B.Eng. Patrick Ziegler

Single-Channel Speech Dereverberation

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

Masterstudium Telematik

eingereicht an der

Technischen Universität Graz

Betreuer

Franz Zotter, Dr.rer.nat. DI.

Institut für elektronische Musik und Akustik

Graz, Oktober 2016

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Datum	Unterschrift

Danksagung

An dieser Stelle möchte ich mich bei all denjenigen bedanken, die mich während der Anfertigung dieser Masterarbeit unterstützt und motiviert haben. Allen voran bedanke ich mich bei meinen Eltern Claudia und Walter Ziegler, die mich sowohl geistig als auch finanziell stets auf meinem Lebensweg unterstützt haben und mir diese Arbeit ermöglichten. All meinen Freunden danke ich besonders für den starken emotionalen Rückhalt über die Dauer meines gesamten Studiums. Ein besonderer Dank gilt auch allen Teilnehmern und Teilnehmerinnen meines Hörversuchs, ohne die dieser Arbeit ein wesentlicher Teil fehlen würde. Mein Dank gilt ihrer Informationsbereitschaft und ihren interessanten Antworten auf meine Beispielaufnahmen. danke an das arbeitsumfeld

Dank an das Arbeitsumfeld

Zuerst gebührt mein Dank Herrn Dr. Franz Zotter, der meine Masterarbeit seitens des IEM (Institut für Elektronische Musik und Akustik) betreut und begutachtet hat. Für die hilfreichen Anregungen, Ideen und die konstruktive Kritik bei der Erstellung dieser Arbeit möchte ich mich herzlich bedanken. Die Kooperation mit der sonible GmbH hat maßgeblich zur Auswahl dieses Themas beigetragen und somit möchte ich mich auch bei den dortigen Arbeitskollegen für die Betreuung bedanken. Ihre fachliche und finanzielle Unterstützung ermöglichte diese Arbeit. Ein besonderer Dank gilt dabei Herrn Alexander Wankhammer, der stets professionell auf meine Fragen eingegangen ist.

Abstract

We often find audio recordings in which there is too much reverberation, but still the recording is important enough to require improvement and post processing to reduce the reverberation. There are already a few methods for so called dereverberation of audio recordings especially for speech signals with particularly varying dereverberation quality depending on different aspects of reverberation. All of them are based on the same fundamental blind deconvolution problem in which neither the original signal nor the impulse response of the system is known. This thesis discusses a method that combines different approaches based on homomorphic deconvolution and spectral subtraction defined by statistical and power spectral analysis. Optimal settings and algorithmic combinations are found by technical and perceptual evaluation of sound quality and reverberation level.

Zusammenfassung

In gängigen Tonaufnahmen können bedingt durch die Aufnahmesituation störende Anteile an Nachhall und frühen Reflexionen enthalten sein, die während der Aufnahme nicht immer vermeidbar sind. In diesen Aufnahmen ist es wünschenswert, die störenden Raumeinflüsse bestmöglich durch eine Audionachbearbeitung zu unterdrücken. Hierfür gibt es in der Literatur bereits verschiedene Ansätze, welche sich vor allem auf Sprachsignale spezialisiert haben. All diesen Ansätzen liegt das Problem zugrunde, dass weder das Originalsignal, noch die Impulsantwort des Aufnahmesystems bekannt sind. In dieser Arbeit wird die Kombination verschiedener Ansätze basierend auf homomorpher Entfaltung, statistischen Analysen und Nachhallschätzung untersucht und daraus ein kombinierter Algorithmus zur Nachhallkompensation entwickelt. Technische und perzeptive Messverfahren zur Bestimmung der Tonqualität und Nachhall dienen schließlich zur Optimierung des Algorithmus.

Contents

Chapter 1

Introduction

In many sound recording conditions such as interviews, lecture recordings, teleconferencing systems, or live recordings, the recorded signals are often corrupted by reverberation. Depending on the physical parameters of a recording room, such as size and absorption coefficients, reverberation occurs due to multi-path propagation of the sound source. Not only the clean speech signal is recorded over a microphone, but also source parts reflected by the walls which cause the reverberation. The main perceptual effects of reverberation are the coloring of the source signal and an echo. In most cases these effects are disturbing and unwanted. It is desirable to remove these reverberating parts from the recording in a post-processing step. Due to the time dependent variation of the relative strength of reverberation compared to the desired signal, it is impossible to suppress such effects by manually parameterizing equalizers. Thus, it is necessary to develop an algorithm that is able to suppress or cancel out reverberation. The suppression or cancellation of reverberation is called dereverberation. The following chapters introduce an overview of current dereverberation techniques and describe the proposed algorithm in detail.

