \frac{E\left[e(k)\varepsilon(k)\right]}{E\left[|e(k)|^2\right]} \tag{4.19}$$

or:

$$\mu_{\text{opt}} = 1 \quad \text{for } n(k) \equiv 0 \tag{4.20}$$

Whereas in order to guarantee for the convergence of the LMS algorithm, the step size should

[2] so that the approximation $E\left[\frac{|e(k)|^2}{\|\mathbf{p}(k)\|^2}\right] \approx \frac{E\left[|e(k)|^2\right]}{E\left[\|\mathbf{p}(k)\|^2\right]}$is justified. [Hay 02]

[3] Note that for $n(k) \equiv 0$ we have: $e(k) = \varepsilon(k)$

meet the following condition [Hay 02, Wid 85]:

$$0 < \mu < \frac{2}{\lambda_{\max}} \tag{4.21}$$

where $\lambda_{\max}$ is the largest eigenvalue of the input correlation matrix $\mathbf{R}^4$. It means that in the LMS algorithm the step size should be inversely proportional to the signal power. Since the optimal convergence speed of the LMS algorithm is achieved for small eigenvalue spread, $\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)$, with increasing eigenvalue spread the convergence of the LMS algorithm will not be optimal anymore [Hay 02]. If we want to increase the convergence speed by increasing the step size, there might be the risk of disturbing the condition in equation 4.21. However, for the NLMS algorithm and for stationary signals, analytical expressions for input eigenvalues are not needed. NLMS can be seen as a special case of the LMS, for which the step size is re-parameterized as (see equation 4.6):

$$\tilde{\mu} = \frac{\mu}{\|\mathbf{p}(k)\|^2} \tag{4.22}$$

and therefore, the step size condition of equation 4.14 is independent of the signal characteristics [Slo 93]. [Tar 88] also showed the better convergence behavior of the NLMS over the LMS for big step size values, however the NLMS shows worse mean square deviations. As a rule of thumb, the best step size for the LMS and NLMS algorithms with respect to the convergence speed, is the one which is half of the maximum stable value. But in practice, usually smaller step size values are chosen to guarantee the stability in the presence of environmental transient noise and disturbances [Ben 01].

4.3 Perfect sweep as optimal excitation of the NLMS

According to figure 4.2 and equation 4.13, for an impulse response $\mathbf{g}$ of length N all N components of the distant vector, $\mathbf{d}(k)$, can be eliminated if N consecutive inputs $\mathbf{p}(k), \mathbf{p}(k-1), ..., \mathbf{p}(k-N+1)$ are orthogonal in the N-dimensional vector space [Ant 94, Ant 08]. Assuming that the input vector $\mathbf{p}(k)$ is a white noise so that the consecutive vectors $\mathbf{p}(k), \mathbf{p}(k-1), ..., \mathbf{p}(k-N+1)$ are each of infinite length, these vectors are orthogonal in the infinite vector space. However, for real applications, the vectors are of finite length and the orthogonality in the N-dimensional vector space is not guaranteed. Since the weight vector $\mathbf{h}(k)$, which changes at each iteration, depends on the past input vectors $\mathbf{p}(k), \mathbf{p}(k-1), ..., \mathbf{p}(k-N+1)$, if these successive input vectors are independent over time, $\mathbf{h}(k)$ will be independent of $\mathbf{p}(k)$ and the adaptation process will converge after sufficient number of iterations to the optimal

[4] $\mathbf{R} = E\left[\mathbf{p}(k)\mathbf{p}(k)^T\right]$

solution [Wid 85]. In applications of NLMS such as echo cancelation, in which the adaptation process is driven by speech signals, the convergence speed is drastically reduced due to colored signals. White noise of finite length as excitation signal provides a better but still not optimal convergence speed. The key to the optimal adaptation is that the N consecutive vectors of the excitation signal are orthogonal to each other. If this condition is met, the NLMS algorithm is able to identify the unknown impulse response of the length N (in a noiseless environment and for $\mu = 1$) after N iterations.

An alternative class of excitation signals beside the white noise stimulus is the so called Perfect Sequences (PSEQs), which offer interesting properties regarding convergence speed of the NLMS adaptive filtering. Perfect Sequences are periodically repeated pseudo noise signals. The main and most important characteristic of PSEQs is that their periodic autocorrelation function $R_{\mathrm{pp}}(i)$ vanishes for all out-of-phase values, it means that all N-phase shifted sequences of the signal with period N are orthogonal [Luk 88]. For the PSEQ $\mathbf{p}(k)$ of period N we have:

$$R_{\mathrm{pp}} = \sum_{k=0}^{N-1} p(\mathrm{i})p(k+\mathrm{i}) = \begin{cases} \|p(\mathrm{i})\|^2 & \text{for i mod N} = 0 \\ 0 & \text{else} \end{cases} \tag{4.23}$$

For this reason, perfect sequences fulfill the requirement of an optimal excitation for the NLMS algorithm [Ant 95]. [Ant 08] showed the advantage of exciting the system with PSEQs in comparison to excitation with white noise. Using the logarithmic distance vector [5] as a measure, figure 4.3 shows the LTI system response to a sudden change at iteration $k = 3000$.

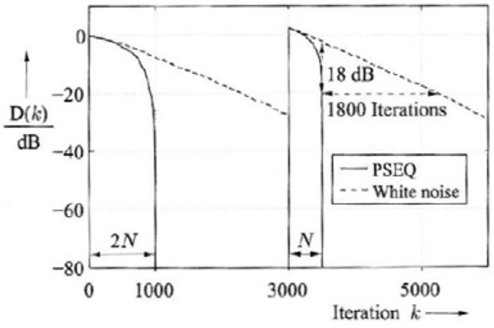


Figure 4.3: Logarithmic distance vector for an LTI system with PSEQ and noise excitation with a sudden change at $k = 3000$ for $\mu = 1$, $N = 500$, and $n(k) \equiv 0$. [Ant 08]

As the environmental noise is eliminated and with the assumption that the length of the adaptive filter is the same as the period N of the PSEQ, the NLMS algorithm is able to reduce the system distance to the desired value within N iterations, whereas with white noise excitation,

[5] The logarithmic distance vector is defined as: $10log_{10}\dfrac{\|\mathbf{g}-\mathbf{h}(k)\|^2}{\|\mathbf{g}\|^2}$ [Ant 08]

is takes much longer. Notice that the initialization phase takes $2N$ iterations. The extra N-samples delay is caused by the N empty filter states of the unknown system at the beginning. Since for most applications the excitation should have a high energy efficiency (or a small crest factor), binary sequences are most preferable, also for ease of implementation. However, no perfect binary sequence of length $N > 4$ can exist [Luk 88]. Therefore, ternary sequences, whose symbols are from the set $[-1, 0, 1]$ are more promising, such as Ipatov sequences [Ipa 79], or odd perfect sequences [Luk 95], both of which can however be constructed only for limited possible lengths. In addition, pseudo noise sequences offer only theoretically the ideal excitation signal in sense of energy efficiency, as there are limitations using them, as already discussed in chapter 3.1. As a result, sweeps perform as a more preferable choice in acoustical measurement tasks. Despite a non-zero crest factor of 3 dB, sweeps can be used at higher amplitudes than pseudo noise sequences through a distortion free measurement [Mul 08] and can also be designed, in contrast to pseudo noise sequences, for any lengths. The problem of using sweeps as excitation signal for the NLMS adaptive filtering system identification is that sweeps do not show the perfect impulse-like autocorrelation function as a perfect sequence. Telle et al. [Tel 10] introduced a new class of PSEQs, the so called perfect sweep, which combines the characteristics of a sweep signal and a PSEQ. A perfect sweep can be constructed as a linear sweep in time and/or in frequency domain. For the construction of a linear sweep in the time domain, a sinus signal is calculated with a phase which increases with a fixed rate per time unit [Mul 01]. As shown in figure 4.4, if the linear sweep is repeated periodically, it is not a perfect sweep because the spectrum is not white at frequencies near beginning and near the nyquist frequency.

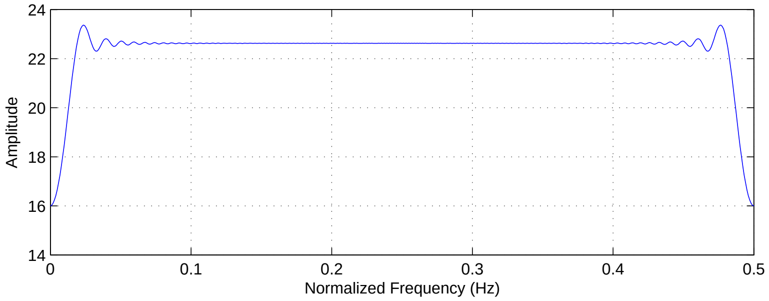


Figure 4.4: Spectrum of the linear sweep, constructed in the time domain.

In order to achieve a completely white spectrum the perfect sweep should be constructed in frequency domain by setting the spectrum amplitude to a constant value and designing a linearly increasing group delay, as already discussed in chapter 3.1. The perfect sweep is then directly calculated by taking the inverse Fourier transform of this signal. For periodically repetitions of the signal, by taking the inverse Fourier transform of the signal, high frequencies

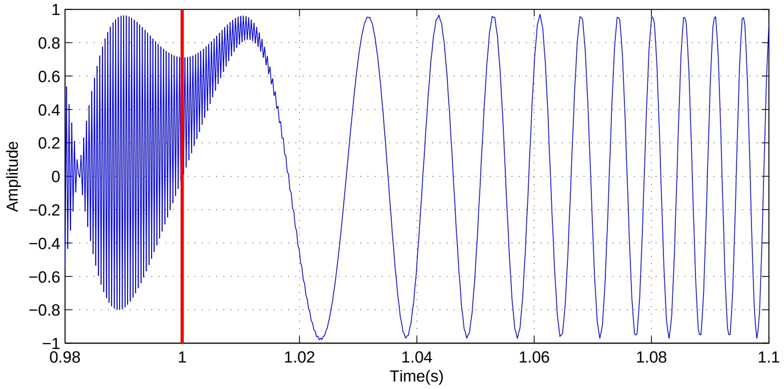


Figure 4.5: The effect of circular convolution for a sweep generated in frequency domain after taking the inverse Fourier transform, passing from one period to the next. The period of the sweep is 1 second.

fold back on the low frequencies, influencing the beginning part of the next period, as shown in figure 4.5. According to [Tel 10], extra actions such as zero padding should nevertheless be avoided, since using such corrections, the sweep is not perfect anymore. Avoiding corrections will guarantee a continuous transition from one period to the next.

4.4 The choice sequence period

Up to this point the length N_{g} of the unknown impulse response $\mathbf{g}$, the length N_{h} of the filter $\mathbf{h}$ and the length N_{p} of the excitation signal $\mathbf{p}$ (in the case of PSEQs, the period of the sequence) were considered to be the same. Actually, $N_{\mathrm{g}} = N_{\mathrm{h}} = N_{\mathrm{p}}$ should be met in order for the NLMS to behave optimally for $\mu = 1$ and $n(k) \equiv 0$ [Ant 95]. According to the nature of the NLMS algorithm, the period of the PSEQ, N_{p}, has to match the length of the filter N_{h}, it means ,$N_{\mathrm{p}} = N_{\mathrm{h}}$. If the period of the PSEQ is smaller ($N_{\mathrm{h}} > N_{\mathrm{p}}$), not all directions of the distance vector $\mathbf{d}(k)$ can be excited. And if the period of the PSEQ is larger ($N_{\mathrm{h}} < N_{\mathrm{p}}$), it is shown that the convergence speed of the system will degrade [Ant 08, Ant 95]. Since PSEQs can be constructed for different lengths, the condition $N_{\mathrm{p}} = N_{\mathrm{h}} = N$ does not represent any constraint. But it is often a problem to meet for the other condition, $N_{\mathrm{g}} = N$, because, the length of the impulse response of the system may be unknown. Assuming an impulse response $\mathbf{g}$ of length N_{g} as $\mathbf{g} = \left(g_0, g_1, ..., g_{\mathrm{N_g\text{-}1}}\right)$, the identification of the filter coefficients for $\mu = 1$ and $n(k) \equiv 0$ is [Ant 08]:

$$h_{\mathrm{k \ mod \ N}}(k) = \delta\left(\mathrm{k \ mod \ N}\right) * \hat{g_{\mathrm{k}}} \qquad \mathit{with} \quad \hat{g_{\mathrm{k}}} = \begin{cases} g_{\mathrm{k}} & k < N_{\mathrm{g}} \\ 0 & k \geq N_{\mathrm{g}} \end{cases} \tag{4.24}$$

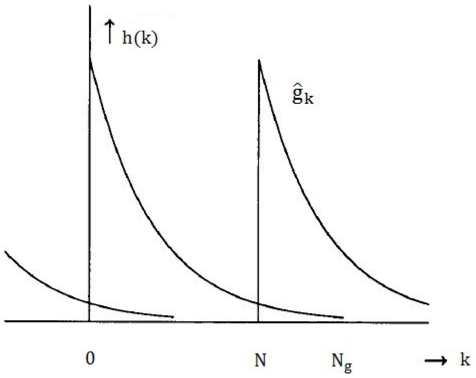


Figure 4.6: Time aliasing effect in the NLMS system identifiation due to filter length shorter than the length of the system impulse response [Ant 94] (Parameter names in figure were changed for uniformity).

δ denotes the unit impulse and $*$ is the convolution operator. The interpretation of equation 4.24 is shown in figure 4.6 for $N_{\mathrm{g}} > N$. As a result, the impulse response wraps back on itself and causes a time aliasing error [Ant 94]. So, the length of the filter, N, must be long enough. On the other hand it should be noted that large filter lengths also increase the convergence time. Therefore the optimal choice for the filter length and the PSEQ period is given by $N_{\mathrm{g}} = N_{\mathrm{h}} = N_{\mathrm{p}}$.

4.5 Single channel acquisition of HRTFs using the NLMS adaptive filter

Enzner [Enz 08] introduced a method for continuous-azimuth acquisition of HRTFs using NLMS adaptive filtering system identification. As depicted in figure 4.7, the subject of interest (artificial head or real person) rotates continuously in the horizontal plane as the loudspeaker at a fixed elevation φ plays back the excitation signal and the microphones at the entrance of the subject's blocked ear canal capture the signals. The acoustical transfer path between the loudspeaker and the microphones is considered as a linear time varying system with the assumption of small non-linear behavior of the measurement setup. Under this assumption, for this acoustical path a time varying impulse response for the left and right ear, h^{i}_{φ} ($\mathrm{i} = l$ or r), can be considered so that the signals captured by the two microphones, $y^{\mathrm{i}}(k)$ at each time k (or its corresponding azimuth θ_{k}) can be described as:

$$y^{\mathrm{i}}(k) = \sum_{m=0}^{N} h^{\mathrm{i}}_{\varphi}(m, \theta_k)\mathbf{X}(k-m) + n^{\mathrm{i}}(k) \tag{4.25}$$

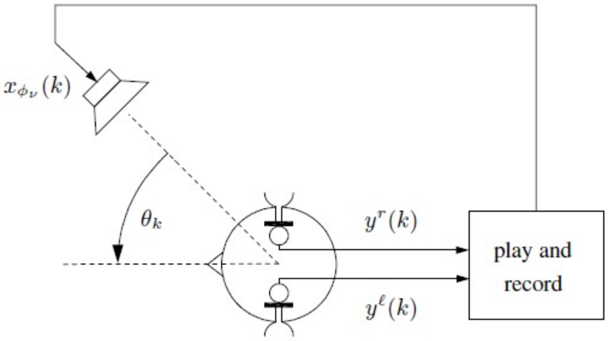


Figure 4.7: Measurement setup for single channel acquisition of HRTFs using NLMS adaptive filtering. [Enz 08]

X denotes the most recent N samples of the known excitation signal x:

$$\mathbf{X} = (x(k), x(k-1), ..., x(k-N+1))^T, \tag{4.26}$$

and $h^{\mathrm{i}}_{\varphi}(\theta_k)$ denotes the impulse response of the acoustical path or simply the HRIR at elevation φ and azimuth θ_k with a length of N samples:

$$h^{\mathrm{i}}_{\varphi}(\theta_k) = \left(h^{\mathrm{i}}_{\varphi}(0, \theta_k), ..., h^{\mathrm{i}}_{\varphi}(N-1, \theta_k)\right) \tag{4.27}$$

$n^{\mathrm{i}}(k)$ models the independent noise, which is present at the left and right microphones including the environmental noise and the noise caused by microphones, amplifiers and other measurement equipment as well as the mechanical system used for the continuous rotation. Subject's constant angular speed is $\omega_0 = \frac{2\pi}{T_{360}}$ with T_{360} as the duration of one complete 360° rotation and θ_k corresponds to the quasi-continuous azimuth $\theta_k = \omega_0 k T_s$ with the temporal sampling interval $T_s = \frac{1}{f_s}$ and the sampling frequency f_s. Making the minimum mean square error $(e^{\mathrm{i}}(k))$ between the real microphone signals and the estimated one at each time k, similar to the NLMS recursive equations 4.6 and 4.7, the new update of the estimated HRIR, $\hat{h_{\varphi}}^{\mathrm{i}}$, at iteration $k+1$ will be:

$$\hat{h_{\varphi}}^{\mathrm{i}}(\theta_{k+1}) = \hat{h_{\varphi}}^{\mathrm{i}}(\theta_k) + \mu \frac{e^{\mathrm{i}}(k)\mathbf{X}^T(k)}{\|\mathbf{X}(k)\|^2} \tag{4.28}$$

with the step size μ and the error signal:

$$e^{\mathrm{i}}(k) = y^{\mathrm{i}}(k) - \hat{h_{\varphi}}^{\mathrm{i}}(\theta_k)\mathbf{X}(k) \tag{4.29}$$

The unknown states of $\hat{h_\varphi}^{\mathrm{i}}$ at the beginning ($k = 0$) is replaced with zeros:

$$\hat{h_\varphi}^{\mathrm{i}}(0) = \underbrace{(0, 0, ..., 0)}_{\mathrm{N}} \tag{4.30}$$

The excitation signal **X** is measured via reference measurements with the microphones at the center position of subject's head in his absence [Wnrt 12].

Assuming a 360° revolution time of 20 seconds for this single channel case [Enz 08] and with a sampling frequency of $f_s = 44100$ Hz, it can easily be seen that the HRIRs can be calculated with an azimuth resolution of $\Delta\theta \approx 10^{-4}$ degrees, showing that the azimuths can be considered with a good approximation as continuous. According to this fact there will be a huge amount of data if all HRIRs at all iterations are assumed to be acquired. Actually, since the HRIRs can be calculated off-line after the microphones have captured the signals for one complete rotation, one can sample out and store the HRIRs only at desired azimuths. It should also be noted that the linear convolution model described by equation 4.25 is valid under the assumption that the system changes more slowly than the time, which is needed to change all states of the NLMS filter, in other words, the time-constant of impulse response variations is significantly larger than the memory of the filter, N.

4.6 Multi-channel (3D) acquisition of HRTFs using the NLMS adaptive filter

Enzner [Enz 09] extended the idea of HRTF acquisition using NLMS adaptive filters to the multi-channel case, as shown in figure 4.8. This system consists of an array of M loudspeakers which are fixed on a vertical arc at discrete elevations. The subject of interest is positioned with his head in the center of the arc and is rotated horizontally with in-ear microphones, as the loudspeakers play back simultaneously the excitation signal. The recursive update process of equations 4.28 and 4.29 changes for the case of multi-channel system identification to:

$$\hat{h_{\varphi_\nu}}^{\mathrm{i}}(\theta_{k+1}) = \hat{h_{\varphi_\nu}}^{\mathrm{i}}(\theta_k) + \mu \frac{e^{\mathrm{i}}(k)\mathbf{X}_{\varphi_\nu}{}^T(k)}{\sum_{\varphi_\nu} \|\mathbf{X}_{\varphi_\nu}(k)\|^2} \tag{4.31}$$

$$e^{\mathrm{i}}(k) = y^{\mathrm{i}}(k) - \sum_{\varphi_\nu} \hat{h_{\varphi_\nu}}^{\mathrm{i}}(\theta_k)\mathbf{X}_{\varphi_\nu}(k) \tag{4.32}$$

It should be noted, that the normalization is done with a common summing term, $\sum_{\varphi_\nu} \|\mathbf{X}_{\varphi_\nu}(k)\|^2$, and the same error signal is used to update the state of impulse response for each channel.

A typical problem of multi-channel system identification using NLMS adaptive filters is that the cross correlation between excitation signals might be non-zero. As only one single output

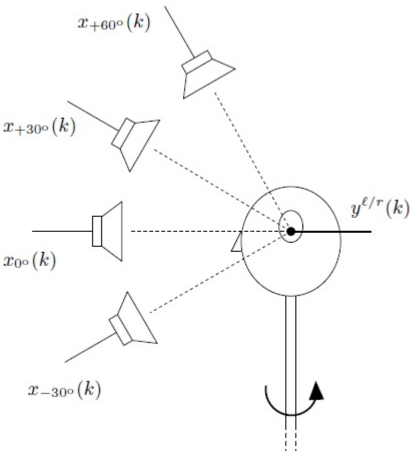


Figure 4.8: Measurement setup for multi-channel acquisition of HRTFs using NLMS adaptive filters. [Enz 09]

signal, y^{i}, is present for the simultaneous identification of all channels, the choice of input excitation signal, $\mathbf{X}_{\varphi_\nu}$, should be met adequately. For such a case, Antweiler [Ant 08] proposed the use of a PSEQ, $\mathbf{X}_{MN}$, with an extended period length of $M \cdot N$. M is the number of systems to be identified and N is the length of the HRIR filters. For the first channel, the originally constructed PSEQ is used as excitation signal, for other channels, a shifted version of this PSEQ is used in a manner that each channel is in N samples shifted against the previous and next channel, so that:

$$\begin{aligned} \mathbf{X}_{\varphi_1}(k) &= \mathbf{X}_{MN}(k) \\ \mathbf{X}_{\varphi_2}(k) &= \mathbf{X}_{MN}(k-N) \\ &\vdots \\ \mathbf{X}_{\varphi_M}(k) &= \mathbf{X}_{MN}(k-(M-1)N) \end{aligned} \tag{4.33}$$

Although the excitation signal has now an extended period duration of $M \cdot N$, since the length of HRIRs for each channel remains the same as in the single channel case, the input vectors used in the recursive equations 4.31 and 4.32 are all of dimension N, meaning that the used excitation signal of each channel in these equations contains only $\frac{1}{M}$ of the whole PSEQ. In this case the excitation signal of each channel has an ideal impulse like autocorrelation function and the excitation signals of all channels are orthogonal with a zero cross correlation which will lead to the optimal excitation for the multi-channel NLMS system identification. Excitation with this kind of extended PSEQ shows a higher convergence speed in comparison to the excitation with white noise, but in contrast to single channel excitation, where the unknown

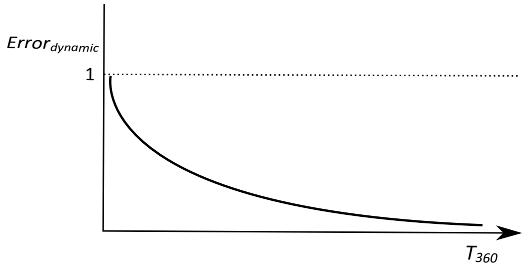


Figure 4.9: Qualitative behavior of dynamic system inaccuracy with revolution time, according to the model of first order Markov process proposed by [Enz 08].

impulse response of length N can be identified after N iterations in a noiseless environment, it will take $M \cdot N$ iterations for the M-channel case to identify M impulse responses, each of length N. Also, in order to balance the time constant of the NLMS filter and the variability of the system between the two cases of single- and multi-channel excitation, the revolution time should be extended from T_{360} in the single-channel case to $M \cdot T_{360}$ for the multi-channel excitation case with M channels [Enz 13].

4.7 Accuracy of dynamic HRIR acquisition - a measurable quantity

According to Haykin [Hay 02], after enough number of iterations, the mean square deviation $\mathcal{D}$ can be approximated as [6]:

$$\mathcal{D}(k) \simeq \mu \text{Error}_{\text{noise}} + \frac{1}{\mu} \text{Error}_{\text{dynamic}} \tag{4.34}$$

According to equation 4.34, there are two error sources which contribute to the inaccuracy of the system identification: $\text{Error}_{\text{noise}}$, which is caused by observation noise and contributes linearly with step size to inaccuracy, and $\text{Error}_{\text{dynamic}}$, which is due to variability of the dynamic rotation of the system and affects the inaccuracy inversely with step size. Enzner [Enz 08] proposed a model based on first order Markov process to describe the dynamic behavior of the rotating system. According to this model, $\text{Error}_{\text{dynamic}}$ decreases exponentially as a function of rotation time in the form of $1 - e^{-\frac{const.}{T_{360}}}$, as qualitatively depicted in figure 4.9.

This behavior implies that longer rotation times reduce the inaccuracy, however the improvement of accuracy with increasing rotation time gets slower with longer T_{360}. In a noiseless situation, the first term on the right-hand side of the equation 4.34 disappears and the optimal step size value with respect to accuracy will be the maximum possible value within the stability

[6] To recall: $\mathcal{D}(k) = E\left[\|\mathbf{d}(k)\|^2\right] = E\left[\|\mathbf{g} - \mathbf{h}(k)\|^2\right]$

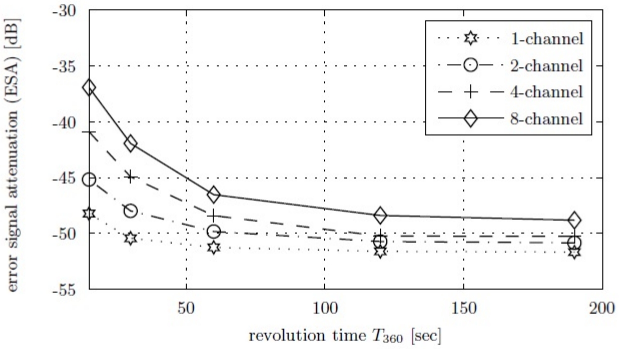


Figure 4.10: ESA for dynamical HRTF measurement for different channel numbers and rotation times. [Enz 13]

range, it means μ=1, as also discussed in section 4.2. In the presence of environmental noise, the optimal step size is theoretically the value for which the two terms on the right-hand side of equation 4.34 contribute equally to the deviation $\mathcal{D}$. However, practically, the deviation $\mathcal{D}$ cannot be accessed directly for noisy environments. Since the only available signals are the error signal[7] and the captured signals y^{i}, Enzner introduced the Error Signal Attenuation (ESA), which maps the theoretical deviation onto a measurable inaccuracy [Enz 08], and can be achieved from the variance of the two signals $e(k)$ and y as [Enz 13]:

$$ESA = 10log_{10}\frac{\sigma^2{}_e}{\sigma^2{}_y}/N(dB) \tag{4.35}$$

As can be seen in figure 4.10, ESA decreases with increasing rotation time and for the multi-channel case, the ESA increases in comparison to single-channel case especially for short rotation times. The results of figure 4.10 are from a dynamical HRTF measurement with white noise excitation, done by Enzner et al. [Enz 13]. According to the authors, there is also another source of error, arising from the finite length of the adaptive filter, which has led to the stopping improvement of the accuracy for longer rotation times. The truncation of the last samples of HRIRs leads to this error and causes results with under modeled low frequency. According to Antweiler et al. [Ant 09], this error is in case of white noise excitation a non-systematic error, which appears as an audible noisy like disturbance signal. Therefore, smaller step size values should be used to reject this error, whereas in case of excitation with PSEQs, this error is systematic with less disturbing fluctuations from one iteration to the next. In comparison to the white noise excitation, PSEQs can result in better ESAs despite truncated filter lengths.

[7] To recall: $e(k) = (\mathbf{g} - \mathbf{h}(k))^T\, \mathbf{p}(k) + n(k)$ - for the case of HRIR measurement: $e^{\mathrm{i}}(k) = y^{\mathrm{i}}(k) - \hat{h_{\varphi}}^{\mathrm{i}}(\theta_k)\mathbf{X}(k)$

4.8 Chapter summary

In this chapter, the NLMS as a widely used adaptive filtering algorithm was introduced and discussed with respect to its adaptation mechanism and stability. In addition , the best choices for the filter length and excitation signal to achieve the best performance of the algorithm were mentioned. It has been shown, that exciting with perfect sequences, due to their impulse-like autocorrelation function, leads to the optimal performance in the sense of adaptation speed. Considering the attractive properties of sweeps for acoustical measurement tasks, perfect sweeps, as proposed by [Tel 10], can be used to gather the advantages of perfect sequences and sweep excitations. Further on, the method proposed by Enzner [Enz 08] was introduced. This method uses NLMS system identification to acquire HRTFs after a measurement with a continuously rotating subject and enables to cover all azimuths (with a very high resolution) within one complete rotation of 20 seconds duration for one sound source at a fixed elevation. This method can also be extended to a multi-channel case [Enz 09], however, modifications should be considered on the excitation signal to attain the best performance of the adaptation. In addition, the multi-channel HRTF acquisition necessitates longer rotation times. One important parameter involved in the NLMS adaptation process is the step size, which was also discussed in this chapter. On the one hand, choosing the correct step size can guarantee the stability of the adaptive filter, on the other hand, the accuracy of the adaptation system has been shown to vary with environmental noise and the dynamic behavior of the time varying system (rotating subject during the measurements) and the step size has an influence on this variation. Generally, large step size values work in favor of systems with short measurement times (corresponding to fast rotations for the continuous HRTF acquisition system) to allow the quick tracking of the time varying system. In contrast, in the presence of noise, a smaller step size is preferable due to the ability of noise rejection. Furthermore, the length of the NLMS filter has an influence on the adaptation speed and the accuracy. The best performance of the system mentioned above can only be achieved considering all involved parameters and finding a tradeoff between them.

Chapter 5

Modeling HRTF measurements

The acoustical path between the loudspeakers and the microphones during HRTF acquisition can be modeled as illustrated in figure 5.1.

x depicts the excitation signal (exponential or perfect sweep). $H_{loudspeaker}$ represents the transfer function of the loudspeaker and the effect of harmonic distortions. The former performs generally as a 2^{nd} order highpass filter and the latter can be modeled using Volterra series. The details are presented in sections 5.1 and 5.2. H_{air} includes the effects of differences in the sound speed and the attenuation by the air. Viscosity, heat conduction and thermal relaxation are the reasons of energy attenuation during sound propagation through the air. The amount of attenuation is also dependent on the distance which the sound travels and the humidity has also a significant influence, which can cause up to 0.1dB/m attenuation of sound level at high frequencies [Vor 08]. Since the attenuation differs for different frequencies, this can lead to slight spectral colorations. Another effect is the change in the sound speed due to temperature fluctuations. However, the distance between the loudspeakers and the microphones

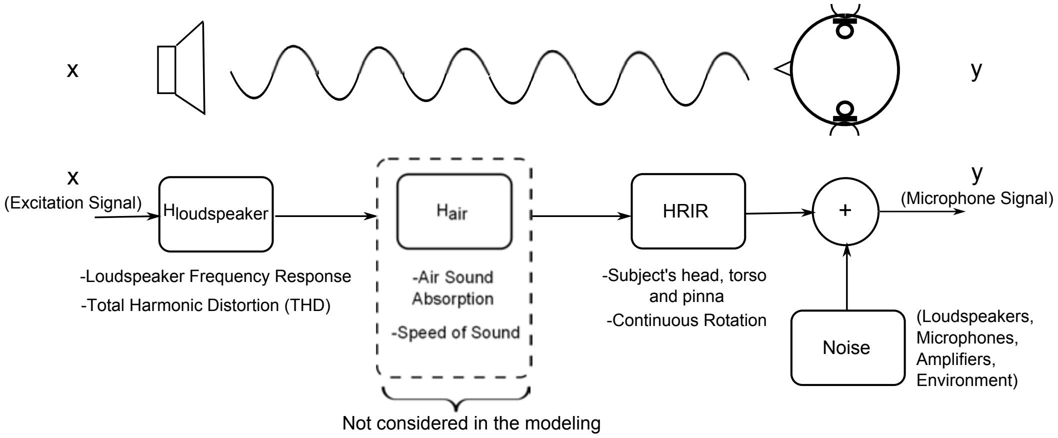

x
y
x
(Excitation Signal)
Hloudspeaker
Hair
HRIR
+
y
(Microphone Signal)
-Loudspeaker Frequency Response
-Total Harmonic Distortion (THD)
-Air Sound Absorption
-Speed of Sound
-Subject's head, torso and pinna
-Continuous Rotation
Noise
(Loudspeakers, Microphones, Amplifiers, Environment)
Not considered in the modeling

Figure 5.1: Transfer path model between the loudspeakers and the microphones.

in the simulations is considered to be in the range of 1 to 2 meters and due to the reduced measurement time (as it is the main goal of the implemented methods), the effect of sound air attenuation and sound speed variation can be neglected. The microphones at the blocked ear canal capture the direct sound from the loudspeakers as well as the reflections from the subject's head, pinna, shoulders and torso. To model the subject himself, an existing HRTF data set is used, which will be presented in more details in section 5.3. Furthermore, subject's position against the sound sources varies permanently in form of a continuous horizontal rotation, making the system get time-variant. Section 5.4 discusses the modeling of the time varying rotating system. Inevitably, the measurement is affected by unwanted interferences of the noise in the room which is caused by the environment, loudspeakers and other measurement equipment. This noise is assumed as a time-invariant signal and is modeled as described in section 5.5.

5.1 A model for linear and high order nonlinear loudspeaker transfer functions

Loudspeakers are the most important source of nonlinear distortions in an acoustic measurement- and transmission line [Goe 08]. The dominant elements in a loudspeaker which cause nonlinearities are mostly related to the voice coil excursion and the amplitude of the input signal such as stiffness of the suspension or the force factor (electromagnetic driving component) as functions of coil displacement, or electromagnetic driving forces as a function of coil displacement and electrical current [Kli 92]. Such nonlinearities generate signal components which do not exist in the exciting input signal and appear generally in the spectrum as integer multiples of the applied fundamental frequencies (harmonic distortion) or as linear combination of the applied fundamentals (intermodulation distortions). One established figure of merit for the level of the harmonics is the Total Harmonic Distortion (THD), which is defined, among other existing definitions, as the ratio of the sum of the power of all existing harmonic components to the power of the fundamental and can be given in decibel or percents as [Mul 08]:

$$THD = 100\% \frac{\sqrt{\sum_{k=2}^{N} I_k}}{I_1} = 20dB \log \frac{\sqrt{\sum_{k=2}^{N} I_k}}{I_1} \tag{5.1}$$

Another common definition for THD compares the harmonic content to the total rms value of the signal. In case of power systems, equation 5.1 offers the better definition [Shm 05]. However, for small THD values these two definitions do not show considerable differences.

The loudspeaker can be seen as a nonlinear system which has, besides the linear impulse response, a set of higher order harmonic impulse responses, also known as kernels. The transfer function of an electro-dynamic loudspeaker, which defines the relationship between the output sound pressure, p, and the input electric voltage, u, is given by [1]:

$$\frac{p}{u} = \underbrace{Bl\frac{1}{R_i + R_a + j\omega L}}_{\text{electrical behavior}} \cdot \underbrace{\frac{1}{j\omega m + r + \frac{D}{j\omega}}}_{\text{mechanical behavior}} \cdot \underbrace{\frac{j\rho_0\omega\pi a^2}{4\pi d}}_{\text{acoustical behavior}} \qquad (5.2)$$

According to equation 5.2, the behavior of an electro-dynamic loudspeaker over frequency can be divided into three parts. The mechanical part represents a resonance frequency at $\omega_0 = \sqrt{\frac{D}{m}}$ with 6dB and -6dB per octave roll-off before and after this resonance frequency. The electrical part behaves as a first order low pass filter (-6dB per octave roll-off) with a cut-off frequency at $\omega_k = \frac{R_i+R_a}{L}$. And finally the acoustical part of the transfer function rises continuously with 6dB per octave over frequency. These altogether result in a frequency response with 12dB/octave increase up to resonance frequency and 6dB/octave fall after the electrical cut-off frequency [Mos 09].

The increase of 12dB/octave was modeled in this thesis with a second order Butterworth high pass filter and the electrical cut-off frequency was assumed to be outside the audible frequency range. In order to have a choice for the cut-off frequency of this Butterwort filter, the measurement results of five loudspeaker drivers were used. These loudspeaker drivers, which were measured as a part of another Master thesis, are listed in table 5.1

Loudspeaker driver model	**Abbreviation in this thesis**
Fountek FR89EX	FOUN
Monacor SPH_30X/f	MON
Peerless 830984	P8
Peerless NE65W-04	PNE
Tangband W3-881SJF	TB_W3

Table 5.1: Loudspeaker drivers used to choose a proper loudspeaker model.

Figure 5.2 shows the result of sensitivity measurements together with the synthesized 2$^{\text{nd}}$ order high pass filter. The cut-off frequency of the high pass filter was varied until the 12dB/octave roll-off of the filter matched approximately the behavior of all five drivers. The curves corresponding to different measurements were shifted vertically against the high pass filter to a proper position to make the comparison easier, therefore, the vertical axis does not supply any special information and figure 5.2 shows rather a comparison between the roll-off behavior of the drivers and that of the high pass filter. According to this comparison, a cut-off frequency of 180Hz was chosen for the model of the linear frequency response of the loudspeaker.

In order to find a proper general model for the loudspeaker's nonlinear behavior, the results of THD measurements for loudspeaker drivers of table 5.1 were studied. The measurement

[1] The derivation of Eq. 5.2 as well as description of its parameters are presented in appendix B

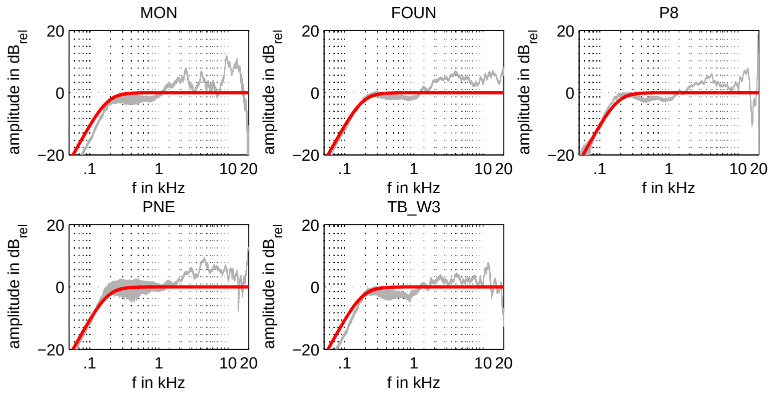


Figure 5.2: Sensitivity measurement results of the five loudspeaker drivers of table 5.1. The bold curve shows the synthesized high pass Butterworth filter.

results, which were available to be used for the modeling in this thesis, included the THD measurements for each driver at three different excitation powers and were carried out for harmonic components up to the fifteenth order. For each harmonic order, the results of the five drivers, each excited at three different power levels, were plotted. Again, the goal was to find a model which matches in average possibly all THD curves for all harmonic orders. This model was achieved by applying a low shelf filter to the before mentioned 2$^{\mathrm{nd}}$ order high pass butterworth which was chosen as the linear impulse response. The result was in addition low pass filtered according to the order of the harmonic response, to avoid the Aliasing effects which could occur after the high order response is raised to the power of the harmonic order (see the Hammerstein model of nonlinearity and equation 5.5). The shelf frequency and the amount of the increase below the shelf frequency were the two degrees of freedom which were to be changed until a proper match to the measurement curves was achieved. The measurement curves were again shifted against the model to make a comparison possible. Each curve was shifted vertically until it contained the same energy as the model. The results of the THD measurements for orders k=2 to k=7, together with the chosen model, are shown in figure 5.3. As a result of this comparison, the shelf frequency was chosen at 1 kHz with an increase of 50dB for frequencies below it.

Having chosen the parameters for the low shelf filter, the next step was to weight the harmonic responses properly, similar to the behavior of the five loudspeaker drivers. To this aim, the energy content of the harmonic responses of orders k=2 to k=15 for all drivers and for all excitation powers were calculated, which are shown in figue 5.4. Then, a polynomial of second order was fitted to the results. The curve fitting was also repeated for the case of omitting the outliers out of the results (this was the case for the two models Peerless NE65W-04 and Tangband W3-881SJF for high and middle levels of excitation). However, since the maximum

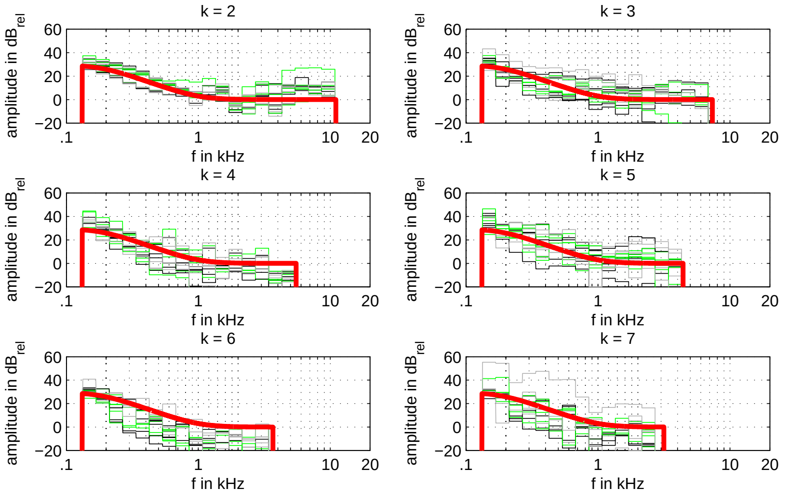


Figure 5.3: The results of the THD measurements for loudspeaker drivers listed in table 5.1 for harmonic orders k=2 to k=7. The bold curve shows the model which was chosen as harmonic loudspeaker responses.

difference to the case, where all results were considered, was about 4dB at relatively higher harmonic orders (k=8 and above), the curve, which was fitted to all measurement results, was considered as acceptable. Using these models and assuming that the number of high order responses, K_{max}, is limited, the high order responses were weighted according to the results shown in figure 5.4. Next, a uniform amplification was applied to them so that the sum of their energy in relation to the energy of the linear response led to a given THD value according to equation 5.1. Figure 5.5 shows the spectrum of the linear and high order impulse responses for the modeled loudspeaker with K_{max}=5 and 3% Total Harmonic Distortion.

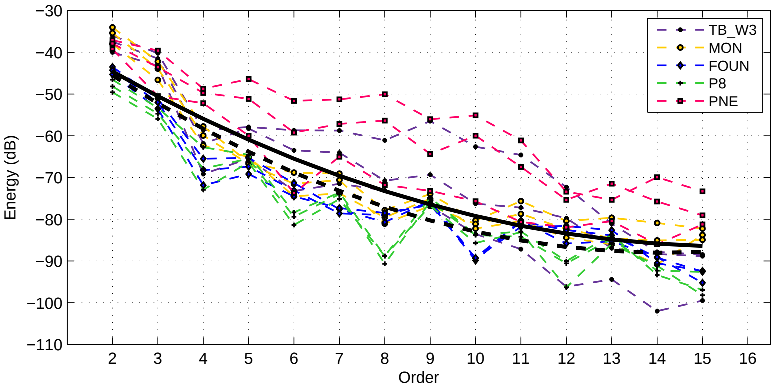


Figure 5.4: Modeling the energy attenuation of harmonic impulse responses. The bold curve was fitted including all measurement points, the dashed curve in the case of omitting the results for TB_W3 and PNE as outliers.

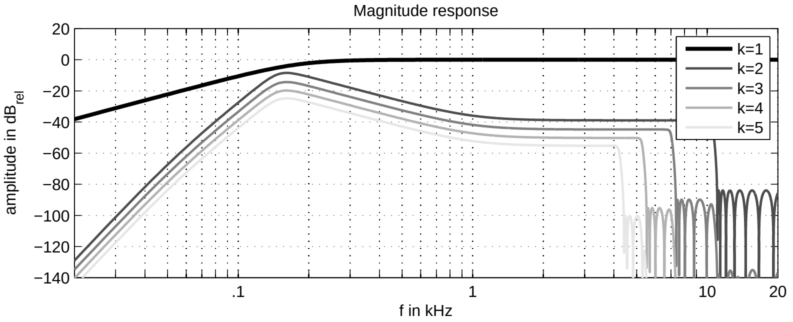


Figure 5.5: Spectrum of the linear and harmonic impulse responses for the modeled loudspeaker, resulting 3% Total Harmonic Distortion. Harmonic orders k>5 were neglected.

5.2 Hammerstein model for loudspeaker nonlinearities

A time invariant nonlinear system with input x and output y can be described using Volterra series as [Sch 80]:

$$\begin{aligned} y(t) = &\int_{-\infty}^{\infty} h_1(\tau_1)x(t-\tau_1)\,d\tau_1 + \iint_{-\infty}^{\infty} h_2(\tau_1,\tau_2)x(t-\tau_1)x(t-\tau_2)\,d\tau_1 d\tau_2 \\ &+ \iiint_{-\infty}^{\infty} h_3(\tau_1,\tau_2,\tau_3)x(t-\tau_1)x(t-\tau_2)x(t-\tau_3)\,d\tau_1 d\tau_2 d\tau_3 + \dots \end{aligned} \tag{5.3}$$

h_1 is the linear impulse response of the system and h_ns are the Volterra kernels or the higher order harmonic impulse responses. The Volterra model of equation 5.3 can be simplified further with the assumption that the memory effects appear in the linear part and the nonlinear part is purely algebraic. The simplest nonlinear model is the Hammerstein model as shown in figure 5.6 which can be described as [Zel 12]:

$$y(t) = h_{lin} * w(t) = \int h_{lin} w(t-k)\,dk \tag{5.4}$$

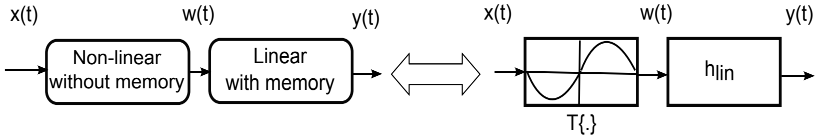


Figure 5.6: Block diagram of Hammerstein nonlinear model (Idea of the figure was adopted from [Zel 12]).

Regarding the linearity of h_{lin} as well as the time invariance of memory-less nonlinear operator T in figure 5.6, and assuming up to order P of nonlinear impulse responses, y can be written as:

$$y(t) = \int h_{lin} T\left(x(t-k)\right) dk = \int h_{lin} \left[\sum_{p=1}^{P} a_p x^p(t-k) \right] = \sum_{p=1}^{P} a_p \int h_{lin} x^p(t-k)\, dk \quad (5.5)$$

According to the fact that in the Hammerstein model the nonlinear part is memory-less, this model can be used for systems where the source of nonlinearity shows time invariant characteristics [Zel 12]. A more generalized Hammersetin model can be achieved by using power filters, which considers various independent linear impulse responses $G_n(f)$, instead of only one impulse response h_{lin} and its corresponding amplification factors a_p. This generalized Hammerstein model is shown in figure 5.7.