1.1 System Description

The effect of reverberation is caused by a superposition of a direct sound wave and its acoustic reflections in the room (see Figure 1.1). The time needed by the direct sound wave to travel from the speaker to the receiver is defined by t_D, which corresponds to the path depicted as the arrow in the middle of Figure 1.1. On the other hand sound waves which take the reflection path over the walls need more time (i.e. $t_R > t_D$) until they reach the microphone because of traveling trough the longer path over the walls. In theory, there is an infinite set of sound propagation paths of different lengths and directions.

In terms of mathematical description a multi-path propagation with a room impulse response (RIR) can be modeled so, that a reverberated signal $x(t)$ can be seen as a convolution of the clean source signal $s(t)$ and the RIR $h(t)$ plus noise $n(t)$.

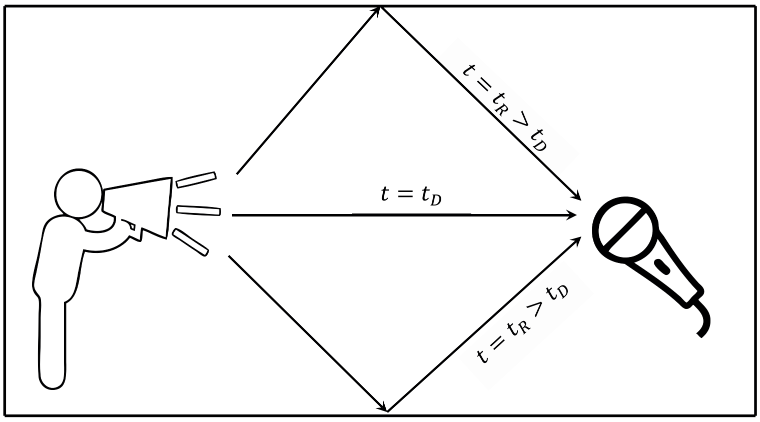


Figure 1.1 – Multi-path propagation of the source sound wave (represented by a speaker) traveling to the receiving microphone.

$$x(t) = s(t) * h(t) + n(t) \tag{1.1}$$

The RIR is divided into a time period where early reflections occur and another period where late reverberation takes place. The late reverberation is a diffuse sound field with exponentially decaying sound energy as shown in Figure 1.2.

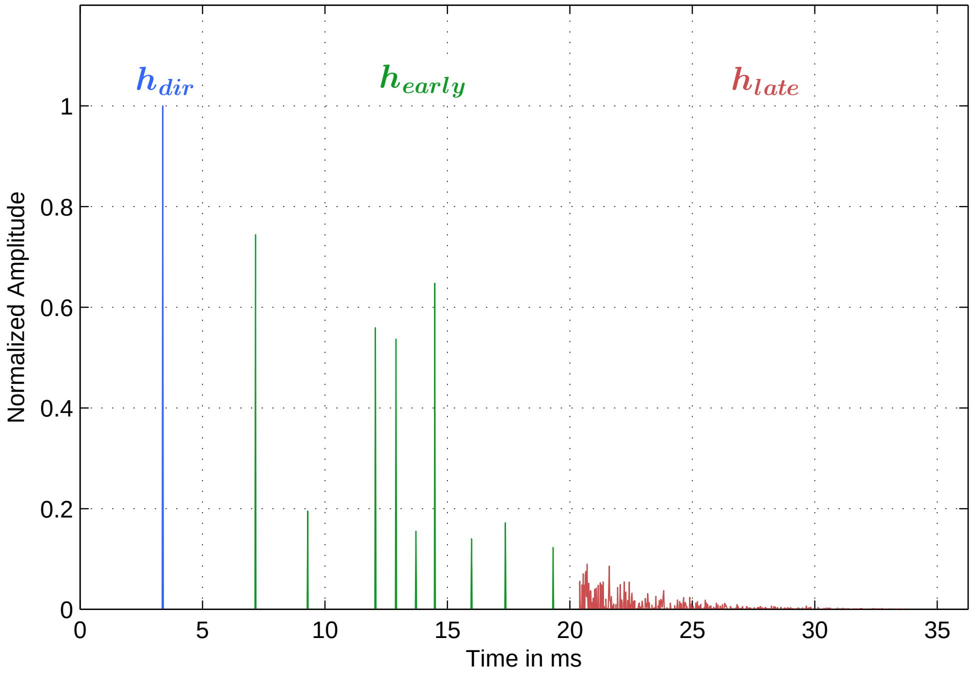


Figure 1.2 – General RIR representation with direct part h_{dir}, early reflections h_{early} and late reverberation h_{late}.

Thus, if we have complete knowledge of the room impulse response of the system at every time instance t, the source signal $s(t)$ might be fully reconstructed. However, in most cases neither the RIR nor the clean speech signal are known. Only the reverberated signal $x(t)$ is available. In general this circumstance is called the blind deconvolution problem. Possible ways to calculate the original speech signal is to estimate the RIR or to define spectral gain functions that make assumptions on the statistical behavior of reverberation. An overview of such techniques is outlined in the next section.