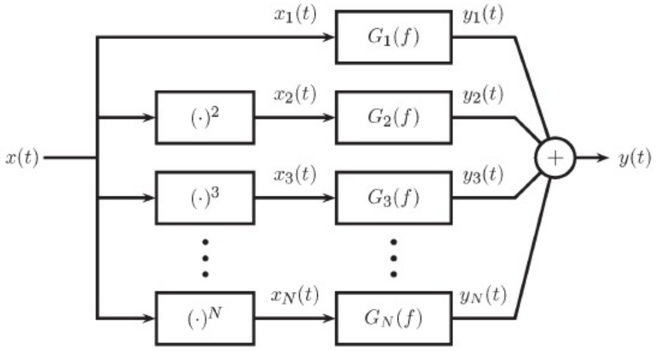


Figure 5.7: Generalized Hammerstein model. [Nov 10]

According to [Nov 10], the linear filters $G_n(f)$ are related to the frequency responses of the system kernels, $H_m(f)$ (up to order N) in the frequency domain, as:

$$H_m(f) = \sum_{n=1}^{N} A_{n,m} G_n(f) \quad (5.6)$$

with the matrix A defined as:

$$A_{n,m} = \begin{cases} \frac{(-1)^{\left(2n + \frac{1-m}{2}\right)}}{2^n - 1} \binom{n}{\frac{n-m}{2}} & \text{for } n \geq m \text{ and } (n+m) \text{ even} \\ 0 & \text{else where} \end{cases} \quad (5.7)$$

H_m can be seen as the system frequency response when only the effect of the input frequency on the m^{th} harmonic of the output is considered . The relation between H and G can be expressed in matrix form as [Nov 10]:

$$\begin{bmatrix} G_1(f) \\ G_2(f) \\ \vdots \\ G_N(f) \end{bmatrix} = \left(A^T\right)^{-1} \begin{bmatrix} H_1(f) \\ H_2(f) \\ \vdots \\ H_N(f) \end{bmatrix} \tag{5.8}$$

So, if the linear transfer function, $H_1(f)$, and the high order transfer functions of the loudspeaker, $H_n(f)$, are known, the weight filters $G_n(f)$ can be calculated by equations 5.7 and 5.8, and the loudspeaker can be modeled by generalized Hammerstein model.

5.3 Modeling the signals captured by in-ear microphones

Apparently, simulations are done in the absence of the real subject. For simulation of HRTF measurements an existing HRTF data bank can be used since any HRTF data bank includes the filtering information of head, pinna and torso of a subject or an artificial head. So, if a new excitation signal is passed through the filters contained in the HRTF data bank, the output is the same as the sound signal which would have been present at the in-ear microphones of the same subject if the HRTF measurement had been repeated for him. The high resolution HRTF data bank, which was available to be used in this thesis, has been measured by Brinkmann et al. for FABIAN [2] head and torso simulator [Lin 06] in the anechoic chamber of Carl von Ossietztky University of Oldenburg [Bri 13]. The grid of source locations in this data bank consists of 11345 points for a fixed 0° head above torso orientation. The grid samples in 2° steps between elevations of $-64°$ and $90°$ and the steps between azimuths were chosen so that the distance between two neighboring points of the same elevation does not exceed two degrees of the greatest circle distance. The grid also includes the horizontal, median and frontal planes. The distance of the sources to the center of the head have been 1.7 m, which implies that the HRTFs can be considered as far field and their dependence on the distance can be neglected [Bru 99]. The HRIRs were acquired point for point by Exponential Sweep measurements. Frequency responses of microphones and loudspeakers as well as the transfer function of measurement equipment were cancelled out. The HRIRs were truncated by 425 samples. The measurement grid used for the modeling was adopted from this data bank. For each point at elevation φ and azimuth θ, the excitation signal was convolved with the corresponding HRIR from the data bank (Ground Truth HRTF data bank) for the left and the right ear at the given elevation and azimuth, to model the signal recorded by the left and right microphones, $y_L(\theta, \varphi)$ and $y_R(\theta, \varphi)$, as depicted in figure 5.8. For source locations at

[2] **F**ast and **A**utomatic **B**inaural **I**mpulse response **A**cquisitio**N**

azimuths which are not contained in the ground truth data set, a representative HRIR was calculated by linear interpolation of HRIRs of the two closest existing azimuths. The elevation of the sound sources stayed the same as the elevations of the measured points in the ground truth data set.

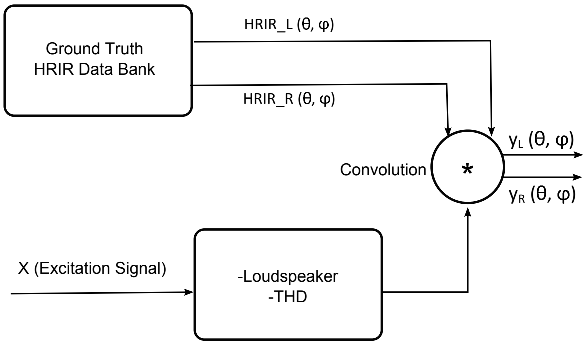


Figure 5.8: Modeling the left and the right microphone signals by convolving the excitation signal with the HRIRs from data bank (Ground Truth Data Bank).

5.4 Modeling the subject's continuous rotation

The model shown in figure 5.8 can be applied to a system-by-system measurement, where the position of the source relative to the microphones stays unchanged during one single measurement. In this case, the ground truth HRIR filters do not change during the whole convolution. In continuous azimuth measurements however, the subject rotates continuously as the loudspeaker (or a number of loudspeakers in case of a multichannel measurement) plays back the excitation signal. If the subject rotates at a speed of $\Delta\theta/dt$ in the horizontal plane (with temporal sample interval $dt = \frac{1}{f_s}$), the HRIR filters change in the time section between t_0 and $t_0 + dt$ from $HRIR(\theta_0, \varphi)$ to $HRIR(\theta_0 + \Delta\theta, \varphi)$, meaning that during the convolution with the excitation signal, the HRIR filter changes constantly. This is the case of non-stationary filtering. Time varying convolutional filtering can be described as non-stationary convolution or non-stationary combination[3] [Mar 98]. To model the subject's continuous rotation, the non-stationary combination was considered. In this case, the non-stationarity of the time-varying impulse response $h(t)$ and its relation to the input signal $u(t)$ and the output signal $v(t)$ is

[3] For a brief discussion on non-stationary convolution and non-stationary combination, see appendix C

described in convolutional matrix form as [Mar 98]:

$$\begin{bmatrix} \vdots \\ v_0 \\ v_1 \\ v_2 \\ \vdots \end{bmatrix} = \cdots \begin{bmatrix} \vdots \\ h_0(t_0) \\ h_1(t_1) \\ h_2(t_2) \\ \vdots \end{bmatrix} u_0 + \begin{bmatrix} \vdots \\ h_{-1}(t_0) \\ h_0(t_1) \\ h_1(t_2) \\ \vdots \end{bmatrix} u_1 + \begin{bmatrix} \vdots \\ h_{-2}(t_0) \\ h_{-1}(t_1) \\ h_0(t_2) \\ \vdots \end{bmatrix} u_2 + \cdots \tag{5.9}$$

The input of the non-stationary filter is the excitation signal (perfect or logarithmic sweep), which repeats periodically due to continuous rotation. After a complete rotation of θ=360°, the recording should continue for another L samples (L being the length of the HRIR filter) to take into account the reverberation information after the excitation signal stops. At this point, rotation stops and therefore the non-stationary filtering changes back to stationary convolution. Figure 5.9 shows the steps of non-stationary combination to model one complete rotation. Note that the time variance of HRIR filters is shown as changes in the current azimuth.

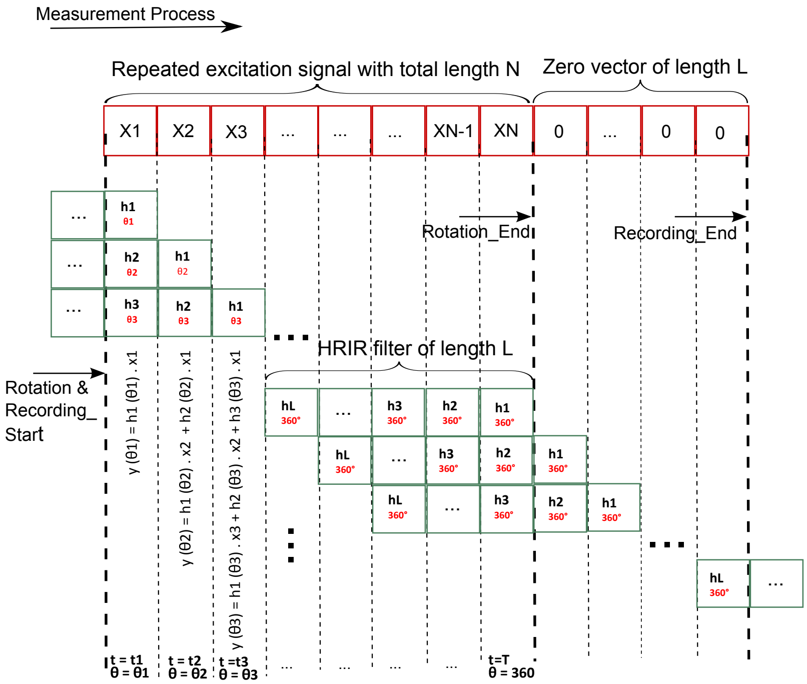


Figure 5.9: Model of subject's one complete rotation with non-stationary combination.

5.5 Modeling the environmental noise

The observed noise at the microphones includes the environmental noise and the noise caused by measurement equipment. The spectral behavior of this noise was adopted from the noise floor measured in the anechoic chamber of Carl von Ossietztky University of Oldenburg [Bri 13]. As can be seen in figure 5.10, the amplitude of the spectrum gets higher for low frequencies which is based on the fact that the isolation of the anechoic chamber decreases for these frequencies.

To have a model for the noise, first a random sequence with standard normal distribution was generated using MATLAB[4] random generator. By applying a low shelf filter to this sequence, it was tried to match its amplitude spectrum to the spectrum of the measured noise (figure 5.10). The shelf frequency and its gain were chosen as 1 kHz and 35dB respectively. Again the curves were shifted vertically against each other to make the comparison possible and only the changing behavior of the amplitude over frequency was of interest. The resulted modeled noise was then applied to the modeled microphone signals within a single HRTF measurement simulation with Exponential Sweep method for an extreme source position, facing directly the left ear (φ=0, θ=90°) and was weighted properly until a desired peak SNR could be read from the HRIRs of the left ear. For this end, an exponential sweep of order 16 (65536 samples $\approx$ 1.5 seconds at f_s=44100Hz) was chosen as the excitation signal. The results of the HRIR for the left ear are shown in figure 5.11 for the case of 60dB and 90dB peak SNR. As this noise was considered to be independent of time and also independent of other signals during the HRTF measurement, it was applied once, at the end of the simulation of one complete rotation, to the modeled microphone signals.

[4] from The MathWorks, Inc.

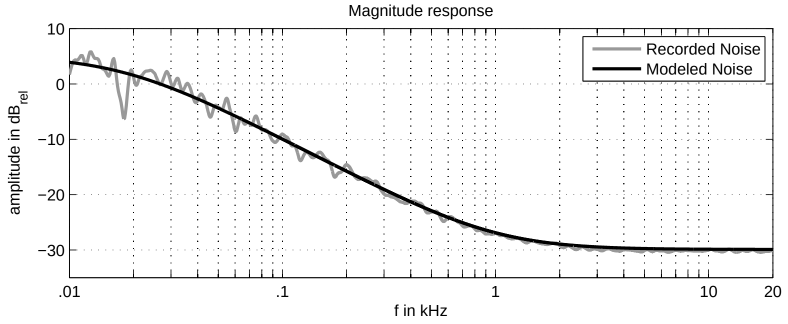


Figure 5.10: Model of the observed noise in the measurement room.

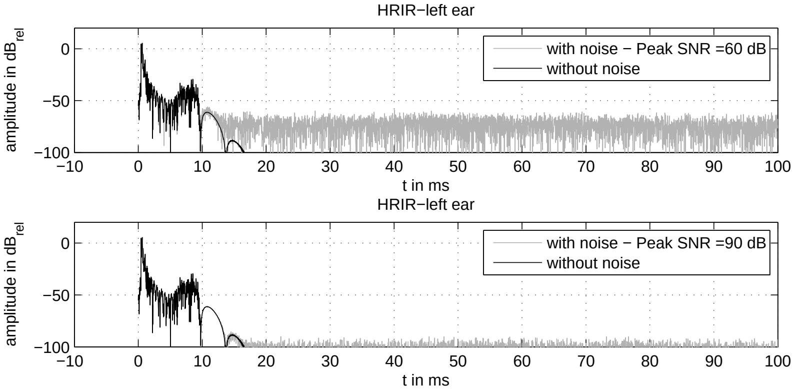


Figure 5.11: Results for a single simulated HRTF measurement with exponential sweep method for the left ear, modeling a peak SNR of 60dB (top) or 90dB (bottom).

Chapter 6

Evaluation of HRTF measurement approaches

In this chapter, a system for the fast measurement of individual head-related transfer functions was simulated. The HRTF measurement was modeled as described in the previous chapter. To obtain the HRTFs, the two discussed system identification algorithms, optimized MESM and NLMS adaptive filtering, were implemented. Each algorithm was first studied separately with respect to its response to varying measurement parameters. Then, the two algorithms were compared regarding their performance. In section 6.1 the criteria are discussed, by which the results have been evaluated. Section 6.2 offers a discussion on the simulated measurement setup and the considerations which have been taken into account for the implemented algorithms. An introduction of the measurement parameters and excitation signals are included as well. The results are presented and discussed in section 6.3.

6.1 Evaluation criteria

The evaluation of results is based on the comparison to the ground truth HRIR data bank which was also used to model the signal captured by the in-ear microphones (see chapter 5.3). The aim is to find out, to what extend the methods for fast HRIR acquisition (the two presented algorithms) differ from a traditional system - by -system measurement. The evaluation criteria are the differences in the ITD and ILD between simulated and ground truth HRIRs (lateralization blur). These differences are named as ILD-error and ITD-error in this thesis. The ILD in dB was calculated as the difference between the rms values of the left and the right HRIR:

$$ILD(\theta, \varphi) = 20log_{10}\frac{rms(HRIR_{Left}(\theta, \varphi))}{rms(HRIR_{Right}(\theta, \varphi))} \tag{6.1}$$

And the ILD-error was calculated as:

$$\text{ILD-error}(\theta, \varphi) = |ILD_{Simulated}(\theta, \varphi) - ILD_{Groundtruth}(\theta, \varphi)| \tag{6.2}$$

The ITD is defined as the difference between the onsets of the left and the right HRIR. Here, the onset was assumed as the ponit, at which the HRIR reaches half of its maximum amplitude. The ITD-error was accordingly calculated as:

$$\text{ITD-error}(\theta, \varphi) = |ITD_{Simulated}(\theta, \varphi) - ITD_{Groundtruth}(\theta, \varphi)| \tag{6.3}$$

An increase in ILD- or ITD-error, which exceeds the audible thresholds of lateralization blur could mean that the implemented method would lead to localization errors which will not happen if the conventionally acquired HRIR (ground truth data set) is used. In the evaluations, the audible threshold levels for lateralization blurs were considered as 0.6 dB and 11 μ seconds for ILD- and ITD-errors respectively. To have an evaluation on the spectral differences between the two HRIR data sets, the method used by Schärer et al. [Sch 09] was considered. This method is based on the auditory filter banks model by Moore [Moo 95], which describes the behavior of human's auditory system with a model consisting of 40 overlapping Equivalent Rectangular Bandwidth (ERB) bandpass filters. For the evaluations, the spectral powers of simulated HRIR and that of the ground truth data (for any source position at elevation φ and azimuth θ) were first filtered by the auditory filter $C(f, f_c)$ with the central frequency f_c and then compared to each other in dB as:

$$E_{Left}(f_c, \theta, \varphi) = 10log_{10} \left[\frac{\int C(f, f_c)|HRTF_{Simulated-Left}(\theta, \varphi)|^2 df}{\int C(f, f_c)|HRTF_{Groundtruth-Left}(\theta, \varphi)|^2 df} \right] \tag{6.4}$$

This difference was calculated for all filter banks for which the central frequency fell between 180Hz and 20kHz (the beginning of the high pass characteristic of the modeled loudspeaker, and the upper limit of the auditory range respectively; on the whole, $N_{auditoryfilter}$= 37 auditory filters) for the left and the right ear. The results were then added and finally averaged to achieve a single value for a given source location, which is named ERB-error in the simulation results:

$$\text{ERB-error}(\theta, \varphi) = \frac{1}{N_{auditoryfilter}} \sum_{f_c} |E_{Left}(f_c, \theta, \varphi)| + |E_{Right}(f_c, \theta, \varphi)| \tag{6.5}$$

To calculate the ERB-filters, `MakeERBFilter` and `ERBFilterBank` from Matlab Auditory Toolbox by Slaney [Sla 98] were used. Schärer et al. [Sch 09] used the spectral comparison of equation 6.4 (without the averaging step of equation 6.5) to study the performance of different inverse filtering methods, which are used to compensate for recording and reproduction systems. They considered deviations greater than 1dB to be perceivable. Minnaar et al. [Min 05] calculated the absolute difference between the magnitudes of interpolated and measured HRTFs at 94 frequency bins, distributed logarithmically on the frequency range of interest, and took the average of the summations for the left and the right ear as a single value to describe the spectral differences. They also reported an audible threshold of 1dB, which was observed through listening tests. Brinkmann et al. [Bri 14] used the single value obtained by equations 6.4 and 6.5 in their investigations on the effect of the head above torso orientation in HRTFs. As their perceptual evaluations resulted, deviations in the range of 0.5 to 1 dB could be considered as slightly perceivable. For the evaluation of the simulation results, ERB-errors less than 1dB were considered as acceptable.

As already stated in chapter 4.5, HRTF acquisition with NLMS leads to a quasi continuous data set. Therefore, in the case of NLMS, the simulated HRIRs can be extracted with a good precision for the measurement points which correspond directly to the HRIR grid of the ground truth data bank. However, as will be discussed shortly later, this is not the case for optimized MESM. For this method, a point to point comparison to the ground truth data set necessitates interpolations. Therefore, optimized MESM suffers in the evaluation from interpolation errors.

6.2 Measurement structure and parameters

The simulated measurement setup is shown in figure 6.1. It consists of a vertical arc with 39 omnidirectional loudspeakers at fixed elevations with 4° steps between $-64°$ and 88°. Subject of interest is located with his head in the center of the arc and is rotated on a turntable continuously in the horizontal direction with microphones at the blocked ear canal entrance [Ham 96]. The distance between the speakers and the subject's head is assumed to be between 1 and 2m and the HRIR dependence on distance is neglected.

For both cases of NLMS and optimized MESM, all loudspeakers were assumed to have the same linear and nonlinear behavior of up to fifth order (as modeled in chapter 5.1). The measurement system worked in its weakly-nonlinear range so that the resulted THD did not exceed 3%. Loudspeakers were assumed to be the main source of nonlinearities. Air sound absorption and phase changes due to sound propagation were neglected. Subject's head remained motionless during the whole measurement and the revolution speed was constant. The sampling rate was 44100Hz and the simulated measurement was assumed to be done in an anechoic chamber.

Frequency response of the in-ear microphones was not modeled, as a result, there was also no need to simulate the reference measurement. The frequency response of the loudspeaker was not compensated from the simulation results, however, as mentioned above, the ERB-error

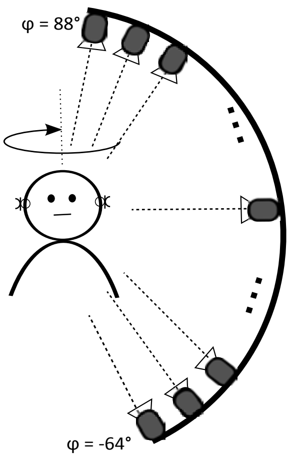


Figure 6.1: Simulated measurement setup.

was evaluated for frequencies already above 180Hz.
The simulations were implemented in Matlab[1].

6.2.1 Considerations on optimized MESM and continuous rotation

As already discussed in chapter 3.5, the two crucial parameters of optimized MESM, namely the sweep rate, r_s, and the time delay between subsequent excitations, τ_w, depend on other involved parameters. To calculate the best case for r_s and τ_w, the ITA-Toolbox, developed at the Institute of Technical Acoustics, RWTH Aachen University was used [Die 13b]. The function `optimized` from the class `itaMSTFinterleaved` in this toolbox receives as input the parameters, which should have been set beforehand through calibration measurements[2]. Then, it looks within a given range of sweep rates and time delays for all possible τ_ws and r_ss, which satisfy the prerequisites of optimized MESM [3], and chooses the best pair (τ_w,r_s), which minimizes the time to measure all M channels once[4]. Due to continuous rotation in the simulations here, subject's azimuth changes even during the excitation of one single channel. This leads inevitably to vagueness in the resulted HRTF after the deconvolution. Subject's azimuth also changes in the time between two subsequent excitations of the same channel. Therefore, two constraints were considered on the revolution speed:

1. During the excitation of each channel the subject must not rotate more than 1° in the azimuthal direction. It means that the sweep rate should correspond to a movement

[1] from The MathWorks, Inc.

[2] These parameters are: length of the HRIR (τ_{DUT}), length of the room impulse response (τ_{IR}), the optional safety time (τ_{sp}), number of the channels (M), number of the harmonic orders (K_{max}), energy decay of the harmonic responses (a_k), Signal to Noise Ratio, frequency range of the exponential sweep and the minimum and maximum limits for the sweep rate, within which the function should look for the best r_s.

[3] See chapter 3.5, equations 3.22 and 3.23

[4] Equation 3.19: $T_{OPT} = T + (M-1)\tau_w + \tau_{st}$

of maximum 1° rotation, by which the changes caused by the rotation remain near the minimum value of the localization blur [Bla 97].

2. Subject's movement between two consequent excitations of the same channel must not exceed 2° of rotation. This constraint aims at an azimuth resolution of 2° or less for the resulted HRIR data set and is based on the highest needed horizontal resolution, which enables interpolations between existing HRTFs without causing audible errors [Min 05].

To calculate the best possible rotation speed, the time for one complete rotation, T_{360}, was calculated for all (τ_w,r_s) pairs, taking into account the two above mentioned constraints. Finally, the pair, for which the calculated T_{360} was the shortest, was chosen. From this pair, the r_s was given to the function `ita_generate_exact_sweep` in the class `itaMSTFinterleaved` to generate the excitation signal and τ_w was used to put the delay between the excitations of consequent channels. As a result, the HRIR grid for optimized MESM could not be set freely and differed for different situations depending on the resulted r_s and τ_w.

Another issue concerning optimized MESM and continuous rotation is that the deconvolution with the known excitation signal leads to HRIRs for given channels (at fixed elevations) which however cannot be assigned to a defined azimuth, but to a range of all azimuths which the subject has passed during the excitation of each channel. Since the frequency of the exponential sweep changes with time during rotation, each azimuth in this range corresponds to another part of the frequency spectrum. This phenomenon was neglected here. Instead, it was decided to choose a proper representative azimuth, to which the resulted HRIR for a given channel could then be assigned. This azimuth was chosen as the point at which the exponential sweep reaches the geometric mean of the two frequencies 3kHz and 16kHz, namely the frequency range within which the predominant individual features with respect to pinna structure are included. For a given non-varying rotation speed and a sweep with defined frequency range and sweep rate, this representative azimuth can be calculated easily.

6.2.2 Different noise levels for NLMS and optimized MESM

To investigate the influence of environmental noise, three cases were considered for the NLMS method: the idealistic case of noiseless environment (infinite SNR), the case of comparativeley quiet environment (90dB peak SNR) and the case of a poorer peak SNR of 60dB[5]. For optimized MESM and according to equation 3.26[6], the case of infinite or very high SNR would imply that the harmonic impulse responses are nearly as long as the linear impulse response. In such cases, the resulted HRIR grid tends to its lowest resolution, leading to an increase in

[5] Peak SNR at one single ES measurement as described in chapter 5.5

[6] Equation 3.26: $\tau_{IR,k} = \frac{SNR-a_k}{SNR}\tau_{IR}$

the interpolation errors. As shown in figure 6.2, there is a slight improvement in the spectral comparison to the ground truth data from the case with 60dB SNR to the case of 90dB. However, the results start to get worse, if SNR is further increased to 150dB. For this reason and in order to avoid the errors which are related to interpolations, the case of infinite SNR was not considered for optimized MESM. Instead, and at the same time in order to attain a sense for the effect of increasing noise, the case of 40dB peak SNR was chosen beside 60dB and 90dB for optimized MESM.