1.2 Overview of Dereverberation Techniques

There are already many works that present different dereverberation techniques. However, it is often very hard to find an appropriate overview due to the discrepancy between various approaches. It is not meaningful to describe all of the methods in detail, although there are some clearly defined differences between some techniques. One way to classify dereverberation algorithms is to split them into reverberation cancellation and reverberation suppression.

Reverberation Cancellation If the RIR is known, the clean speech signal can be fully reconstructed under certain conditions. Algorithms of this kind try to completely cancel out the influence of the reverberation by directly calculating the room impulse response. As outlined in the previous section, in most cases the RIR is unknown and has to be approximated with assistance of existing speech models. Reverberation cancellation methods using homomorphic deconvolution are described in [1], [2] and [3]. Also harmonicity based dereverberation (HERB) [4], [5] and [6] is used to cancel out reverberation. Inverse filtering by Bussgang Method [7] could also be treated as a cancellation technique. One of the problems of such methods is that due to the calculation of the amplitude and phase of the RIR, these algorithms are sensitive and prone to phase errors even at low levels of dereverberation.

Reverberation Suppression On the other hand there are methods that use statistical models and the general characteristics of speech signals. The levels of the reverberant parts of the signal are estimated and suppressed subsequently by using a spectral gain function or an appropriate filter instead of directly calculating an impulse response. Although it is not possible to completely cancel out the reverberation of a signal using these algorithms, they are more robust to pre-ringing and annoying *musical noise*. Furthermore, inaccuracies of human hearing such as frequency and time resolution could be used as a motivation of suppression and to regard reverberation cancellation as being unnecessary. In reverberation suppression, algorithms based on spectral subtraction are common [8], [9], and [10] which suppress reverberation in the short-term spectral domain. Therefore, estimating reverberant parts is done by using a simple reverberation model [11] or higher order statistics [12].

Another important classification of such algorithms is defined by the number of channels. We can distinguish between single-channel and multi-channel applications.

Single-Channel Dereverberation As the name suggests, only one channel of a reverberant signal is assumed. This is a comprehensive method because it uses the properties of only one reverberated signal to calculate the reverberant parts and works therefore also for multi-channel applications. Hence, it is interesting to find an appropriate Single-Channel dereverberation algorithm. On the other hand it is comparably difficult to estimate an impulse response or reverberation from a mono signal as opposed to multi-channel signals. Such Methods are used for example in [13].

Multi-Channel Dereverberation In multi-channel systems such as stereo or ambisonic, most of the dereverberation techniques employ cross-correlation between channels to extract more information about the signal and thus more information about reverberant parts or the RIR. Hereby, the amount of available information increases by an increasing number of channels and therefore multi-channel dereverberation in general leads to better results. In ambisonic systems, beamforming [10] is an appropriate technique to determine the direction of the source signal increasing the ratio between direct sound and reverberation. The main problem of such methods is computational complexity and the fact that most of the recordings are still single-channel or stereo recordings.

The last clearly defined classification of dereverberation algorithms can be done by defining in which time span of the room response the dereverberation becomes effective. There are methods that are only focusing on suppressing late reverberation and such limited to early reflections. Independently of the target segment of the room response, there are algorithms which employ reverberation suppression and algorithms which employ reverberation cancellation. The usage of multi-channel or single-channel is not bounded to a particular segment of an impulse response. Still, a well-defined goal for which part of reverberation the algorithm shall become effective can be a decisive design question.

Figure 1.3 shows a possible classification of blind dereverberation techniques. Its left half presents reverberation suppression algorithms that try to suppress estimated reverberant parts of speech. The right half of the figure shows algorithms that directly estimate the room impulse response. Spatial processing can apply either suppression or cancellation methods after employing e.g. a beamforming algorithm. Homomorphic deconvolution can also be achieved by suppression or cancellation techniques as shown in Chapter 2. The next Section describes which kind of algorithms are discussed in this work.

1.3 Proposed Dereverberation Algorithm

Most of the dereverberation algorithms are focusing on suppressing late reverberation. Just a few approaches try to suppress early reflections. However, in many recordings early reflections have a large influence on coloring the sound and ambient effects. Therefore, we focus on suppressing early reflections in this work. Additionally late reverberation suppression is also considered for evaluation.

The target group for the proposed algorithm are in general speech signals from i.e. radio or documentary recordings. These recordings are often recorded by one microphone. We are not able to assume a multi-channel setting for every recording. Therefore, it is suitable to develop an algorithm with single-channel dereverberation functionality. This also works for multi-channel systems and can be extended and improved by i.e. a beamformer afterwards.