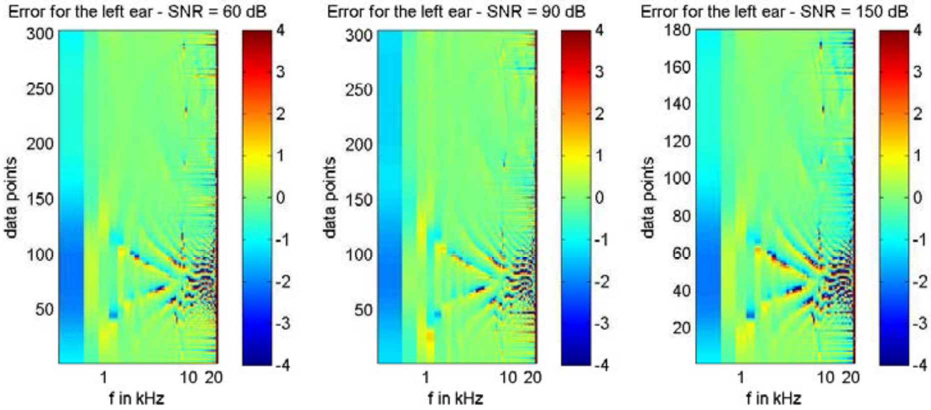


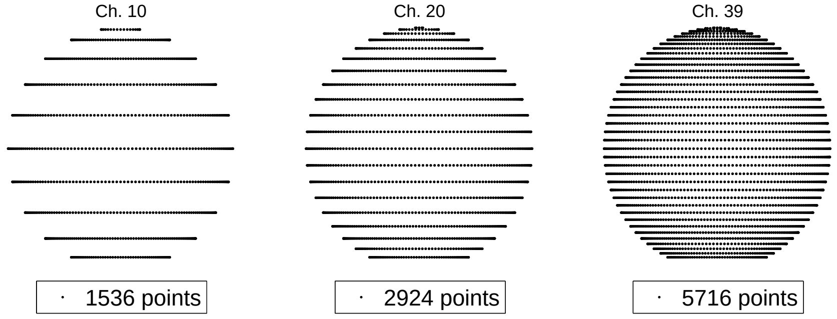
Figure 6.2: Spectral comparison to the ground truth data for 60dB (left), 90dB (middle), and 150dB peak SNR (right): the effect of the interpolation error on optimized MESM for very high SNR values. The vertical axis, named as data points, shows the points assigned to the azimuths from 0° to 360°.

6.2.3 Number of channels, the HRIR grid and the measurement duration

The simulations for both algorithms were accomplished for three cases: 10 channels, 20 channels and 39 channels. In each case the supposed number of channels was distributed equidistantly at elevations between $-64°$ and $88°$ by taking into consideration that only the elevations are allowed which are contained in the grid of the ground truth data set. Figure 6.3 shows the used HRIR grid for the three cases of channel numbers.

The results for the NLMS were acquired directly for these points, whereas the results for optimized MESM were modified after acquisition via interpolation to match these grids.

For NLMS, three revolution times of 1, 5 and 15 minutes were considered. For optimized MESM however, this duration is a result of measurement condition (THD, SNR and number of channels) as well as the constraints regarding the revolution speed, and could not be chosen freely. Table 6.1 shows the resulted revolution time and azimuthal resolution for optimized MESM according to the well-chosen (τ_w,r_s) pair for different simulated cases. The resulted T_{360} durations for optimized MESM imply that the changes in THD or SNR or a change from 10 channels to 20 channels doesn't lead to significantly different revolution times. There is an



Elevation	88°	84°	80°	76°&72°	68°	64°-52°	48°- -48°	-52° - -64°
Az. Resolution	45°	18°	10°	6°	5°	3°	2°	3°

Figure 6.3: Measurement grid for different number of channels.

increase of about 1.5 minutes in T_{360} for the system with 39 channels. A somehow considerable increase can be seen for the system with 39 channels, 3% THD and 90dB measured peak SNR.

Channels	SNR 40 dB			SNR 60 dB			SNR 90 dB		
	THD:0%	1%	3%	THD:0%	1%	3%	THD:0%	1%	3%
10									
T_{360}	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$
(τ_w,r_s)	(0.16,6)	(0.16,6)	(0.16,6)	(0.16,6)	(0.16,6)	(0.16,6)	(0.16,6)	(0.16,6)	(0.16,6)
Az.res.	1°	1°	1°	1°	1°	1°	1°	1°	1°
$\theta_{onesweep}$	1°	1°	1°	1°	1°	1°	1°	1°	1°
20									
T_{360}	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$	$10' \, 7''$	$10' \, 10''$	$10' \, 19''$
(τ_w,r_s)	(0.1,6)	(0.1,6)	(0.1,6)	(0.1,6)	(0.1,6)	(0.1,6)	(0.1,6)	(0.1,5.97)	(0.17,6)
Az.res.	1.18°	1.18°	1.18°	1.18°	1.18°	1.18°	1.18°	1.18°	2°
$\theta_{onesweep}$	1°	1°	1°	1°	1°	1°	1°	1°	0.98°
39									
T_{360}	$11' \, 42''$	$11' \, 42''$	$11' \, 42''$	$11' \, 42''$	$11' \, 42''$	$11' \, 42''$	$11' \, 42''$	$11' \, 42''$	$13' \, 10''$
(τ_w,r_s)	(0.1,5.19)	(0.1,5.19)	(0.1,5.19)	(0.1,5.19)	(0.1,5.4)	(0.1,5.4)	(0.1,5.19)	(0.1,5.4)	(0.1,5.22)
Az.res.	2°	2°	2°	2°	2°	2°	2°	2°	2°
$\theta_{onesweep}$	1°	1°	1°	1°	0.95°	0.95°	1°	0.95°	0.88°

Table 6.1: Revolution time (T_{360}), the well-chosen (τ_w, r_s) pair, the originally calculated azimuthal resolution (Az.Res), and the rotation corresponding to one single sweep ($\theta_{onesweep}$) for different measurement conditions of optimized MESM.

6.2.4 Considerations on NLMS due to the undefined filter states at the beginning

Due to the unknown states of the NLMS filter at the beginning stage of adaptation, it takes a few iterations for the NLMS algorithm to converge to the desired response. If the measurement is supposed to begin directly at the beginning azimuth (θ=0° in our simulations), the results for the first few azimuths of the HRTF grid will be missing or might be incomplete. As a result, before starting the measurement, subject's position was rotated back to an azimuth

θ_{Preset}<0°, as shown in figure 6.4. So, after rotating θ_{Preset}, where the subject passes the azimuth θ=0°, the NLMS filters will have reached their adaptation stage for the whole length of the filter. The HRIR extraction and storage will start as before at θ=0°. θ_{Preset} was set at a point which corresponds to 5 seconds ≡ 220500 samples[7]. Therefore the actual measurement time for NLMS will be 5 seconds longer than the time for one complete rotation.

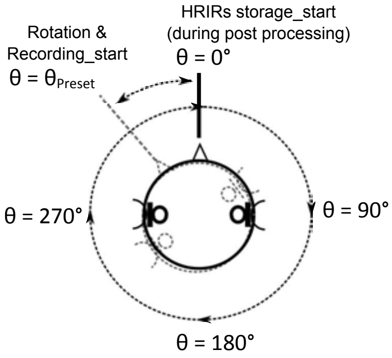


Figure 6.4: Rotating the subject back to θ_{Preset} before starting the measurement, to compensate for the unknown start states of NLMS filters.

6.2.5 Excitation signal and HRIR acquisition

The excitation signal for optimized MESM was an exponential sweep (20Hz < f < 22.05kHz), generated by the function `ita_generate_exact_sweep` from ITA-Toolbox for the chosen sweep rate. Figure 6.5 shows the excitation signal for a system consisting of 10 channels, 3% THD and 60dB SNR, the recorded microphone signal of 1° rotation (all channels excited once) and the result of the deconvolution of these two signals. The HRIR length was set to 176 samples ≡ 4ms. The length of the room impulse response was assumed to be 100ms and 1ms was considered as the safety time before and after the HRIR. The sweep rate range, within which the Toolbox should look for the optimal sweep rate was set to 3< r_s <6, corresponding to sweep order between 16 and 17. The energy attenuation values shown in figure 5.4 were weighted according to the THD value and were considered as energy decay coefficients of the harmonic impulse responses, a_k. The HRIRs were separated later with a rectangular window function. The process shown in figure 6.5 was repeated until the measurement was completed for 360° rotation. For the NLMS algorithm, the excitation signal was a perfect sweep generated in frequency domain as described in section 4.3 (0 < f < 22.05kHz). The excitation signal, generated to measure 10 channels, as well as the recorded microphone signal after a complete rotation of 60 seconds duration, and the result of the multi-channel system identification with the NLMS algorithm are shown in figure 6.6. The NLMS filter length (HRIR length) was set

[7] −30° for T_{360}= 1 minute, −6° for T_{360}= 5 minutes and −2° for T_{360}= 15 minutes

to 156 samples ≡ 3.5 ms. The excitation signal shown in figure 6.6 was repeated periodically for the first channel until one complete rotation of 360° was finished. The excitation signals of the second to the tenth channel were the same, with the difference of a circular shift of 156 samples with respect to their neighbouring channels. For the post processing, three different step size values were considered: 0.25, 0.5 and 1.

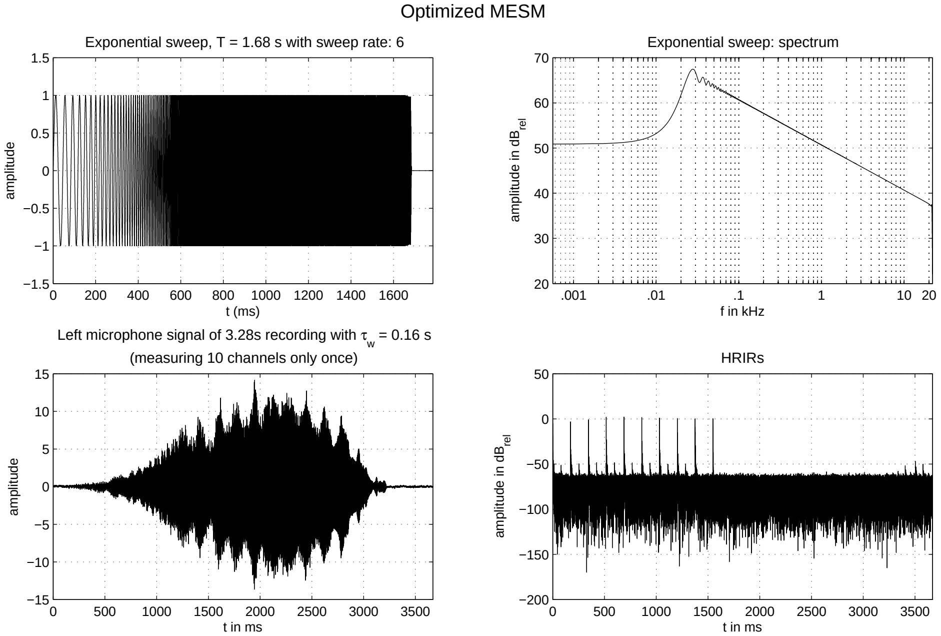


Figure 6.5: Optimized MESM for exciting 10 channels only once, with 3% THD and 60dB SNR: exponential sweep in time and frequency domain (top), microphone signal (bottom left) and the HRIRs after deconvolutoin (bottom right).

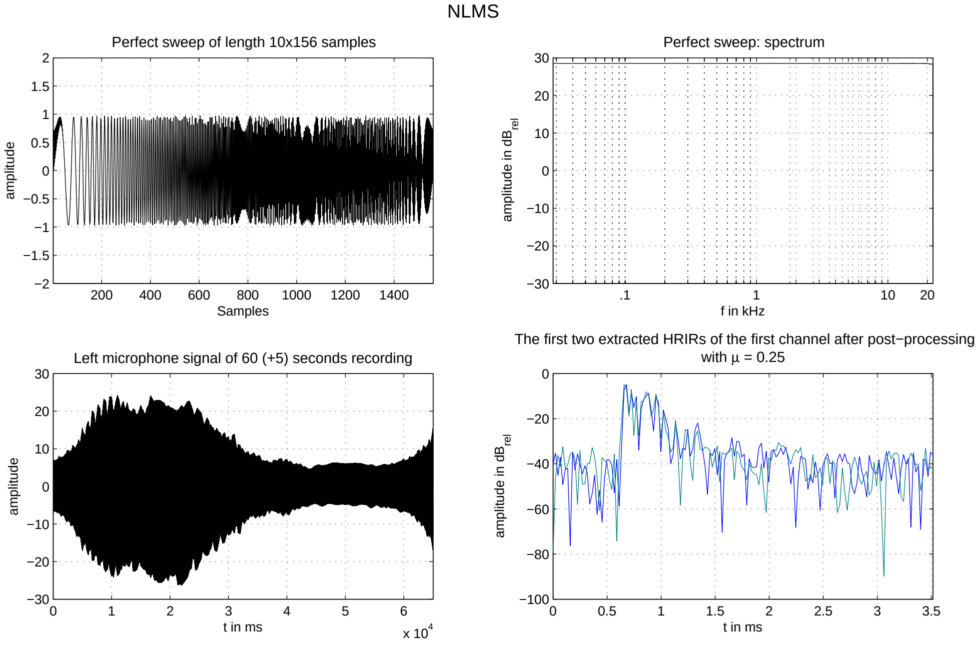


Figure 6.6: NLMS for exciting 10 channels within a complete rotation of 60 seconds duration, with 3% THD and 60dB SNR: perfect sweep in time and frequency domain (top), simulated microphone signal (bottom left) and the first two acquired HRIRs through system identification (bottom right).

6.2.6 An overview of the parameter set used in the simulations

Table 6.2 summarizes the parameter set used in the simulations.

<table>
<tr><td colspan="3">Channel numbers: 10, 20 and 39 channels
THD: 0%, 1% and 3%
Max. order of nonlinearities: 5</td></tr>
<tr><td></td><td>optimized MESM</td><td>NLMS</td></tr>
<tr><td>Excitation</td><td>exponential sweep
freq. Range: 20 - 22050 Hz</td><td>perfect sweep
freq. Range: 0 - 22050Hz</td></tr>
<tr><td>HRIR filter length</td><td>$\tau_{DUT} = 0.004$ s (176 samples)</td><td>NLMS filter length = 0.0035 s (156 samples)</td></tr>
<tr><td>T_{360}</td><td>between ca. 10 and 13 minutes (see table 6.1)</td><td>1, 5, and 15 minutes</td></tr>
<tr><td>SNR</td><td>90, 60, and 40 dB</td><td>∞, 90, and 60 dB</td></tr>
<tr><td>Other parameters</td><td>$\tau_{IR} = 0.1$ s
$\tau_{sp} = 0.001$ s
$3< r_s <6$
a_k: according to the loudspeaker model and THD value (see section 5.1)</td><td>μ: 0.25, 0.5, 1
$\theta_{Preset} \equiv$ 5 seconds of rotation</td></tr>
</table>

Table 6.2: A summary of parameter set used in the simulations.

6.3 Simulation results

Due to the large number of results for different combinations of parameters, the dependency on the azimuth and elevation within the resulted HRTF data bank was not considered. Instead, the 95th percentile of each data set was chosen, that means, the value below which 95% of the results for all azimuths and elevations within a given set of fixed parameters (THD, SNR, number of channels and step size (for NLMS)) fall. In addition, applying the Lilliefors test over the results showed that none of the error results (ILD-, ITD-, and ERB-errors) fall under a normal distribution curve. Therefore, instead of the mean value, the median value of each data set was considered. However, as can be seen in the figures shown in the next section, the behavior of the results for 95th percentile and for median values is in the same direction. Moreover, the judgement, whether the errors exceed the audible thresholds or not, will be made according to the more strict 95th percentile values. Therefore, only these values are considered in the discussions. The median values are nevertheless depicted to provide an additional glimpse at the results.

6.3.1 Optimized MESM simulation results - evaluation

Since the measurement grid for optimized MESM differs according to involved parameters, and in order to avoid the presence of interpolation errors, any comparison is only possible for results, for which the same azimuthal resolution is calculated. According to table 6.1 the original azimuthal resolution shows no changes for different cases within the results of the system consisting of 10 loudspeaker channels (1° for all cases). The same applies for the system with 39 channels (2° azimuthal resolution for all cases). For the system with 20 channels the original resolution is 1.18° for all cases, except for the case with 3% THD and 90dB peak SNR, which has an azimuthal resolution of 2°. Therefore, any analysis should only be done within the results for a fixed number of channels and the effect of differing channel numbers cannot be determined directly. Figure 6.7 summarizes the results for optimized MESM for ILD-, ITD-, and ERB-errors for different number of channels, THD- and SNR values. According to figure 6.7, apart from the exceptional case of 20 channels with 3% THD and 90dB peak SNR, for all other cases the variation in the THD-value leads to deviations less than 0.03dB for ILD-error and less than 0.07dB for ERB-error. There is no difference in the results of ITD-error.

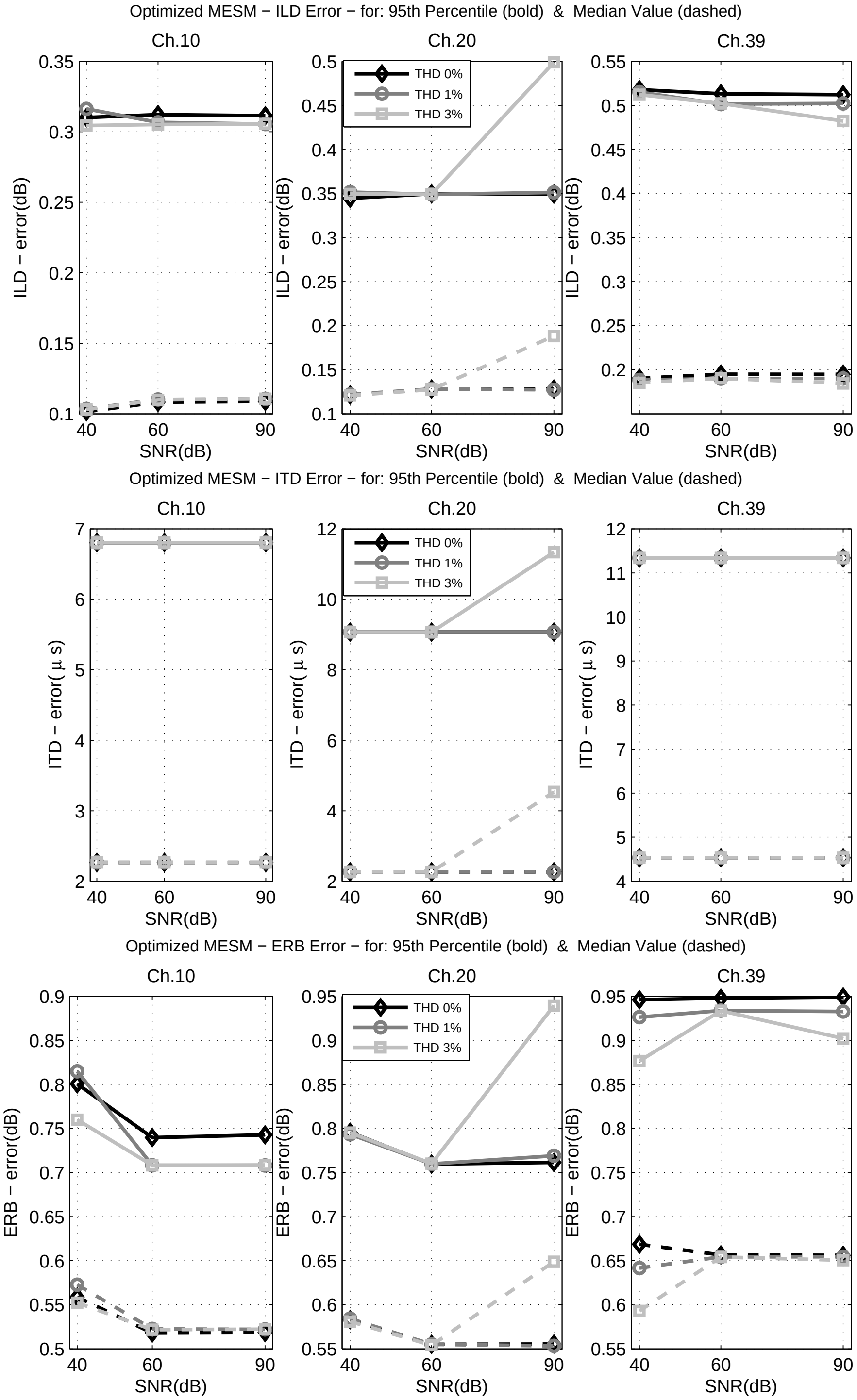


Figure 6.7: Optimized MESM-ILD-error(top), ITD-error(middle) and ERB-error(Bottom) for different THD and SNR values for 10 channels(left), 20 channels(middle) and 39 channels(right)- 95th percentile (bold) and median value (dashed).

The outlier case within the results of the system with 20 channels shows a somehow higher deterioration of about 0.15dB for ILD-error, 2,26 μsec. for ITD-error and 0.18dB for ERB-error. The influence of variations in the SNR value is negligible for ILD- and ITD-errors. In the case of ERB-error, there is a slight improvement for increased SNR (with the exception of the case with 39 channels and 3% THD), however, the deviations are less than 0.06dB and can actually be neglected.

6.3.2 Optimized MESM simulation results - discussion

The simulation results point out a neutral effect of the variations in THD or SNR on the accuracy of optimized MESM. The fact that the outlier in the results of the system with 20 channels corresponds exactly to the case of different azimuthal resolution proposes the idea that the worsening results in this case are related to a consistent error caused by interpolation. This outlier case represents at the same time an interesting point because it has the same azimuthal resolution of 2° as the system with 39 channels and offers therefore a qualitative estimate of the effect of varying number of channels. The comparison between the results of 20 and 39 channels in this situation implies however very small and negligible differences between the two cases (about 0.01dB for ILD-error, 0.03dB for ERB-error, and almost no difference for ITD-error). This implies that the variation in the number of channels has also no effect on the quality of the results and the slight increase in errors from the case with 10 channels to the system with 20 channels could also correspond to the small change in the azimuthal resolution (from 1° for 10 channels to 1.18° for 20 channels).

The performance of the simulated optimized MESM measurement with respect to the perceivable error thresholds is shown in figure 6.8 for the case of 0% THD. The 95% percentile values of the ILD-, ITD-, and ERB-errors fall for all cases in the acceptable region or only slightly out of it. As already remarked in [Die 13a], optimized MESM shows a robust behavior despite the presence of noise or nonlinear distortions and is comparable to a sequential exponential sweep measurement.