In Chapter 2 two dereverberation approaches that employ cepstrum based techniques to estimate reverberation are outlined. In Chapter 3, a reverberation suppression method based on spectral subtraction is presented. The algorithms are reviewed and the most suitable algorithm will be tested in Chapter 4 for evaluation.

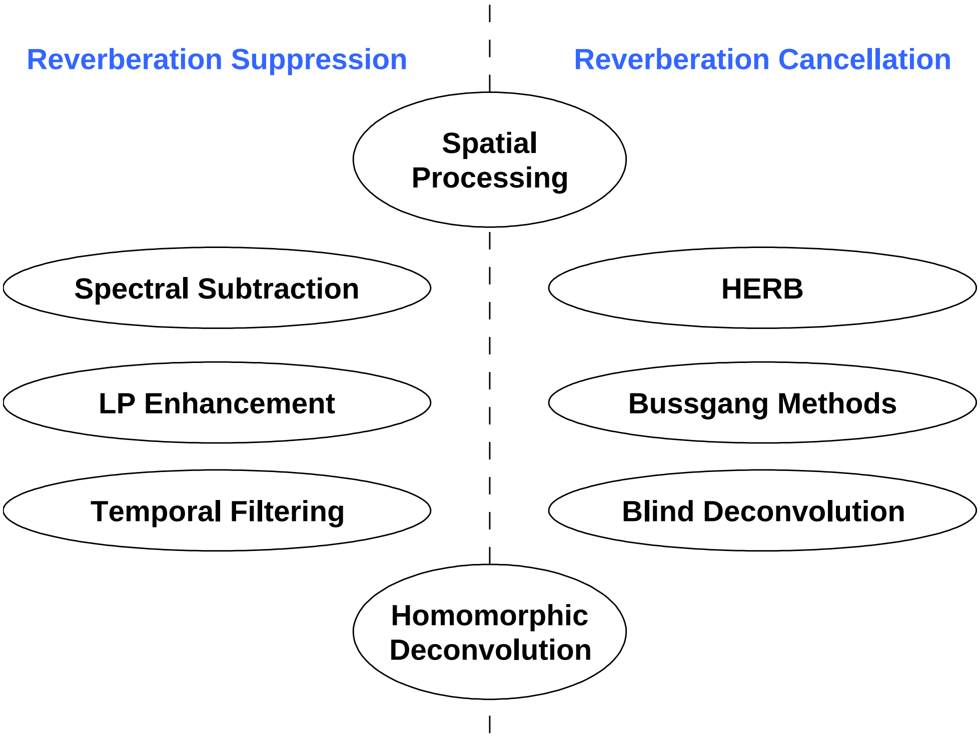


Figure 1.3 – Overview of dereverberation techniques

Chapter 2

Homomorphic Deconvolution

Homomorphic deconvolution transforms multiplicative components into additive components and uses linear filtering techniques in the log-spectral domain for the deconvolution procedure. It is used to separate the source signal from the room impulse response. In the log-spectral domain, the real cepstrum and the complex cepstrum are distinguished. While the real cepstrum is calculated only by the magnitude spectrum, the complex cepstrum also includes phase information but comes along with a certain computational complexity. This Chapter outlines two well-known homomorphic deconvolution approaches using the cepstrum technique for speech dereverberation.

2.1 Real-Cepstrum Based Dereverberation

In theory speech can be described as a combination of source excitation signal (i.e. an impulse for voiced or white noise for unvoiced sounds) and an anechoic vocal tract filter. While the vocal tract filter represents the formant frequencies and produces a spectral envelope, the excitation signal exhibits quasi-periodic harmonic ripples which represent the fundamental frequency. This means, that the vocal tract filter corresponds to "slowly changing" and the excitation signal to "rapidly changing" spectral values. The real-cepstrum is able to separate these values by an inverse Fourier transform of the log magnitude spectrum. It is given by

$$\hat{c}_x(t) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln[|X(\omega)|] e^{j\omega t} d\omega, \tag{2.1}$$

where $X(\omega)$ is the spectrum of a signal $x(t)$ with frequency ω. In terms of discrete signal processing the spectrum $X(k)$ of a discrete signal $x(n)$ is defined by

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-2\pi j \frac{k \cdot n}{f_S \cdot N}} \tag{2.2}$$

with frequency bin index k, N as signal length and f_s as sampling frequency. Thus, we can also write down the discrete real-cepstrum $\hat{c_x}(q)$ of $x(n)$ with quefrency bin index q by

$$\hat{c_x}(q) = \text{iDFT}\left[\ln[|X(k)|]\right]. \tag{2.3}$$

For example, Figure 2.1 shows the real-cepstrum of a female speech signal. The cepstral coefficients located around the origin correspond to the vocal tract filter and the formant frequencies. The peak in the higher quefrency range represents the fundamental frequency f_0. It is therefore possible to detect speech components by analyzing these coefficients and the cepstral peak.

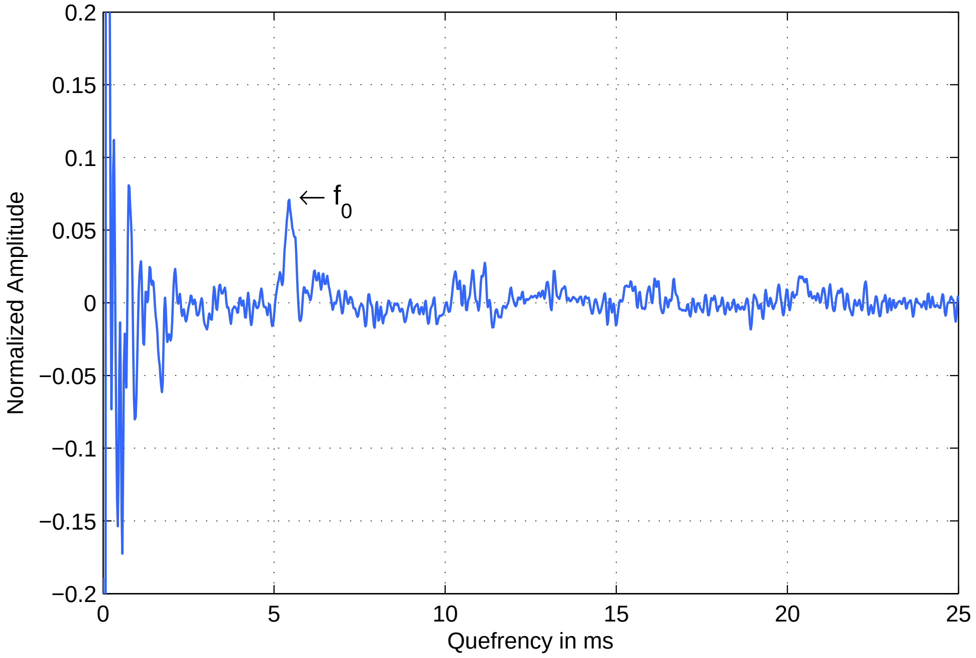


Figure 2.1 – Real-cepstrum of simulated female speaker saying "aaa"

2.1.1 Cepstral Liftering

Assuming that a reverberated signal is given, the real-cepstrum is corrupted by reverberation. However, there are regions in the real-cepstrum where speech is not present but reverberation can be. The idea of the real-cepstrum based dereverberation is to filter out certain regions in the cepstral domain. In terms of the cepstrum this is called liftering. A simple way to do this is to extract the coefficients located around the origin and to suppress the rest of the cepstrum except the range around the peak that represents f_0. An appropriate lifter can be defined as a real-valued quefrency-depended gain function $G(q)$ and is used as a simple multiplication as

$$\hat{c_y}(q) = \hat{c_x}(q) \cdot G(q). \tag{2.4}$$

An example for a gain function $G(q)$ as lifter is shown in Figure 2.2.

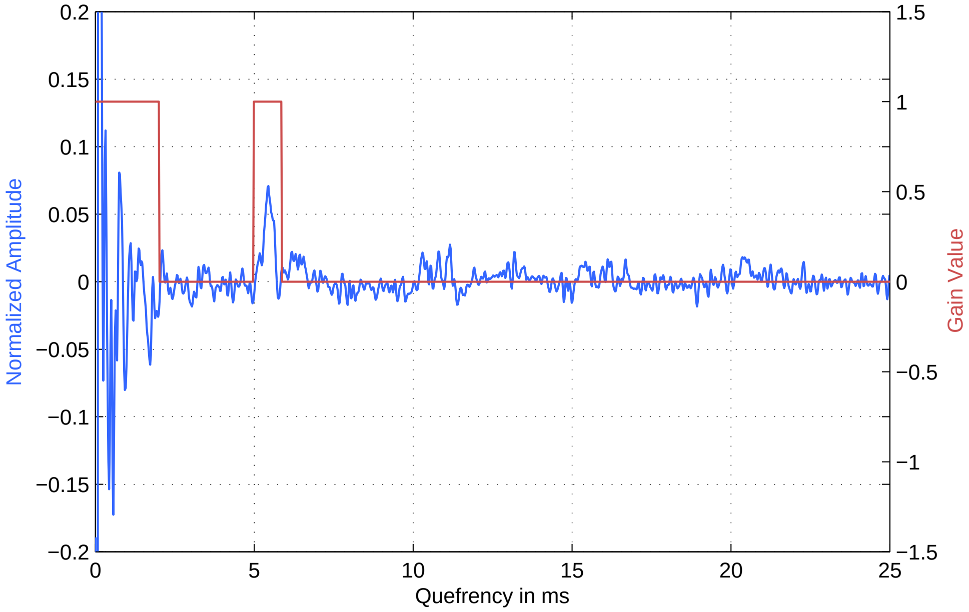


Figure 2.2 – Real-cepstrum of simulated female speaker saying "aaa" with lifter gain function $G(q)$ at a sampling frequency of $f_s = 44.1kHz$

The peaks between the origin and the f_0 range rely on the effects of reverberation that are suppressed by the cepstral lifter and $\hat{c}_y(q)$ can be used as an initial point of dereverberation.