6.3.3 NLMS simulation results - evaluation

figure 6.9 shows the ILD-, ITD-, and ERB-errors, for different number of channels, THD- and SNR values, for system identification with the NLMS adaptive filter algorithm. The errors decrease generally with improved SNR and increase for higher number of channels. More on the behavior of errors with SNR or number of channels as well as rotation time will be discussed shortly later. First of all, the comparison of the results shows that different values of applied THD do not cause significant changes in the results. Figure 6.9 shows the results

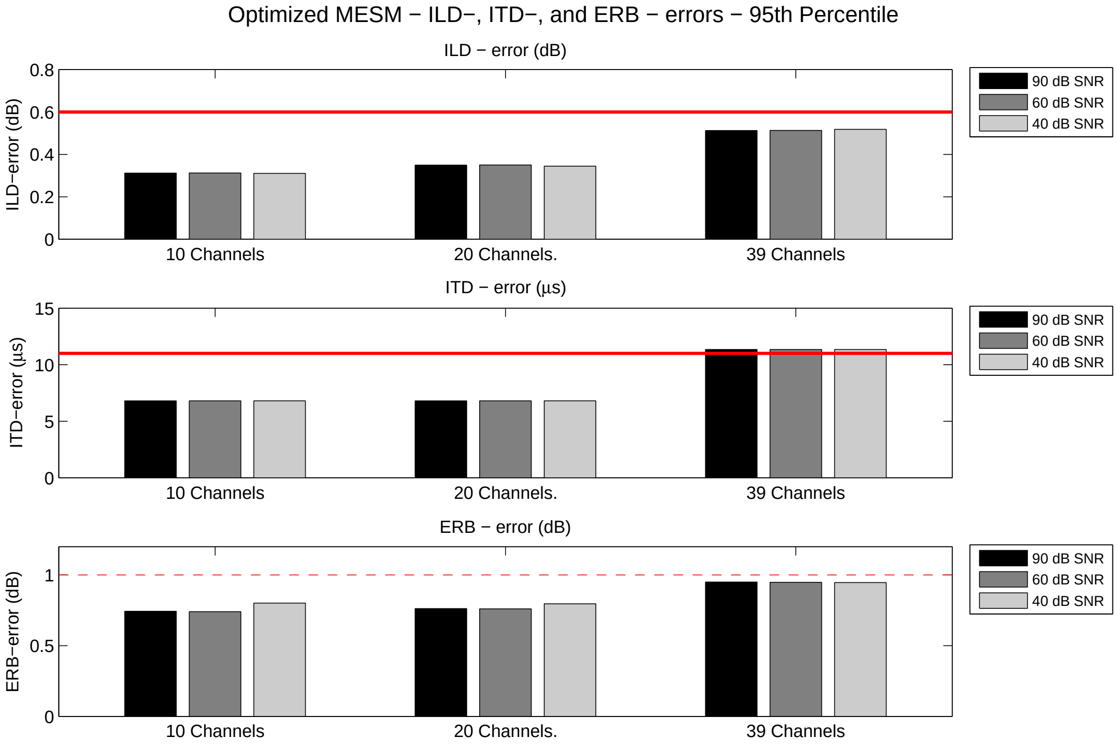


Figure 6.8: Optimized MESM performance: Comparison to the perceivable threshold: ILD-error (top), ITD-error (middle) and ERB-error(bottom).

for the case of step size μ=0.5 and rotation time T_{360}=5 minutes. But, the study of other step size values and rotation times indicates that the deviation between the results for different THD values is less than 0.05 dB for ILD-error, less than 2,26 μseconds for ITD-error and less than 0.25 dB for ERB-error. As a result and for the sake of feasibility, for the rest of the discussion only the case of THD = 0% is considered. The complete set of results for different THD values are available in appendix D.

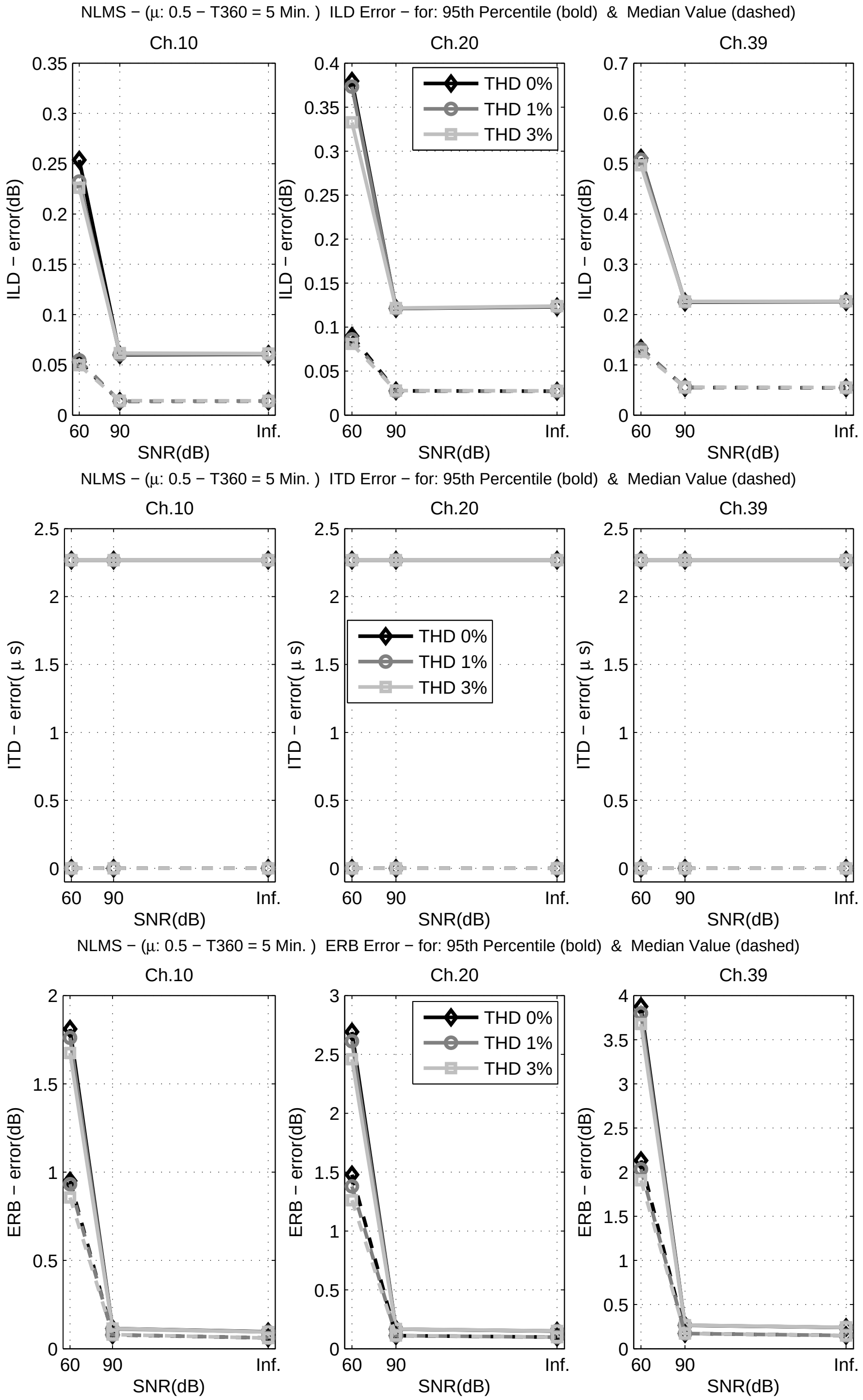


Figure 6.9: NLMS-ILD-error(top), ITD-error(middle) and ERB-error(Bottom)-dependency on THD for different SNR values for 10 channels(left), 20 channels(middle) and 39 channels(right)- 95th percentile (bold) and median value (dashed)- μ=0.5, T_{360}=5 minutes.

NLMS ILD-error

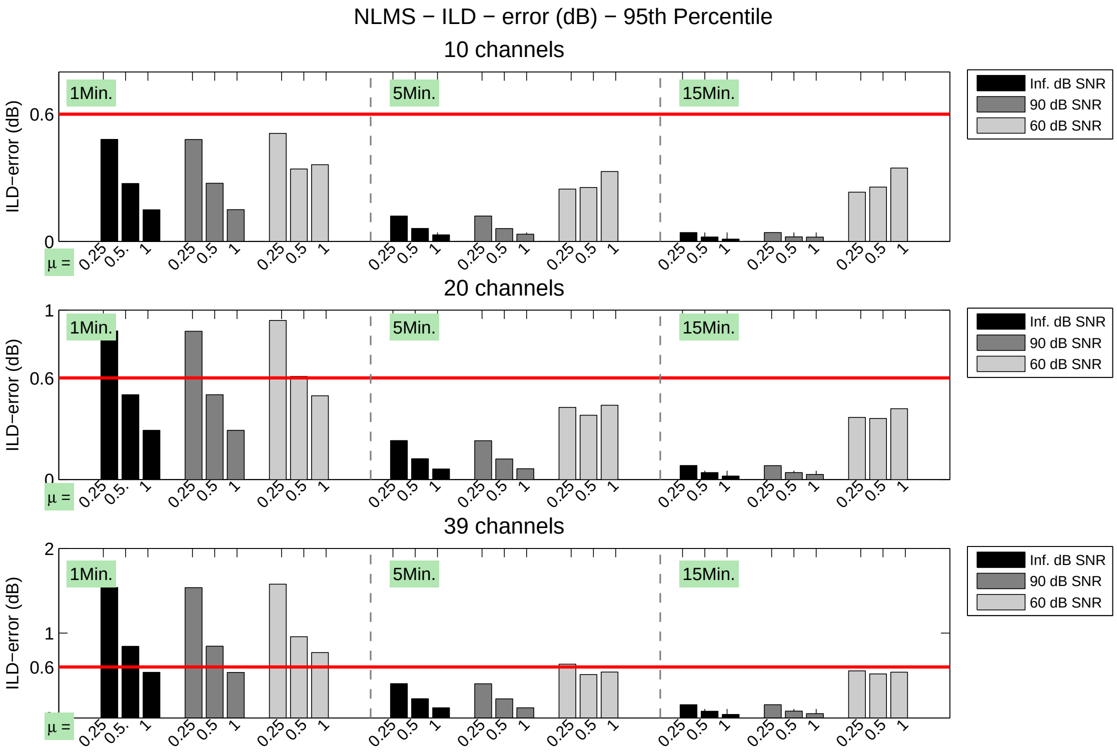


Figure 6.10: NLMS ILD-error and its comparison to the perceivable threshold, for 10 channels (top), 20 channels (middle) and 39 channels (bottom), for 1 Min. (left), 5 Min. (middle) and 15 Min. (right) rotation time, for different SNR and step size values.

According to figure 6.10:

- The ILD-error increases with increasing number of channels. This increase is sharper for shorter T_{360}.
- The ILD-error decreases for longer T_{360}. There is however a saturation in the improvment with incerasing T_{360} and especially in the presence of noise, there is no significant difference between T_{360}=5 and 15 minutes.
- The ILD-error decreases with increasing SNR (with no significant difference between infinite and 90dB SNR). However, for T_{360}=1 minutes the error due to the short measurement time outweights the error caused by SNR.
- The ILD-error decreases with increasing step size. This effect slows down for longer T_{360}. For the case of 60dB SNR the effect of variances in the step size diminishes for T_{360}=5 and 15 minutes.

- For the system with 10 channels, at any case, the 95% percentile value of ILD-errors falls in the acceptable region. For 20 or 39 channels, this meets only for T_{360} of 5 and 15 minutes. For large numbers of channels and short rotation times (T_{360}=1 minutes), this situation can only be achieved with sufficient SNR and a properly chosen step size ($\mu \rightarrow 1$).

NLMS ITD-error

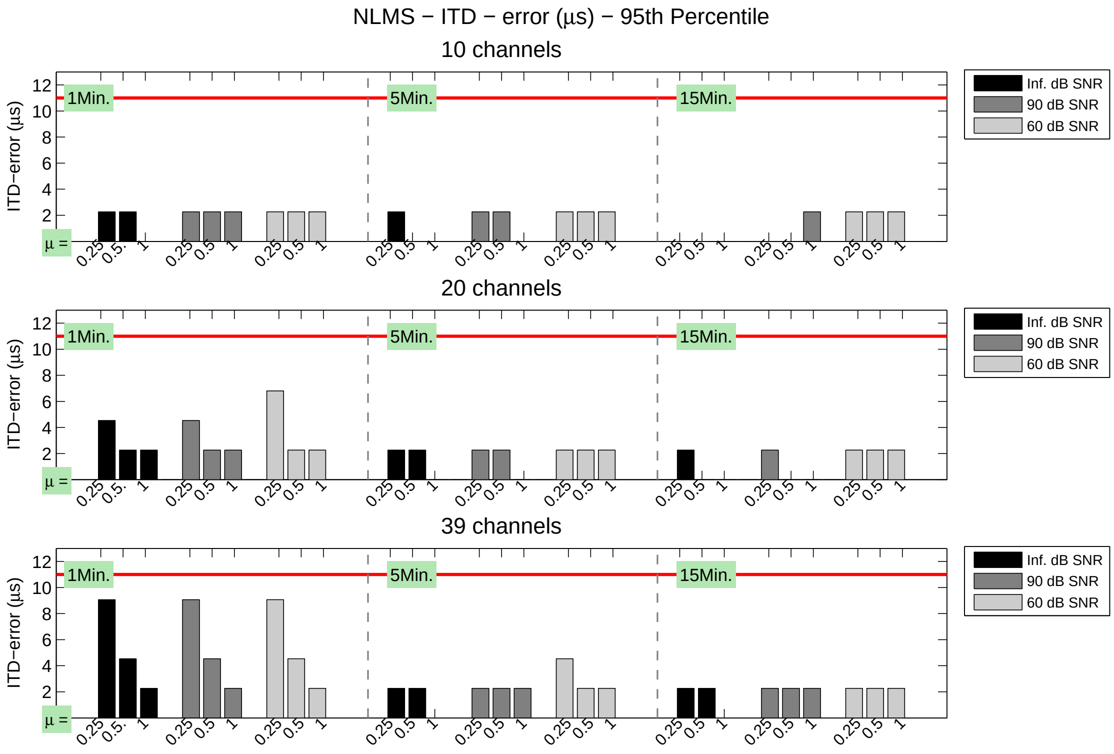


Figure 6.11: NLMS ITD-error and its comparison to the perceivable threshold, for 10 channels (top), 20 channels (middle) and 39 channels (bottom), for 1 Min. (left), 5 Min. (middle) and 15 Min. (right) rotation time, for different SNR and step size values.

According to figure 6.11:

- The ITD-error increases with increasing number of channels. This increase is sharper for shorter T_{360}.
- The ITD-error decreases for longer T_{360}. However, there is no significant differences between T_{360}=5 and 15 minutes.
- The ITD-error decreases with increasing SNR. For T_{360}=1 minutes, the error due to the

short measurement time outweighs the effect of the SNR.

- The ITD-error decreases with increasing step size. For longer T_{360}, the effect of variances in step size vanishes.
- At any case, the 95% percentile value of ITD-errors falls in the acceptable region.

NLMS ERB-error

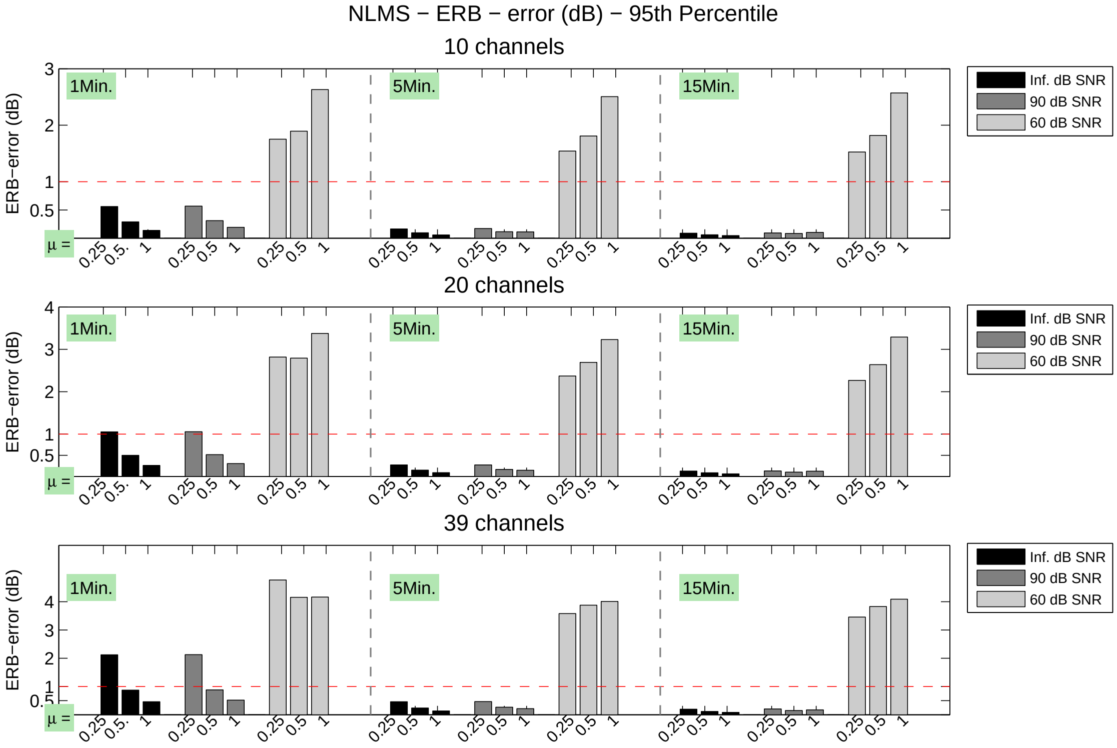


Figure 6.12: NLMS ERB-error and its comparison to the perceivable threshold, for 10 channels (top), 20 channels (middle) and 39 channels (bottom), for 1 Min. (left), 5 Min. (middle) and 15 Min. (right) rotation time, for different SNR and step size values.

According to figure 6.12:

- The ERB-error increases with increasing number of channels. This increase is sharper for shorter T_{360}.
- With 90dB or infinite SNR, the ERB-error decreases for longer T_{360}. With 60dB SNR, there are no significant changes in the ERB-error for different T_{360}.
- The ERB-error decreases with increasing SNR (with no significant difference between infinite and 90dB SNR).

- With 90dB or infinite SNR, the ERB-error decreases with increasing step size, whereas with 60dB SNR, μ<1 is often better.

- The 95% percentile value of ERB-error can be considered as acceptable if there is sufficient SNR (90dB and more) and the step size is chosen properly ($\mu \rightarrow 1$). For 60dB SNR, the 95% percentile value of ERB-error at any case exceeds the perceivable threshold.

6.3.4 NLMS simulation results - discussion

The results of all three evaluation criteria (ILD-, ITD-, and ERB-errors) imply that the error increases with increasing number of loudspeaker channels and decreases for longer revolution times. The improvement with longer rotation times slows down, the longer the revolution time T_{360} gets, especially for the cases with poorer SNR. These results are in good accordance with the behavior of Error Signal Attenuation (ESA).

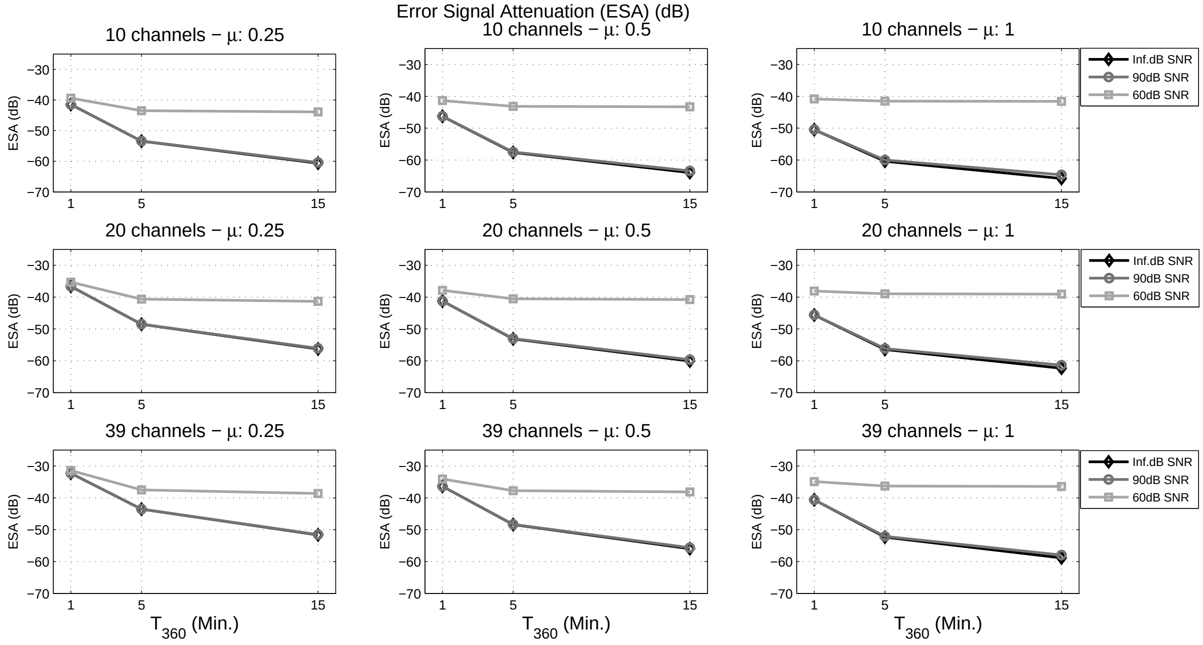


Figure 6.13: Error Signal Attenuation as a function of T_{360} for different SNR values, for 10 channels (top), 20 channels (middle) and 39 channels (bottom), for μ=0.25 (left), μ=0.5 (middle) and μ=1 (right).

As can be seen in figure 6.13, the improvement rate of ESA gets slower with increasing T_{360}. For 60dB peak SNR the improvement from T_{360}= 5 minutes to 15 minutes almost stops. As already discussed in chapter 4.7, the ESA performs as a proper representative of the NLMS inaccuracy. The dropping rate of improvement in the simulation results with increasing T_{360} corresponds to the improvement of NLMS accuracy and its dependency on the modeled dynamic behavior of the system.

For noiseless environments or with 90dB SNR, different evaluation criteria (ILD-, ITD-, and ERB-errors) show a similar behavior with respect to the step size: For all of them the choice of larger step size is applicable to improve the results for a given measurement setup (with defined number of channels and rotation time). For $T_{360} = 1$ minute and with the small step size of μ=0.25, the results of ILD- and ITD-errors for infinite or 90dB SNR are mostly as high as the results for 60dB SNR. Therefore, especially for faster rotation times, a larger step size gains more importance.

For 60dB SNR, the results don't show a common tendency with step size variations. While for the ILD- and ITD-errors the effect of step size variances diminishes for longer T_{360}, for the ERB-error there still can be up to 1dB improvement, which is due to variations in the step size. This can even be observed at T_{360}=15 minutes. In addition, the best result in the ERB-error is often the case for μ<1. It seems that the ERB-error is more correlated with the environmental noise than with the error due to time variations and profits from the noise rejection effect of smaller step size values. However, according to figure 6.12, with 60dB SNR, the ERB-error always exceeds the perceivable threshold. Even variances in the step size cannot pull down the errors below this threshold. Therefore, the best performance of the NLMS, for which all three evaluation criteria fall at the same time in the acceptable region, can only be achieved for infinite or 90dB SNR.

For a real measurement, the SNR might be somewhere between 60dB and 90dB. Since it is not clear, how the behavior of results with respect to the step size would be, it is advisable to choose a smaller step size at any case to improve the results in favor of the ERB-error. As already mentioned, for rotation times of 5 minutes and more, the step size alteration doesn't cause significant changes in the ILD- and ITD-error. Even for smaller step size values, these errors still fall below the audible thresholds. So, if the revolution time is long enough (at least 5 minutes), a smaller μ is still able to keep track of time variations and offers at the same time the advantage of smoothing effects in case of unwanted sudden disturbances. However, for T_{360}=5 minutes, the step size should not become very small, because, for very small step size values the accuracy of the NLMS algorithm begins again to deteriorate (see figure 6.14). As a result, step size values smaller than 0.25 are not recommended.