2.1.2 Signal Reconstruction

After cepstral liftering, the reconstruction of a dereverberated signal $y(t)$ is the last mandatory task. Since the computation of the real-cepstrum according to Equation (2.3) does not include phase information, the original phase of $X(k)$ for the spectrum $Y(k)$ has to be employed to reconstruct the signal. Therefore, the spectrum of the dereverberated signal $y(t)$ is defined by

$$Y(k) = |Y(k)| \cdot e^{j\arg[X(k)]}, \tag{2.5}$$

where the magnitude spectrum $|Y(k)|$ is calculated by the inverse real cepstrum of $\hat{c}_y(q)$ through

$$|Y(k)| = e^{\mathrm{DFT}[\hat{c}_y(q)]}. \tag{2.6}$$

An inverse Fourier transformation leads then to the dereverberated signal

$$
\begin{aligned}
y(n) &= \mathrm{iDFT}\left[e^{\mathrm{DFT}[\hat{c}_y(q)]} \cdot e^{j\mathrm{arg}[X(k)]}\right] &\quad (2.7)\\
&= \mathrm{iDFT}\left[e^{\mathrm{DFT}[\hat{c}_y(q)]+j\mathrm{arg}[X(k)]}\right]. &\quad (2.8)
\end{aligned}
$$

As already shown, the original phase has to be used for signal reconstruction. Since the phase is not taken into account and unaffected by the algorithm, the effect of this simple lifter is low. In particular, if more complex impulse responses are considered, the reverberation also influences the cepstral coefficients around the origin. Because of that a more accurate and detailed model to achieve better reverberation suppression is needed. The next Section discusses a dereverberation technique based on the complex cepstrum.

2.2 Complex-Cepstrum Based Dereverberation

In the time domain, a reverberated speech signal can be described as the convolution of a clean speech signal and an impulse response assuming insignificantly small noise. This convolution converts into a multiplication in the spectral domain. Further, if the complex cepstrum of a reverberated spectrum is calculated, the multiplication turns into a summation in the cepstral domain due to the logarithm in the spectral domain. The complex cepstrum $\hat{x}(t)$ of a signal $x(t)$ can be calculated as

$$
\hat{x}(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln[X(\omega)]e^{j\omega t}d\omega, \tag{2.9}
$$

and in terms of discrete signal processing

$$
\hat{x}(q) = \mathrm{iDFT}\left[\ln[X(k)]\right]. \tag{2.10}
$$

If the complex logarithm $\ln[X(k)]$ of the spectrum $X(k)$ defined by the multiplication of $S(k)$ and $H(k)$ is considered, the multiplication turns into a summation.

$$
\begin{aligned}
\ln(\mathrm{DFT}[x(n)]) &= \ln[X(k)] &\quad (2.11)\\
\ln[X(k)] &= \ln[S(k)H(k)] &\quad (2.12)\\
&= \ln[S(k)] + \ln[H(k)]. &\quad (2.13)
\end{aligned}
$$

After computing the complex cepstrum $\hat{s}(q)$ of a clean speech spectrum $S(k)$ and the complex cepstrum $\hat{h}(q)$ of the RIR spectrum $H(k)$ with Equation (2.10), the reverberated speech signal $x(n)$ can be written in the cepstral domain:

$$
\begin{aligned}
\hat{x}(q) &= \mathrm{iDFT}\left[\ln[S(k)]\right] + \mathrm{iDFT}\left[\ln[H(k)]\right] &\quad (2.14)\\
&= \hat{s}(q) + \hat{h}(q). &\quad (2.15)
\end{aligned}
$$

Now by complete knowledge of the complex cepstrum of the impulse response, the original clean speech complex cepstrum $\hat{s}(q)$ can be reconstructed by simply subtracting $\hat{h}(q)$ from $\hat{x}(q)$. After transforming the complex cepstrum of $\hat{x}(q)$ back into the time domain by computing the inverse complex cepstrum defined by

$$s(n) = \text{iDFT}\left[e^{\text{DFT}[\hat{s}(q)]}\right], \tag{2.16}$$

the original clean speech signal $s(n)$ is reconstructed. But as shown in the previous chapter, a room impulse response $h(n)$ is not known in general, so $h(n)$ has to be approximated using cepstrum-based estimation techniques.

2.2.1 RIR Representation in the Cepstral Domain

As outlined in the previous Chapter, we are mainly interested in early reflections of the room impulse response. Early reflections are represented by delayed peaks in the time domain. The complex cepstrum of an impulse response also shows early reflections in the form of peaks at the corresponding quefrency. For Example Figure 2.3 shows a simple impulse response $h(n)$ with direct sound impulse $\delta(n)$ at time sample $n = 0$ and one simple reflection impulse after $n = n_0$ samples and the corresponding complex cepstrum. The alternating peaks in the complex cepstrum naturally arise from the computation due to power series expansion of the logarithm

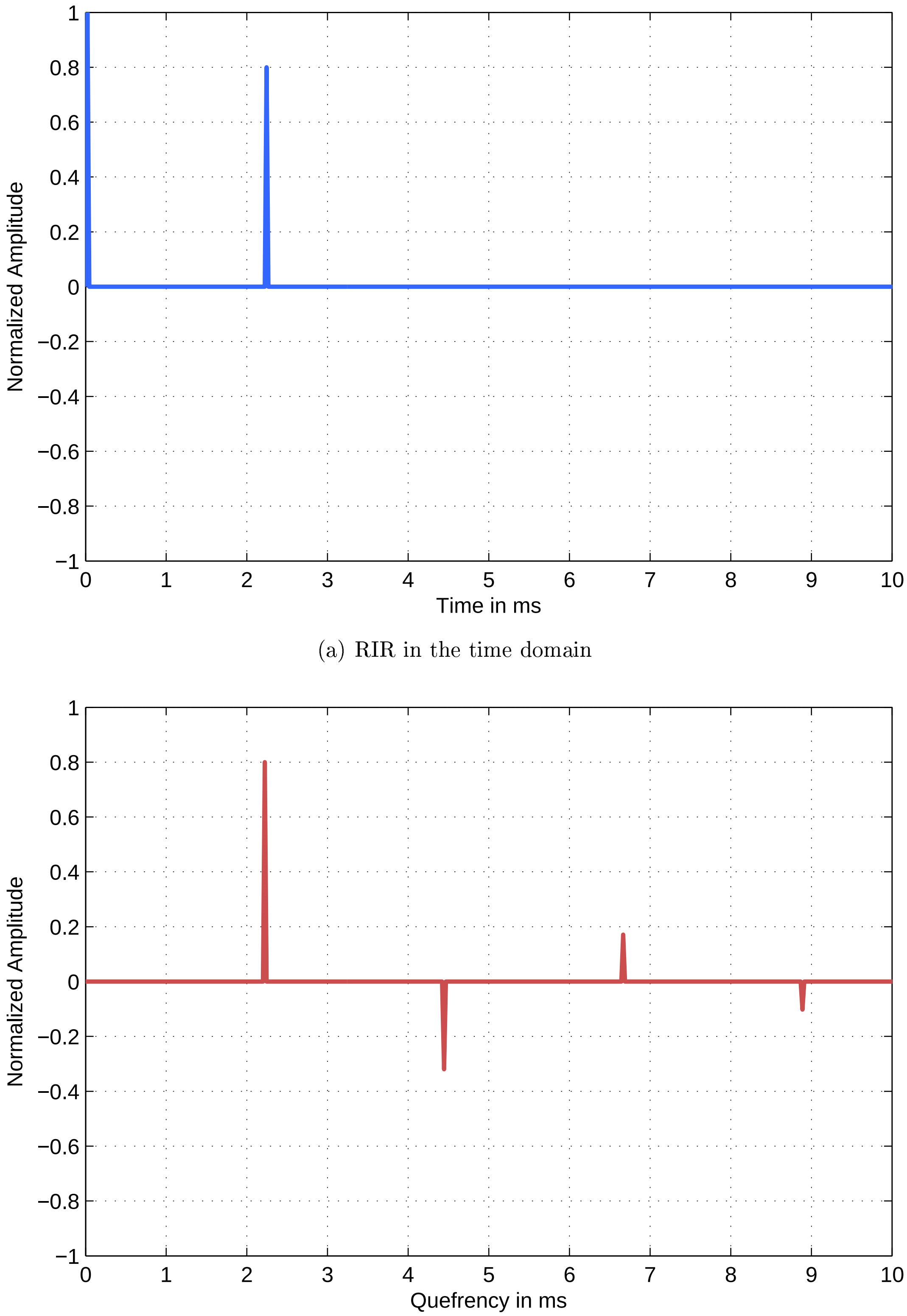
$$\begin{aligned}
h(n) &= \delta(n) + a\delta(n - n_0) && (2.17)\\
H(k) &= 1 + ae^{-jkn_0} && (2.18)\\
\hat{H}(k) &= \ln[1 + ae^{-jkn_0}] && (2.19)\\
&= ae^{-jkn_0} - \frac{a^2}{2}e^{-2jkn_0} + \frac{a^3}{3}e^{-3jkn_0} - ... && (2.20)\\
\hat{h}(q) &= a\delta(n - n_0) - \frac{a^2}{2}\delta(n - 2n_0) + \frac{a^3}{3}\delta(n - 3n_0) - ... && (2.21)
\end{aligned}$$