Besides choosing a smaller step size, there is also another way to improve the results further, at the cost of measurement duration: referring to the simulation results, measuring with less number of channels leads to less errors. So, the idea is to divide 39 channels into two groups of 20 channels (actually one group with 20 channels and another with 19 channels), and perform two separate measurements, each of 5 minutes duration. This will result in a measurement time of 10 minutes, plus the time which is spent between the two measurements to set the

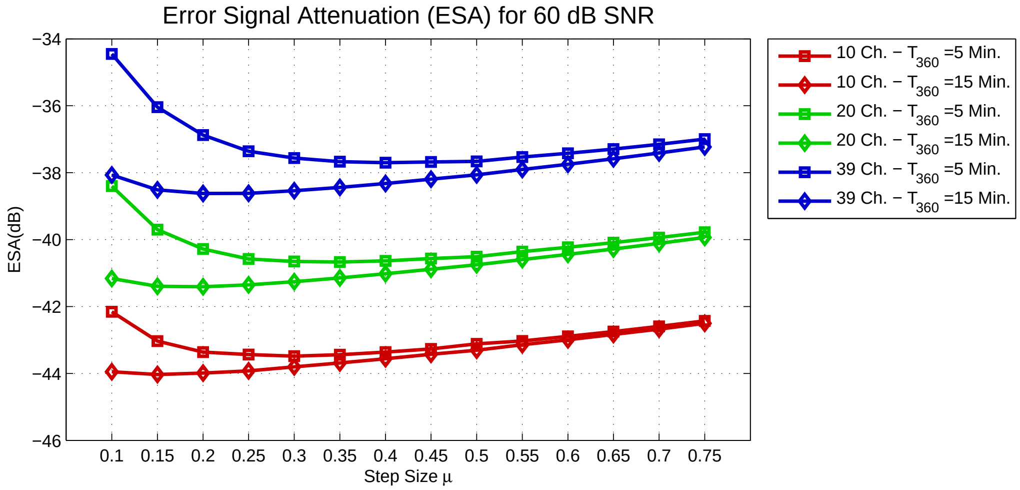


Figure 6.14: Error Signal Attenuation resulted for different step size values, for 60 dB SNR.

system. The total measurement duration could be between 10 to 15 minutes, but there can be up to 1dB improvement in the ERB-error in comparison to one single measurement of 39 channels with 5 or 15 minutes rotation. Although 1dB improvement doesn't offer the required reduction in the ERB-error in case of 60 dB SNR, but breaking the measurement into two parts to take advantage of less involved channels can be considered as an option, which contributes to better results. However, since the HRTF acquisition for all channels through one single measurement is of more interest, an improved SNR, as long as it is possible, offers the better solution.

6.3.5 Optimized MESM or NLMS?

For both implemented methods, there are conditions, under which the algorithms lead to the fulfillment of the constraints included in the evaluation criteria. This meets for optimized MESM, unaffected from variations in the number of channels or SNR, if the less strict evaluation criterion on ERB-error (1dB threshold) is acceptable. At the same time, the measurement duration for optimized MESM varies according to the used parameters and cannot be reduced without loss of accuracy in the azimuthal resolution. Assuming a THD $< 3\%$, it takes at least 11 minutes for optimized MESM to measure 39 channels under 90 dB SNR. On the other hand, without the need of changing the measurement duration, optimized MESM has the advantage, that the quality of the results doesn't change if SNR is reduced to 60 dB or even less.

For NLMS, meeting the constraints of the evaluation criteria all at the same time is only possible for 90dB SNR or more. If only 10 or 20 channels are to be measured, it is possible to obtain satisfactory results even with 1 minute rotation, as long as the step size is set large enough. To measure 39 channels in one single measurement, at least 5 minutes are required.

However, if the SNR is decreased to 60dB, according to the results of ERB-error, the spectral colorations will become audible. In this case, a smaller step size (μ=0.25) can makes the results become better, although this improvement is not enough to reduce the errors below the perceivable threshold. A longer rotation time doesn't lead to significant improvements either. In comparison to optimized MESM, the higher sensitivity of the NLMS to the environmental noise can be seen in figure 6.15. This figure shows, as an example, the HRTF acquired for the source at $\theta = 0°$ azimuth and $\varphi = 0°$elevation for both methods.

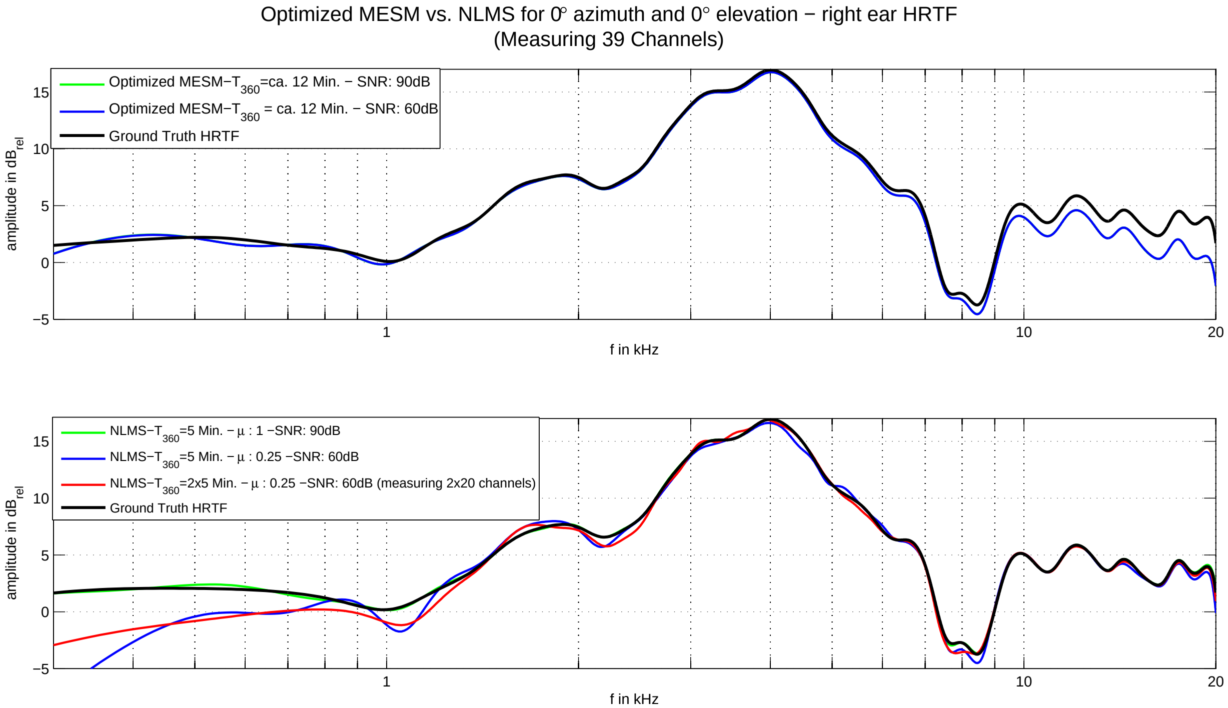

Optimized MESM vs. NLMS for 0° azimuth and 0° elevation − right ear HRTF
(Measuring 39 Channels)

Figure 6.15: Optimized MESM versus NLMS: Simulated HRTF of the right ear for the source position at $\theta = 0°$ azimuth and $\varphi = 0°$elevation. The results for both methods at their best performance (measuring 39 channels in the shortest possible time) under the two SNR conditions of 90dB and 60dB.

Summarizing the discussions, as long as the SNR remains at 90 dB, the NLMS outperforms the optimized MESM regarding the measurement time and HRTF grid resolution. If less channels should be measured, the NLMS can offer satisfactory results even with a 1 minute measurement. For optimized MESM, less channel numbers don't cause any significant reduction in the measurement duration. In addition, for the NLMS, the azimuthal resolution can vary to any precision without the loss of quality or the need for longer measurement times. In contrast, the azimuthal resolution of optimized MESM is limited to the measurement time. Furthermore, as long as one single deconvolution is applied to the data corresponding to a range of azimuths to acquire a single HRTF, this method entails inevitable errors due to continuous rotation. However, if the SNR is decreased to 60dB, the NLMS cannot offer the desirable quality anymore, not even for longer measurement times. In this case, due to robustness in spite of SNR degradation, optimized MESM outperforms the NLMS.

Actually, it is the continuous rotation, which has largely limited the performance of optimized MESM. Otherwise, the stability and robustness of optimized MESM against SNR variations in comparison to NLMS is very remarkable. Of course, in a discrete azimuth measurement, optimized MESM offers results, which, compared to an exponential sweep measurement, are of almost the same quality, while reducing the measurement duration to a large extend. However, as long as there is sufficient SNR, for the continuous rotating system, which was simulated here, the NLMS adaptive filtering is more suitable. It should also be noted, that in the simulations, the measurement room was considered as totally anechoic. In a real measurement however, reflections from objects in the room or from room itself cannot be eliminated. In a real case, a filter length of 156 samples, as used for NLMS in the simulations here, will probably not be sufficient. Therefore, the convergence speed of the NLMS in a real measurement would become limited, as a longer filter length will be needed to cover the reflections and avoid the artifacts due to the truncated length of the impulse response.

Chapter 7

Conclusion

High resolution HRTF data sets demand long and time consuming measurements. Due to the importance of individually measured HRTFs in the improvement of sound localization with binaural signals, a system for the fast measurement of individual head-related transfer functions was simulated and evaluated. This system consisted of a vertical arc of up to 39 loudspeaker canals, with the horizontally continuous rotation of the subject during the measurement, and enabled performing HRTF measurement for 5716 source locations in less than 15 minutes. The sound propagation path between the loudspeaker and the microphone was modeled, including loudspeakers' transfer function and their nonlinear behavior, subject's head, pinna and torso, the continuous rotation of the subject as well as the environmental noise. Two of the existing algorithms were implemented to perform the excitation and acquire the HRTFs from the simulated recordings: the optimized MESM and the NLMS adaptive filtering.

Optimized MESM showed a highly robust performance with respect to variations in the SNR, THD, and the number of the channels. Since optimized MESM is originally supposed for a discrete azimuth measurement, the continuous rotation necessitated limitations on the measurement duration and azimuthal resolution, both of which were dependent on the used parameters and could not be set freely. Since the evaluation of the results was based on the comparison to an existing ground truth HRTF data set with a fixed azimuthal resolution, the results of optimized MESM suffered from interpolation error. However, despite this error, the optimized MESM offered to a very good extend a satisfactory performance with deviations to the ground truth data set, which were below the critical audible thresholds.

For the NLMS adaptive filtering algorithm the results showed also no significant differences for different THD values, but they varied with respect to SNR, measurement duration, number of channels, and step size. In general, the accuracy of the system identification improved with increasing measurement time and SNR, and degraded with increasing number of channels. However, very long measurement times did not necessarily make the results much better. Especially for short measurement times, a larger step size was preferable. Step size values smaller than 1 offered a noise rejection effect. With enough long measurement times (at least 5 minutes), the NLMS algorithm offered results with deviations in the ITD and ILD as well as spectral deviations to the ground truth data, which all fell in the acceptable range. However, as soon as the SNR degraded, the coloration errors presented a problem, independent of revolution time or number of channels.

The comparison of both implemented algorithms showed, that for the modeled HRTF measurement system with continuous rotation, the NLMS algorithm offers the better option, as long as the SNR is sufficiently high. In this case, the NLMS was able to offer satisfactory results within a measurement of 5 minutes duration or even less, with the option of varying azimuthal resolution without any constraints. For acceptable results with the same azimuthal resolution, the optimized MESM required measurement durations at least twice as long as the NLMS algorithm. A reduction in the measurement time for optimized MESM is at the expense of high azimuthal resolution and accuracy. However, since for degraded SNR conditions the NLMS shows audible coloration errors, which cannot be eliminated even with longer measurement durations, in such cases, the optimized MESM outperforms the NLMS due to its robust performance.

There are some points which should be taken into consideration for a real measurement: first of all, the reflections in the measurement room are inevitable and affect the quality of the results for both methods. In particular, the length of the NLMS filter should be modified accordingly. This limits the convergence speed of the algorithm. There is also the need for head tracking, to detect the exact position of the subject's head and considerations given to assigning the HRTFs to the corresponding head-source positions. The latency time between excitation and response of the system should be compensated. In addition, the effect of the transfer functions of the measurement system (Microphone, loudspeakers, amplifiers etc.) should be canceled out and for this aim, reference measurements are necessary. Unwanted sudden disturbances and non optimal performance of the measurement equipment are other constraints. If the new measurement system is used to repeat the HRTF acquisition for a person or an artificial head, for which a formerly measured HRTF dataset already exists, the criteria described in this thesis can be used to evaluate the results. For optimized MESM, this comparison can be improved, if for each frequency bin the corresponding azimuth position of the head is considered for the interpolation. The more accurate and of course the more time consuming choice is to perform psychoacoustic listening tests to verify the performance of binaural signals with the individually measured HRTFs.

Bibliography

[Ahn 08] W. Ahnert and H.P. Tennhardt(2008): "Raumakustik". In: *Handbuch der Audiotechnik (S. Weinzierl)*, Berlin Heidelberg: Springer, pp. 181-266.

[Ajd 07] T. Ajdler, L. Sbaiz and M. Vetterli (2007): "Dynamic measurement of room impulse responses using a moving microphone". In: *J. Acoust. Soc. Am. 122(3)*, pp. 1636-1645.

[Alg 01a] V.R. Algazi, R.O. Duda, D.M. Thompson an C. Avendano (2001): "The CIPIC HRTF Database". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York*, pp. 99-102.

[Alg 01b] V.R. Algazi, C. Avendano and R.O. Duda (2001): "Elevation localization and head-related transfer function analysis at low frequencies". In: *J. Acoust. Soc. Am. 109(3)*, pp. 1110-1122.

[Alg 02] V.R. Algazi, R.O. Duda, R. Duraiswami, N.A. Gumerov and Z. Tang (2002): "Approximating the head-related transfer function using simple geometric models of the head and torso". In: *J. Acoust. Soc. Am. 112(5), Pt.1*, pp. 2053-2064.

[Ant 94] Chr. Antweiler and M. Dörbecker (1994): "Perfect sequence excitation of the NLMS algorithm and its application to acoustic echo control". In: *Annales des Telecommunications 49(7-8)*, pp. 386-397.

[Ant 95] Chr. Antweiler and M. Antweiler (1995): "System identification with perfect sequences based on the NLMS algorithm". In: *International Journal of Electronics and Communication (AEÜ) 49(3)*, pp. 129-134.

[Ant 08] Chr. Antweiler (2008): "Multi-channel system identification with perfect sequences - Theory and applications". In: *Advances in Digital Speech Transmission (R. Martin, U. Heute and C. Antweiler)*, Chichester, UK: John Wiley & Sons Ltd., pp. 171-198.

[Ant 09] Chr. Antweiler and G. Enzner (2009): "Perfect sequence LMS for rapid acquisition of continuous-azimuth head related impulse responses". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 18-21 October, 2009, New Paltz, NY*,pp. 281-284.

[Ant 12] Chr. Antweiler and G. Enzner (2012): "Perfect-sweep NLMS for Time-variant acoustic system identification". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing, Kyoto, Japan*, pp. 517-520.

[Beg 00] D.R. Begault, E.M. Wenzel, A.S. Lee and M.R. Anderson (2000): "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source". In: *108th AES Convention*, Paris, France.

[Ben 01] J. Benesty, T. Gänsler, D.R. Morgan, M.M. Sondhi and S.L. Gay (2001): *Advances in network and acoustic echo cancellation*, Berlin Heidelberg: Springer.

[Bla 97] J. Blauert (1997): *Spatial hearing: the psychophysics of human sound localization*, Revised ed. Camebridge, Massachusetts: MIT Press.

[Bla 08] J. Blauert and J. Braasch (2008): "Räumliches Hören". In: In: *Handbuch der Audiotechnik (S. Weinzierl)*, Berlin Heidelberg: Springer, pp. 87-121.

[Blu 04] A. Blum, B.F.G. Katz and O. Warusfel (2004): "Eliciting adaptation to non-individual HRTF spectral cues with multi-modal training". In: *CFA/DAGA '04*, Strasbourg, pp. 1225-1226.

[Bri 13] F. Brinkmann, A. Lindau, S. Weinzierl, G.Geissler and S. van de Par (2013): "High resolution head-related transfer function data base including different orientations of head above the torso". In: *AIA-DAGA 2013*, Merano, pp. 596-599.

[Bri 14] F. Brinkmann, R. Roden, A. Lindau and S. Weinzierl (2014): "Audibility of head-above-torso orientation in head-related transfer functions". In: *Forum Acusticum*, Krakau, Poland, (accepted article).

[Bru 99] D.S. Brungart and W.M. Rabinowitz (1999): "Auditory localization of nearby sources. Head-related transfer functions". In: *J. Acoust. Soc. Am. 106(3)*, pp. 1465-1479.

[But 77] R.A. Butler and K. Belendiuk (1977): "Spectral cues utilized in the localization of sound in the median sagittal plane". In: *J. Acoust. Soc. Am. 61(5)*, pp. 1264-1269.

[Die 13a] P. Dietrich, B. Masiero and M. Vorländer (2013): "On the optimization of the multiple exponential sweep method". In: *J. Audio Eng. Soc. 61(3)*, pp. 113-124.

[Die 13b] P. Dietrich, M. Guski, J. Klein, M. Müller-Trapet, M. Pollow, R. Scharrer and M. Vorländer (2013): "Measurements and room acoustic analysis with the ITA-Toolbox for MATLAB". In: *AIA-DAGA 2013*, Merano, pp. 1391-1394.

[Enz 08] G. Enzner (2008): "Analysis and optimal control of LMS-type adaptive filtering for continuous-azimuth acquisition of head related impulse responses". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, NV*, pp. 393-396.

[Enz 09] G. Enzner (2009): "3D-Continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 18-21 October, 2009, New Paltz, NY.*

[Enz 13] G. Enzner, Chr. Antweiler and S. Spors (2013): "Trends in Acquisition of individual head-related transfer functions". In: *The technology of binaural listening (J. Blauert)*, Berlin Heidelberg: Springer, pp. 57-92.

[Far 00] A. Farina (2000): "Simultaneous measurement of impulse response and distortion with a swept-sine technique". *Presented at the 108th Convention of the Audio Engineering Society, Preprint 5093*, Paris, France.

[Fel 04] J. Fels, P. Buthmann and M. Vorländer (2004): "Head-related transfer function of children". In: *Acta Acustica united with Acustica 90*, pp. 918-927.

[Fuk 07] K. Fukudome, T. Suetsugu, T. Ueshin, R. Idegami and K. Takeya (2007): "The fast measurement of head related impulse responses for all azimuthal directions using the continuous measurement method with a servo-swiveled chair". In: *Applied Acoustics 68*, pp. 864-884.

[Goe 08] A. Goertz(2008): "Lautsprecher". In: *Handbuch der Audiotechnik (S. Weinzierl)*, Berlin Heidelberg: Springer, pp. 421-490.

[Gon 04] A. González, P. Zuccarello, G. Piñero and M. De Diego (2004): "Simultaneous measurement of multichannel acoustic system". In: *J. Audio Eng. Soc. 52(1/2)*, pp. 26-42.

[Gum 10] N.A. Gumerov, A.E. O'donovan, R. Duraiswami and D.N. Zotkin (2010): "Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation". In: *J. Acoust. Soc. Am. 127(1)*, pp. 370-386.

[Ham 96] D. Hammershøi and H. Møller (1996): "Sound transmission to and within the human ear canal". In: *J. Acoust. Soc. Am. 100(1)*, pp. 408-427.

[Hay 02] S. Haykin (2002): *Adaptive Filter Theory*, 4th ed., Upper Saddle River, NJ: Prentice Hall.

[Ipa 79] V. Ipatov (1979): "Ternary sequences with ideal periodic autocorrelation properties". In: *Radio Engineering Electronics and Physics 24*, pp. 75-79.

[Jin 00] C. Jin, P. Leong, J. Leung, A. Corderoy and S. Carlile (2000): "Enabling individualized virtual auditory space using morphological measurements". In: *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia (2000 International Symposium on Multimedia Information Processing)*, pp. 235-238.

[Kat 01] B.F.G. Katz (2001): "Boundary element method calculation of individual head-related transfer function". In: *J. Acoust. Soc. Am. 110(5), Pt.1*, pp. 2440-2455.

[Kle 93] M. Kleiner, B.I. Dalenbäck and P. Svensson (1993): "Auralization - an overview". In: *J. Audio Eng. Soc. 41(11)*, pp. 861-875.

[Kli 92] W. Klippel (1992): "Nonlinear large-signal behavior of electrodynamic loudspeakers at low frequencies". In: *J.Audio Eng. Soc. 40(6)*, pp. 483-496.

[Lin 06] A. Lindau (2006): *Ein Instrument zur softwaregestützten Messung binauraler Raumimpulsantworten in mehreren Freiheitsgraden*, Magisterarbeit. Technische Universität Berlin, Germany.

[Luk 88] H.D. Lüke (1988): "Sequences and arrays with perfect periodic correlation". In: *IEEE Transactions on Aerospace and Electronic Systems 24(3)*, pp. 287-294.

[Luk 95] H.D. Lüke and H.D. Schotten (1995): "Odd-perfect, almost binary correlation sequences". In: *IEEE Transactions on Aerospace and Electronic Systems 31(1)*, pp. 495-498.

[Mad 00] A. Mader, H. Puder and G.U. Schmidt (2000): "Step-size control for acoustic echo cancellation filters-an overview". In: *Signal Processing 80*, pp. 1697-1719.

[Maj 07] P. Majdak, P. Balazs and B. Laback (2007): "Multiple Exponential Sweep Method for fast measurement of head-related transfer functions". In: *J. Audio Eng. Soc. 55(7-8)*, pp. 623-637.

[Mar 98] G.F. Margrave (1998): "Theory of nonstationary linear filtering in the fourier domain with application to time-variant filtering". In: *Geophysics 63(1)*, pp. 244-259.

[Mas 11] B. Masiero, M. Pollow and J. Fels (2011): "Design of a fast broadband individual head-related transfer function measurement system". In: *Forum Acusticum 2011*, Aalborg, Denmark, pp.2197-2202.

[Min 05] P. Minnaar, J. Plogsties and F. Christensen (2005): "Directional resolution of head-related transfer functions required in binaural synthesis". In: *J. Audio Ang. Soc. 53(10)*, pp. 919-929.

[Møl 92] H. Møller (1992): "Fundamentals of binaural technology". In: *Applied Acoustics 36*, pp. 171-218.

[Møl 95] H. Møller, D. Hammershøi, C.B. Jensen and M.F. Sørensen (1995): "Transfer characteristics of headphones measured on human ears". In: *J. Audio Eng. Soc. 42(4)*, pp. 203-217.

[Møl 96] H. Møller, M.F, Sørensen, C.B. Jensen and D. Hammershøi (1996): "Binaural Technique: Do we need individual recordings?". In: *J. Audio Eng. Soc. 44(6)*, pp. 451-469.

[Møl 97] H. Møller, C.B. Jensen, D.Hammershøi and M.F.Sørensen (1997): "Evaluation of artificial heads in listening tests". In: *102nd AES Convention*, Munich, Germany.

[Moo 95] B.C.J. Moore (1995): "Frequency analysis and masking". In: *Hearing. Handbook of Perception and Cognition (B.C.J. Moore)*, Academic Press, San Diego, pp. 161-205.

[Mos 09] M. Möser (2009): *Technische Akustik*, 8. Aufl., Berlin Heidelberg: Springer.

[Mul 01] S. Müller and P. Massarani (2001): "Transfer function measurement with sweeps". In: *J. Audio Eng. Soc. 49*, pp. 443-479.

[Mul 08] S. Müller (2008): "Messtechnik". In: *Handbuch der Audiotechnik (S. Weinzierl)*, Berlin Heidelberg: Springer, pp. 1087-1169.

[Nic 10] R. Nicol (2010): *Binaural Technology*, New York: Audio Engineering Society Inc.

[Nov 10] A. Novák, L. Simon, F. Kadlec and P. Lotton (2010): "Nonlinear system identification using exponential swept-sine signal". In: *IEEE Transactions on Instrumentation and Measurement 59(8)*, pp. 2220-2229.

[Ota 06] M. Otani and S. Ise (2006): "Fast calculation system specialized for head-related transfer function based on boundary element method". In: *J. Acoust. Soc. Am. 119(5)*, pp. 2589-2598.

[Par 12a] G. Parseihian and B.F.G. Katz (2012): "Morphocons: A new sonification concept based on morphological earcons". In: *J. Audio Eng. Soc. 60(6)*, pp. 409-418.

[Par 12b] G. Parseihian and B.F.G. Katz (2012): "Rapid head-related transfer function adaptation using a virtual auditory environment". In: *J. Acoust. Soc. Am. 131(4)*, pp. 2948-2957.

[Rot 10a] M. Rothbucher, M. Durkovic, H. Shen and K. Diepold (2010): "HRTF customization using multiway array analysis". In: *18th European Signal Processing Conference (EUCIPCO-2010)*, Alborg, Denmark, pp. 229-233.

[Rot 10b] M. Rothbucher, T. Habigt, J. Habigt, T. Riedmaier and K. Diepold (2010): "Measuring anthropometric data for HRTF personalization". In: *2010 sixth Conference on Signal-Image Technology and Internet Based Systems, IEEE Computer Society*, pp. 102-106.

[Sch 80] M. Schetzen (1980): *The Volterra and Wiener theories of nonlinear systems*, New York(NY), USA: Wiley.

[Sch 09] Z. Schärer and A. Lindau (2009): "Evaluation of equalization methods for binaural signals". In: *126th AES Convention*, Munich, Germany.

[Shi 98] B.G. Shinn-Cunningham, N.I. Durlach and R.M. Held (1998): "Adapting to supernormal auditory localization cues". In: *J. Acoust. Soc. Am. 103(6)*, pp. 3656-3676.

[Shm 05] D. Shmilovitz (2005): "On the definition of total harmonic distortion and its effect on measurement interpretation". In: *IEEE Transactions on Power Delivery 20(1)*, pp. 526-528.

[Sla 98] M. Slaney (1998): *Auditory Toolbox*, Version2, Technical Report # 1998-010, Interval Research Corporation.

[Slo 93] D.T.M. Slock (1993): "On the convergence behavior of the LMS and the Normalized LMS algorithms". In: *IEEE Transactions on Signal Processing 41(9)*, pp. 2811-2825.

[Som 89] P.C.W. Sommen and C.J. van Valburg (1989): "Efficient realization of adaptive filter using an orthogonal projection method". In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 940-943.

[Tar 88] M. Tarrab and A. Feuer (1988): "Convergence and performance analysis of the Normalized LMS algorithm with uncorrelated Gaussian data". In: *IEEE Transactions on Information Theory 34(4)*, pp. 680-691.

[Tel 10] A. Telle, Chr. Antweiler and P. Vary (2010): "Der perfekte Sweep-ein neues Anregungssignal zur adaptiven Systemidentification zeitvarianter akustischer Systeme". In: *DAGA 2010*, Berlin, pp. 341-342.

[Tor 11] A. Torras-Rosell and F. Jacobsen (2011): "A new interpretation of distortion artifacts in sweep measurements". In: *J. Audio Eng. Soc. 59(5)*, pp. 283-289.

[Vol 08] F. Völk, F. Heinemann and H. Fastl (2008): "Externalization in binaural synthesis: effects of recording environment and measurement procedure". In: *Acoustics 08*, Paris, France, pp. 6419-6424.

[Vor 08] M. Vorländer (2008): *Auralization. Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, 1st Edition, Berlin Heidelberg: Springer.

[Wei 09] S. Weinzierl, A. Giese and A. Lindau (2009): "Generalized multiple sweep measurement". In: *126th AES Convention*, Munich, Germany.

[Wen 88] E. Wenzel, F. Wightman, D. Kistler and S. Foster (1988): "Acoustic origins of individual differences in sound localization behavior". In: *J. Acoust. Soc. Am. Suppl.1 (84)*,S79.

[Wen 93] E.M. Wenzel, M. Arruda, D.J. Kistler and F.L. Wightman (1993): "Localization using nonindividual head-related transfer functions". In: *J. Acoust. Soc. Am. 94(1)*, pp. 111-123.

[Wid 85] B. Widrow and S.D. Stearns (1985): *Adaptive Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall.

[Wnrt 12] M. Weinert, G. Enzner, J.M. Batke, P. Jax and Chr. Antweiler (2012): "Komfortable Messung und Bereitstellung individueller kopfbezogener Impulsantworten als OpenDAFF". In: *DAGA 2012*, Darmstadt, pp. 703-704.

[Zah 06] P. Zahorik, P. Bangayan, V. Sundareswaran, K. Wang and C. Tam (2006): "Perceptual recalibration in human sound localization: Learning to remediate front-back reversals". In: *J. Acoust. Soc. Am. 120(1)*, pp. 343-359.

[Zel 12] M. Zeller (2012): *Generalized nonlinear system identification using adaptive volterra filters with evolutionary kernels*, Dissertation, Technische Fakultät der Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany.

[Zot 06] D.N. Zotkin, R. Duraiswami, E. Grassi and N.A. Gumerov (2006): "Fast head-related transfer function measurement via reciprocity". In: *J. Acoust. Soc. Am. 120(4)*, pp. 2202-2215.

List of Figures

List of Tables

Appendix A

LMS and NLMS adaptation algorithm

The basic idea of the method of steepest descent is to find an optimal solution $\mathbf{w}_0$ among some unknown vectors $\mathbf{w}$, which satisfies the following condition:

$$J(\mathbf{w}_0) \leq J(\mathbf{w}) \qquad \text{for all } \mathbf{w} \tag{A.1}$$

where the cost function $J(\mathbf{w})$ is a continuously differentiable function of $\mathbf{w}$. For the LMS algorithm, the cost function is the mean square of the deviation of the response of the filter $\mathbf{w}$ to a desirable response. A well suited condition is to assume that the cost function is reduced at each iteration:

$$J(\mathbf{w}(k+1)) < J(\mathbf{w}(k)) \tag{A.2}$$

For the method of steepest descent the adjustment applied to the weight vector $\mathbf{w}$ is in a direction opposite to the gradient vector of the cost function. The gradient vector of the cost function is given as [Wid 85]:

$$\bigtriangledown J(k) = -2\mathbf{C} + 2\mathbf{R}\mathbf{w}(k) \tag{A.3}$$

where $\mathbf{R}$ is the correlation matrix of the input of the filter and $\mathbf{C}$ is the cross-correlation vector betweeen the desired response and the same input. The steepest descent algorithm is described by [Hay 02]:

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \frac{1}{2}\mu \bigtriangledown J(k) \tag{A.4}$$

with the following adjustment going from iteration k to $k+1$:

$$\delta\mathbf{w}(k) = \mathbf{w}(k+1) - \mathbf{w}(k) = -\frac{1}{2}\mu \bigtriangledown J(k) \tag{A.5}$$

k denotes the iteration and μ is the step size. The question is, whether this formulation for steepest descent algorithm satisfies the condition in equation A.1. With the assumption that μ is small, one can use the first order Taylor series expansion around $\mathbf{w}(k)$ to obtain the approximation:

$$J(\mathbf{w}(k+1)) \approx J(\mathbf{w}(k)) + \bigtriangledown J^T(k)\delta\mathbf{w}(k) \tag{A.6}$$

Substituting equation A.5 in A.6 yields:

$$J(\mathbf{w}(k+1)) \approx J(\mathbf{w}(k)) - \frac{1}{2}\mu \left\| \bigtriangledown J(k) \right\|^2 \tag{A.7}$$

According to equation A.7, $J(\mathbf{w}(k+1))$ is smaller than $J(\mathbf{w}(k))$ if μ is positiv. This also shows that the cost function decreases with increasing k, approaching the minimum value at $k = \infty$. The difference between the LMS algorithm and the steepest descent algorithm is that the method of steepest descent uses exact measurements of the gradient vector at each iteration, whereas the LMS algorithm relies on an estimation of the gradient vector. In order to have an estimate of the gradient vector, an instantaneous estimate of $\mathbf{R}$ and $\mathbf{C}$, based on the sample values of the tap input vectors and the desired response, y, can be used as following [Hay 02, Wid 85]:

$$\begin{aligned} \hat{\mathbf{R}}(k) &= \mathbf{p}(k)\mathbf{p}^T(k) \\ \hat{\mathbf{C}}(k) &= \mathbf{p}(k)y^T(k) \end{aligned} \tag{A.8}$$

Substituting these estimations in equation A.3, and again, substituting the result in equation A.4, we can get a recursive relation for updating the tap-weight vector:

$$\mathbf{h}(k+1) = \mathbf{h}(k) + \mu\mathbf{p}(k)e(k) \tag{A.9}$$

with:

$$e(k) = y(k) - \mathbf{p}^T(k)\mathbf{h}(k) \tag{A.10}$$

The adjustment of the filter in LMS method depends directly on the input vector $\mathbf{p}(k)$. Since the LMS algorithm suffers from a gradient noise due to estimations, this problem can get worse for large inputs. The Normalized Least Mean Square method (NLMS) overcomes this problem by normalizing the adjustment at iteration $k+1$ with respect to the squared Euclidean norm of the input vector at iteration k such that:

$$\mathbf{h}(k+1) = \mathbf{h}(k) + \frac{\mu}{\left\|\mathbf{p}(k)\right\|^2}\mathbf{p}(k)e(k) \tag{A.11}$$

with $e(k)$ being set by equation A.10.

Appendix B

Transfer function of an electro dynamic loudspeaker - Derivation

[1] An electro dynamic loudspeaker can mechanically be modeled as a sprind-mass system. The electro dynamic force generates a relative movement between loudspeaker diaphragm and its housing. The housing is relatively heavy and an be seen as stationary. Therefore the important mass m is of the loudspeaker diaphragm which can also include the mass of the coil or other moving parts. This mass is resiliently mounted to the heavy immpbile housing which together build a resonator. On the mass m act three external forces: the exciting electro dynamic force, as well as the resisting forces due to spring stiffness and friction. According to Hooke's law and assuming a velocity-proportional friction and also assuming pure tones, the exciting electro dynamic force can be written as:

$$F = \left(j\omega m + r + \frac{D}{j\omega} \right) \nu \tag{B.1}$$

In which:
ν: velocity of the diaphragm
r : friction coeficient
D: spring stiffness
The resonance frequency of the spring-mass system is at $\omega_0 = \sqrt{\frac{D}{m}}$. On the other hand the electro dynamic force is related to the electric current I, magnetic flux density B and the length of the coil l:

$$F = BIl \tag{B.2}$$

The quivalent circuit for the electrical behavior of the loudspeaker is shown in figure B.1.

[1] From [Mos 09]

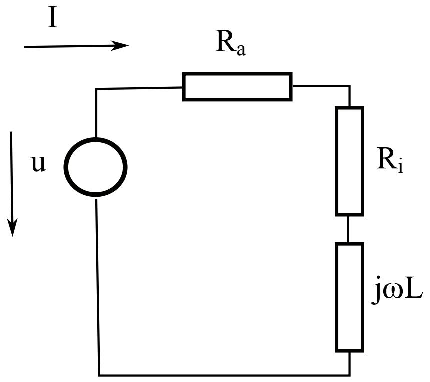


Figure B.1: Equivalent electrical circuit of the electro dynamic loudspeaker.

According to figure B.1 the electrical current can be written as:

$$I = \frac{u}{R_i + R_a + j\omega L} \tag{B.3}$$

with $R_i + j\omega L$ as the impedance of the loudspeaker and R_a as the impedance of the voltage source. Substituting equations B.1 and B.3 in equation B.2 we have:

$$\frac{\nu}{u} = Bl \frac{1}{R_i + R_a + j\omega L} \frac{1}{j\omega m + r + \frac{D}{j\omega}} \tag{B.4}$$

Finally considering the loudspeaker as a spherical volume source one can describe the diaphragm velocity ν in relationship to the spund source as:

$$p = \frac{j\rho_0 \omega \pi a^2 \nu}{4\pi d} e^{j(\omega t - kd)} \tag{B.5}$$

With:
a: radius of the spherical volume source
d: distance to the source
k: wave number
Combining ewuations B.4 and B.5 the transfer function of the loudspeaker is achievd as:

$$\frac{p}{u} = Bl \frac{1}{R_i + R_a + j\omega L} . \frac{1}{j\omega m + r + \frac{D}{j\omega}} . \frac{j\rho_0 \pi a^2}{4\pi d} e^{j(\omega t - kd)} \tag{B.6}$$

Appendix C

Non-stationary convolution and combination

[1] For stationary convolutional filtering the response of the filter to an impulse response is known at any particular time and for any input $u(t)$ the filtered output is the convolution of input signal with the impulse response of the filter $h(t)$:

$$v(t) = \int_{-\infty}^{\infty} h(t-\tau)u(\tau)\,d\tau = h(t) * u(t) \tag{C.1}$$

In discrete form the integral of equation C.1 can be written in matrix form as:

$$\begin{bmatrix} \vdots \\ v_0 \\ v_1 \\ v_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & h_0 & h_{-1} & h_{-2} & \vdots \\ \vdots & h_1 & h_0 & h_{-1} & \vdots \\ \vdots & h_2 & h_1 & h_0 & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \vdots \\ u_0 \\ u_1 \\ u_2 \\ \vdots \end{bmatrix} \tag{C.2}$$

Equation C.2 can also be rewritten after matrix multiplication as [Mar 98]:

$$\begin{aligned} &\vdots \\ v_0 &= \cdots + h_0u_0 + h_{-1}u_1 + h_{-2}u_2 + \cdots \\ v_1 &= \cdots + h_1u_0 + h_0u_1 + h_{-1}u_2 + \cdots \\ v_2 &= \cdots + h_2u_0 + h_1u_1 + h_0u_2 + \cdots \\ &\vdots \end{aligned} \tag{C.3}$$

[1] From [Mar 98]

Equation C.3 can again be rewritten in matrix form, this time by putting the filter coefficients h_k in column matrices:

$$\begin{bmatrix} \vdots \\ v_0 \\ v_1 \\ v_2 \\ \vdots \end{bmatrix} = \cdots \begin{bmatrix} \vdots \\ h_0 \\ h_1 \\ h_2 \\ \vdots \end{bmatrix} u_0 + \begin{bmatrix} \vdots \\ h_{-1} \\ h_0 \\ h_1 \\ \vdots \end{bmatrix} u_1 + \begin{bmatrix} \vdots \\ h_{-2} \\ h_{-1} \\ h_0 \\ \vdots \end{bmatrix} u_2 + \cdots \tag{C.4}$$

According to matrix form of equation C.4 each input sample u_k is used to weight a time shifted version of $h(t)$. Since equation C.4 shows the case of stationary filtering the column vectors at the right hand side of the equation are the same except for a time shift. But in case of non-stationary filtering they vary as the filter varies with the time:

$$\begin{bmatrix} \vdots \\ v_0 \\ v_1 \\ v_2 \\ \vdots \end{bmatrix} = \cdots \begin{bmatrix} \vdots \\ h_0(t_0) \\ h_1(t_0) \\ h_2(t_0) \\ \vdots \end{bmatrix} u_0 + \begin{bmatrix} \vdots \\ h_{-1}(t_1) \\ h_0(t_1) \\ h_1(t_1) \\ \vdots \end{bmatrix} u_1 + \begin{bmatrix} \vdots \\ h_{-2}(t_2) \\ h_{-1}(t_2) \\ h_0(t_2) \\ \vdots \end{bmatrix} u_2 + \cdots \tag{C.5}$$

Considering the time variance of filter h the non-stationary convolutional integral can be written as:

$$v(t) = \int_{-\infty}^{\infty} h(t-\tau, \boldsymbol{\tau}) u(\tau)\, d\tau \tag{C.6}$$

Equation C.6 is known as non-stationary convolution, which describes the non-stationarity as a function if input time τ. Non-stationary convolution preserves the impulse response in the columns of the convolutional matrix (see equation C.2). Non-stationarity can also be considered as a function of output time t, known as non-stationary combination:

$$v(t) = \int_{-\infty}^{\infty} h(t-\tau, \mathbf{t}) u(\tau)\, d\tau \tag{C.7}$$

In the case of non-stationary combination the impulse response is preserved as rows of the convolutional matrix and can be written as:

$$\begin{bmatrix} \vdots \\ v_0 \\ v_1 \\ v_2 \\ \vdots \end{bmatrix} = \cdots \begin{bmatrix} \vdots \\ h_0(t_0) \\ h_1(t_1) \\ h_2(t_2) \\ \vdots \end{bmatrix} u_0 + \begin{bmatrix} \vdots \\ h_{-1}(t_0) \\ h_0(t_1) \\ h_1(t_2) \\ \vdots \end{bmatrix} u_1 + \begin{bmatrix} \vdots \\ h_{-2}(t_0) \\ h_{-1}(t_1) \\ h_0(t_2) \\ \vdots \end{bmatrix} u_2 + \cdots \tag{C.8}$$

Appendix D

Simulation results of NLMS for different THD-values

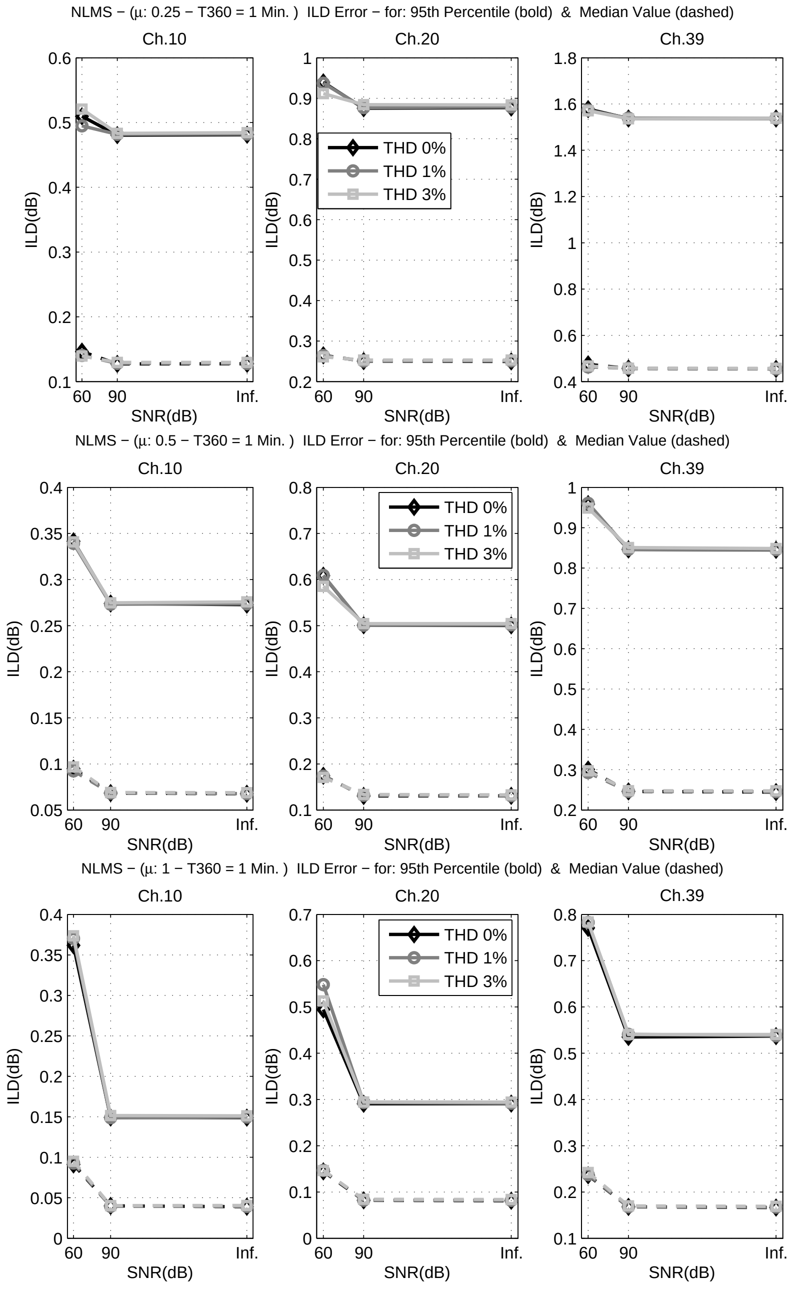


Figure D.1: NLMS-ILD-error for different THD values - Revolution time T_{360}=1 minute, for step size μ=0.25 (top), μ=0.5 (middle) and μ=1 (bottom)

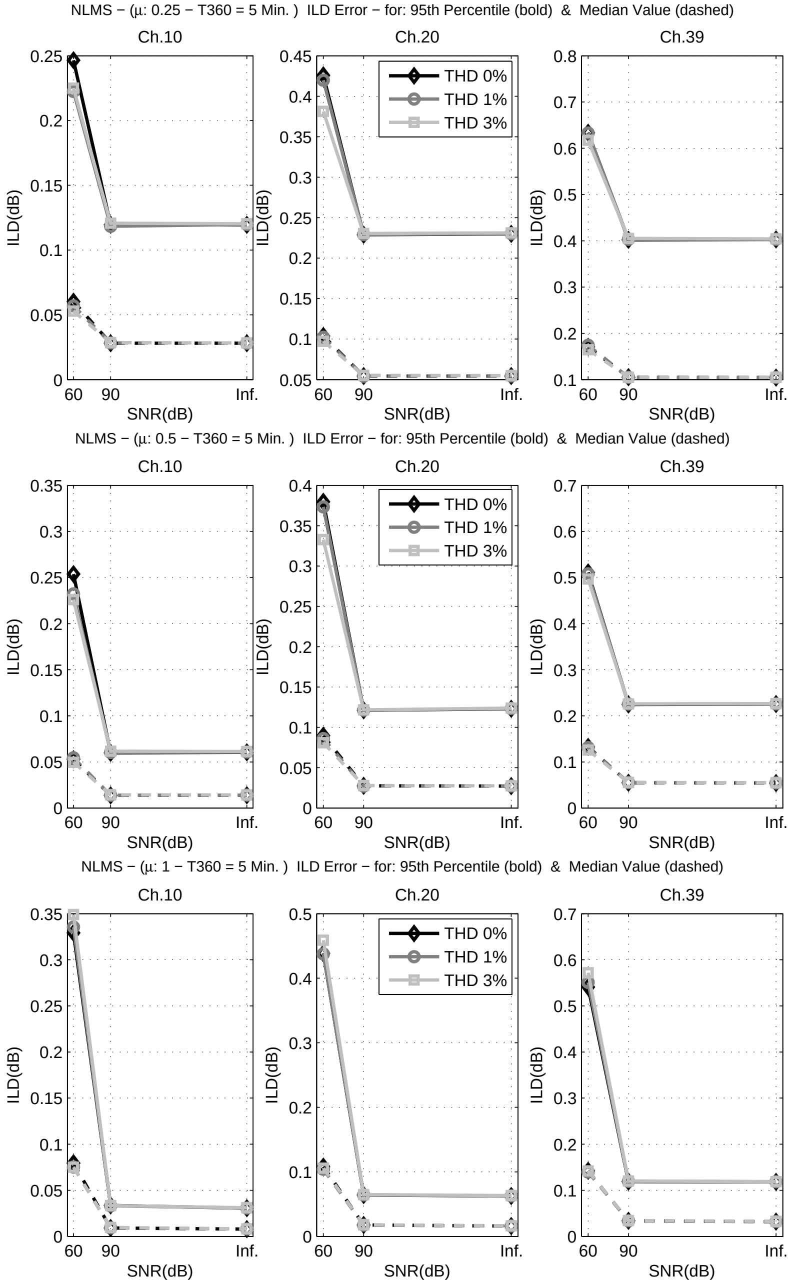


Figure D.2: NLMS-ILD-error for different THD values - Revolution time T_{360}=5 minute, for step size μ=0.25 (top), μ=0.5 (middle) and μ=1 (bottom)

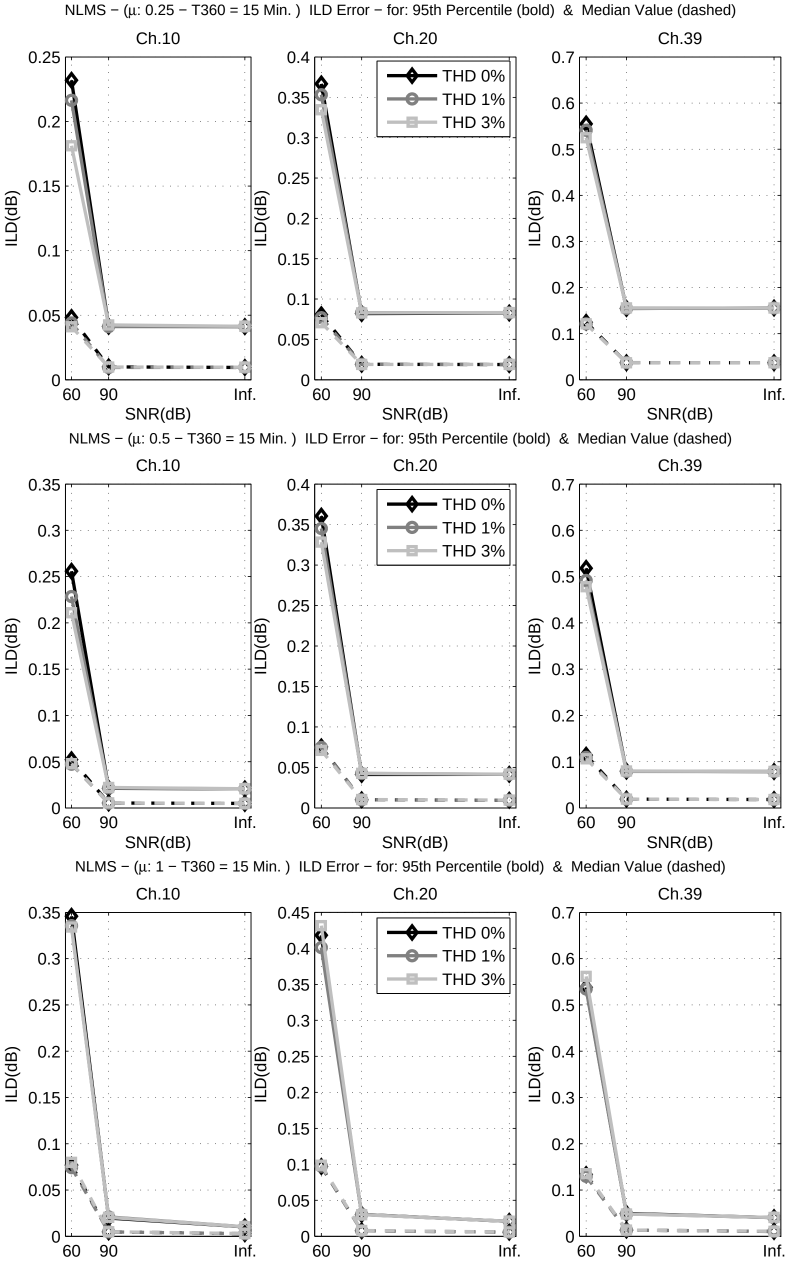


Figure D.3: NLMS-ILD-error for different THD values - Revolution time T_{360}=15 minute, for step size μ=0.25 (top), μ=0.5 (middle) and μ=1 (bottom)

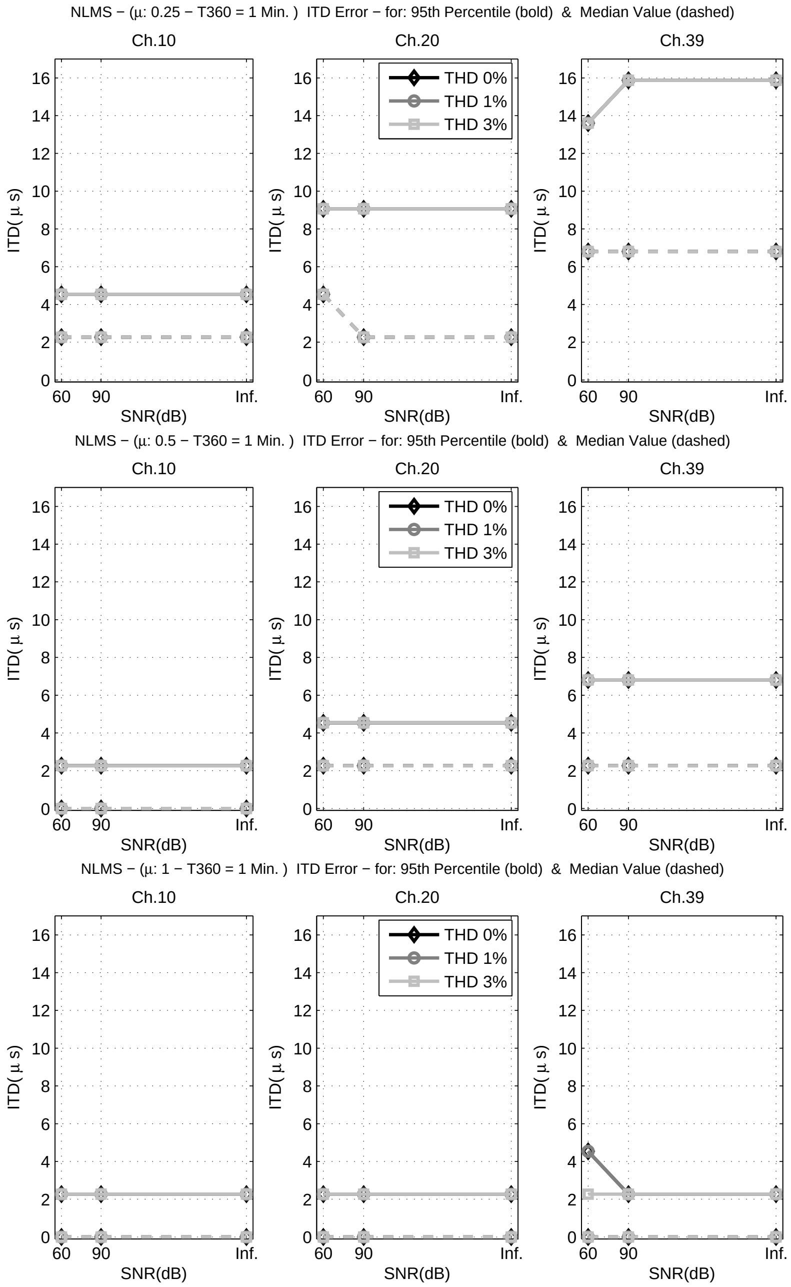


Figure D.4: NLMS-ITD-error for different THD values - Revolution time T_{360}=1 minute, for step size μ=0.25 (top), μ=0.5 (middle) and μ=1 (bottom)

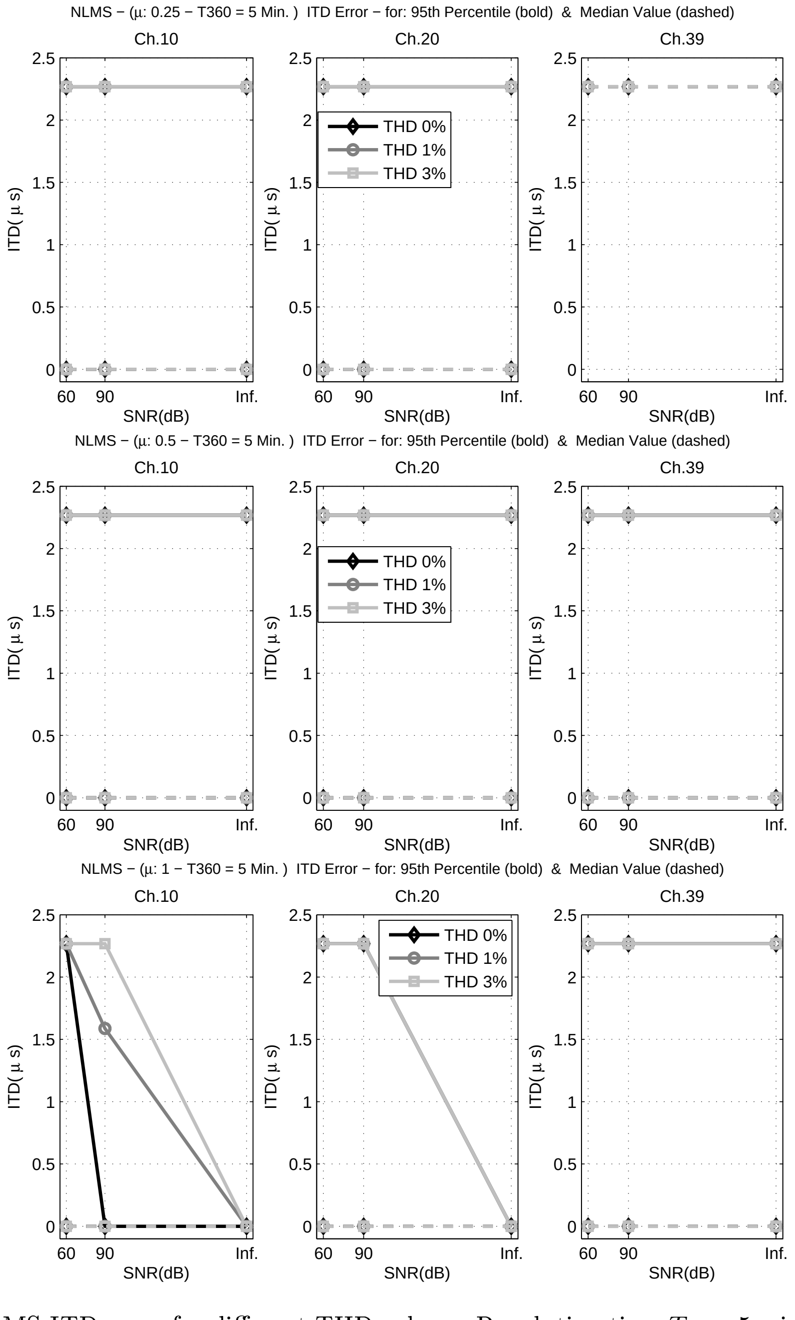


Figure D.5: NLMS-ITD-error for different THD values - Revolution time T_{360}=5 minute, for step size μ=0.25 (top), μ=0.5 (middle) and μ=1 (bottom)

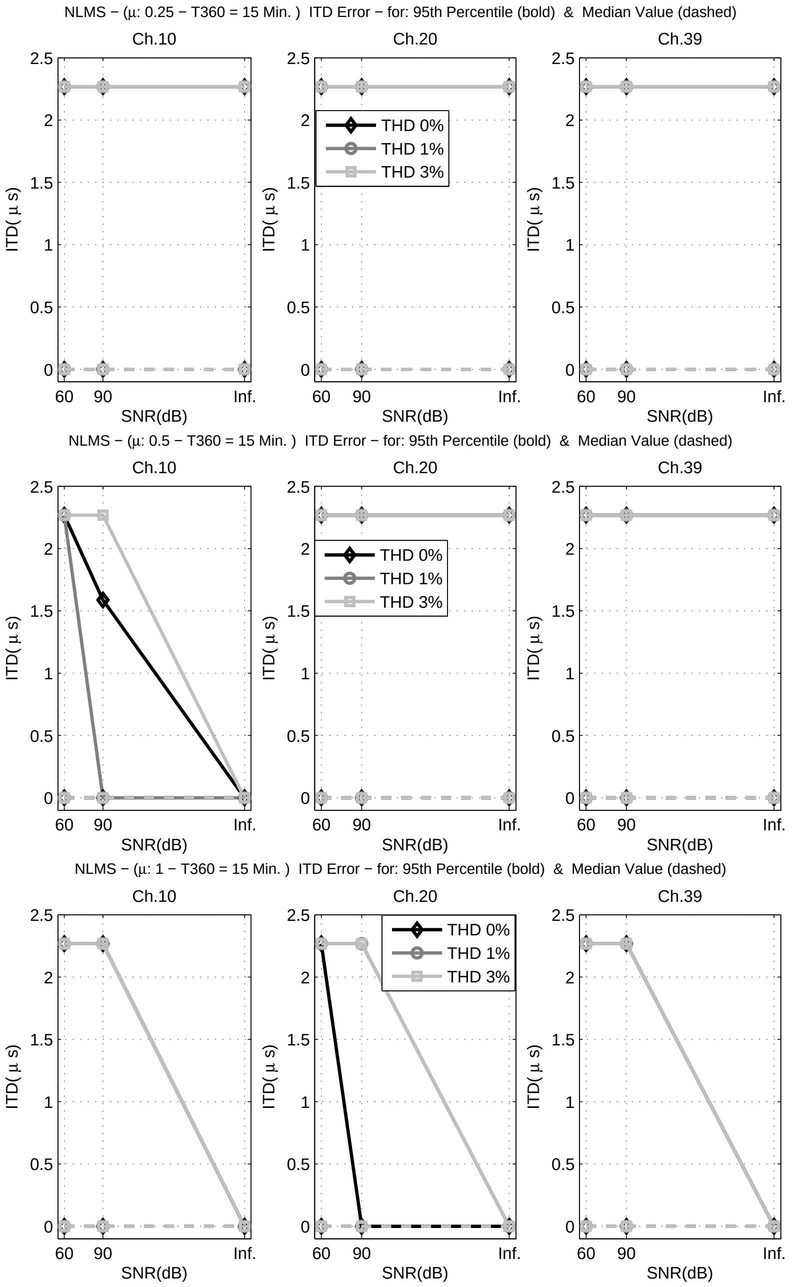


Figure D.6: NLMS-ITD-error for different THD values - Revolution time T_{360}=15 minute, for step size μ=0.25 (top), μ=0.5 (middle) and μ=1 (bottom)

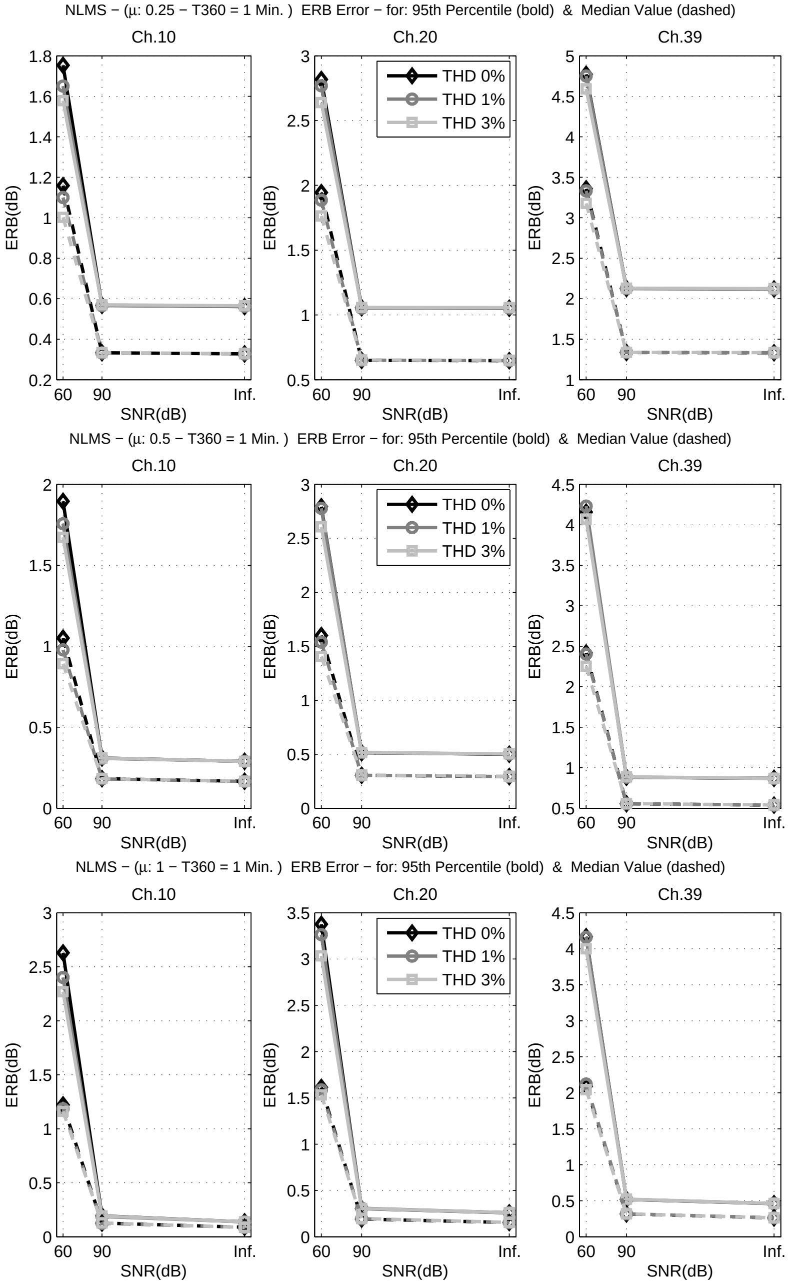


Figure D.7: NLMS-ERB-error for different THD values - Revolution time T_{360}=1 minute, for step size μ=0.25 (top), μ=0.5 (middle) and μ=1 (bottom)

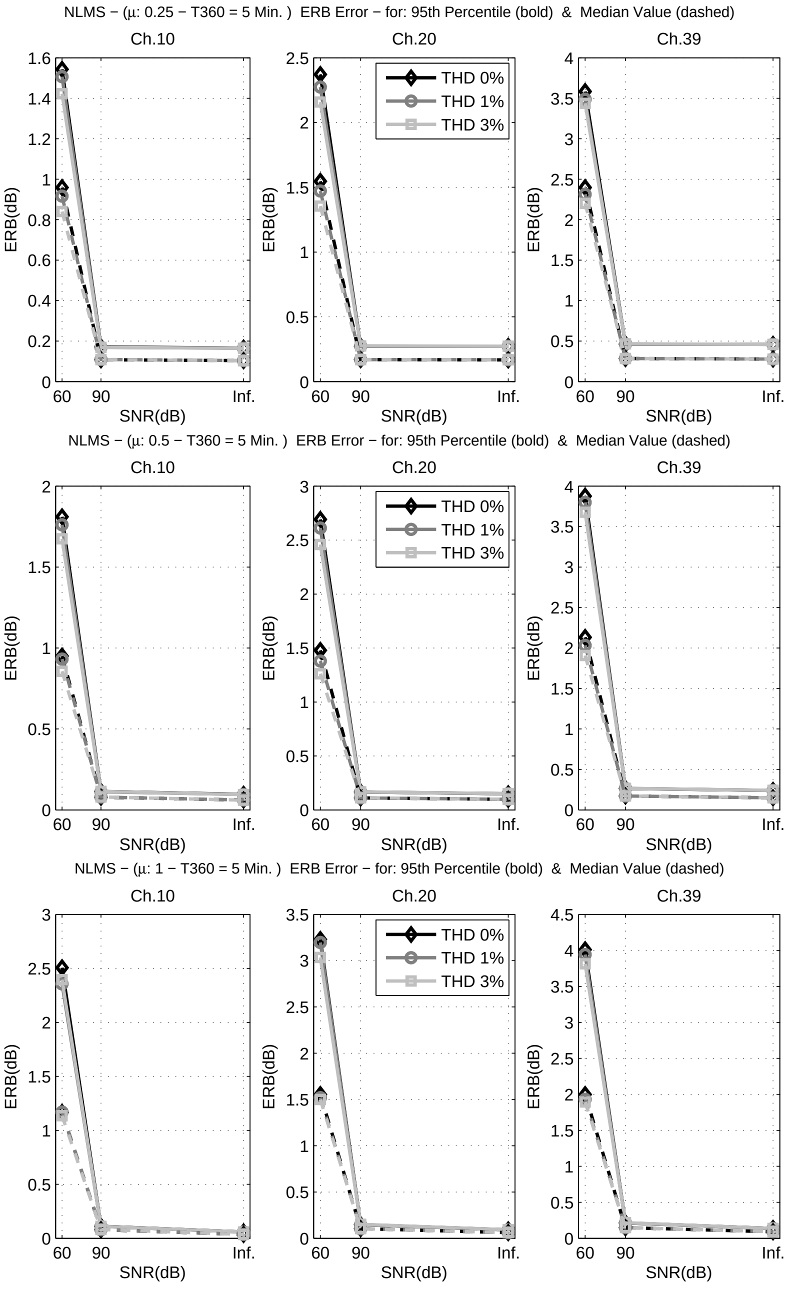


Figure D.8: NLMS-ERB-error for different THD values - Revolution time T_{360}=5 minute, for step size μ=0.25 (top), μ=0.5 (middle) and μ=1 (bottom)

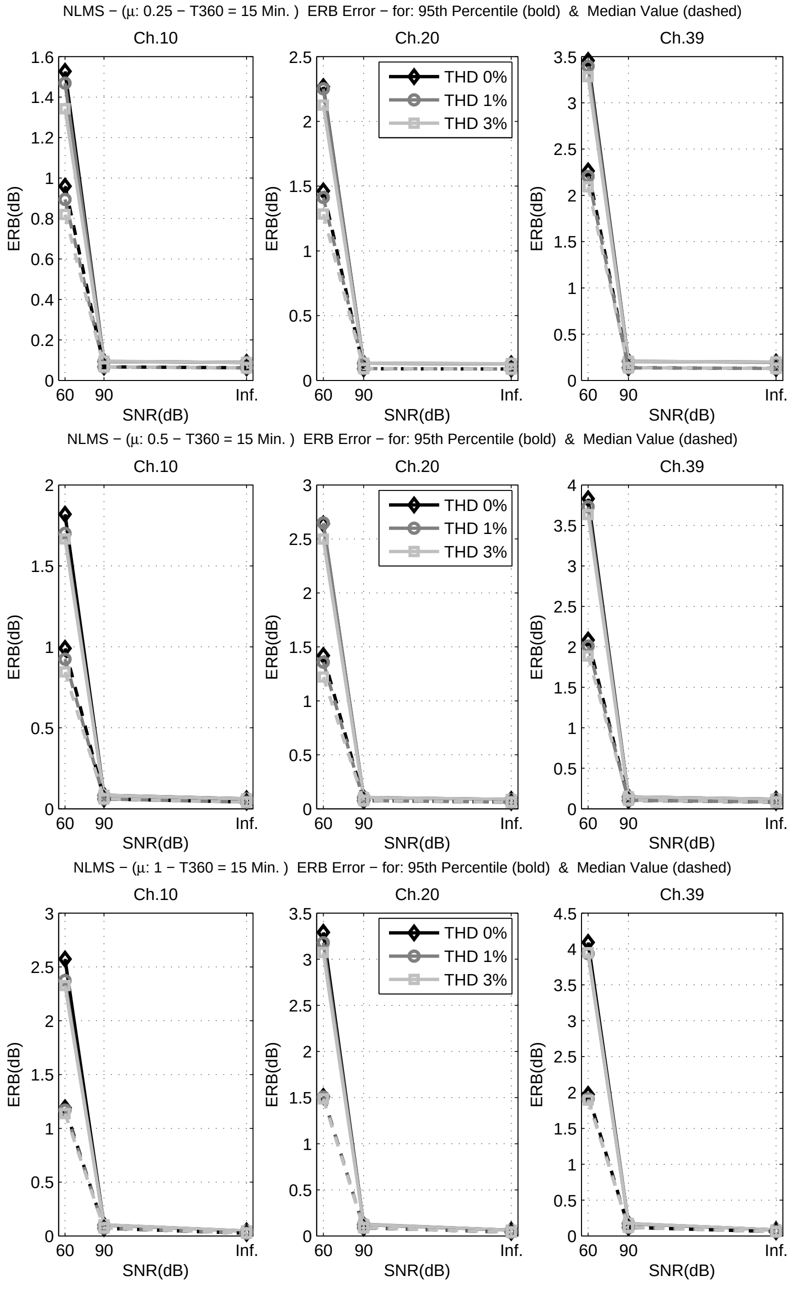


Figure D.9: NLMS-ERB-error for different THD values - Revolution time T_{360}=15 minute, for step size μ=0.25 (top), μ=0.5 (middle) and μ=1 (bottom)