If the complex cepstrum of a signal convolved with that impulse response is computed, the peaks arise at n_0 because of the summation described in Equation (2.15). As reflections occur mainly in the form of peaks in the cepstral domain and interfered additively with the complex cepstrum of a signal, the subtraction of that peaks would lead to the complex cepstrum of the dereverberated speech $\hat{y}(q)$ with

$$\hat{y}(q) \approx \hat{x}(q) - \mathbb{H}_p(q), \tag{2.22}$$

where $\mathbb{H}_p(q)$ is a set of cepstral peaks detected by a simple peak picking algorithm with threshold $\Lambda(q)$ defined by

$$\mathbb{H}_p(q) = \begin{cases} \hat{x}(q), & \text{if } \hat{x}(q) > \Lambda(q) \\ 0, & \text{otherwise} \end{cases}. \tag{2.23}$$



(a) RIR in the time domain

(b) Complex Cepstrum of RIR

Figure 2.3 – Impulse response and corresponding complex cepstrum of a simple comb filter with $n = 100 \mathrel{\hat{=}} 2.2$ ms @44.1 kHz

Figure 2.4 illustrates the complex cepstrum of a short speech segment $x(n)$ convolved with the simple RIR $h(n)$ from above. The red marked cepstral peaks show the additivity between $\hat{h}(q)$ and $\hat{x}(q)$ and mainly represent the RIR. The peaks are clearly detectable and can be subtracted with Equation 2.22. An inverse transformation of the peak subtracted cepstrum leads to a reconstructed, dereverberated signal, but a perfect reconstruction is not possible because the speech fraction of such a complex cepstrum is not known. For a single quefrency bin q_p where a peak was found, the instance of $\mathbb{H}_p(q_p)$ is defined as

$$\mathbb{H}_p(q_p) = \hat{s}(q_p) + \hat{h}(q_p). \tag{2.24}$$

It is therefore not detectable how strong the influence of $\hat{s}(q_p)$ is within $\hat{h}(q_p)$ since we do not know the complex cepstrum of $\hat{s}(q_p)$ and $\hat{h}(q_p)$. Hence, if a more complex RIR is assumed, it is often difficult to find such clearly pronounced peaks in the complex cepstrum due to a insignificantly low amplitude or the superposition with speech in the cepstral domain.

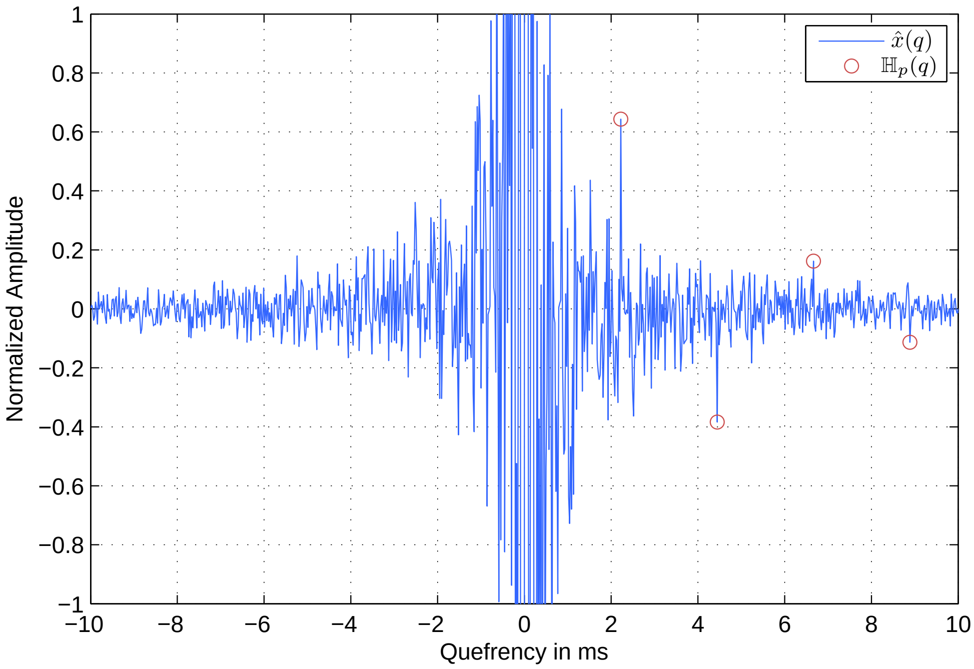
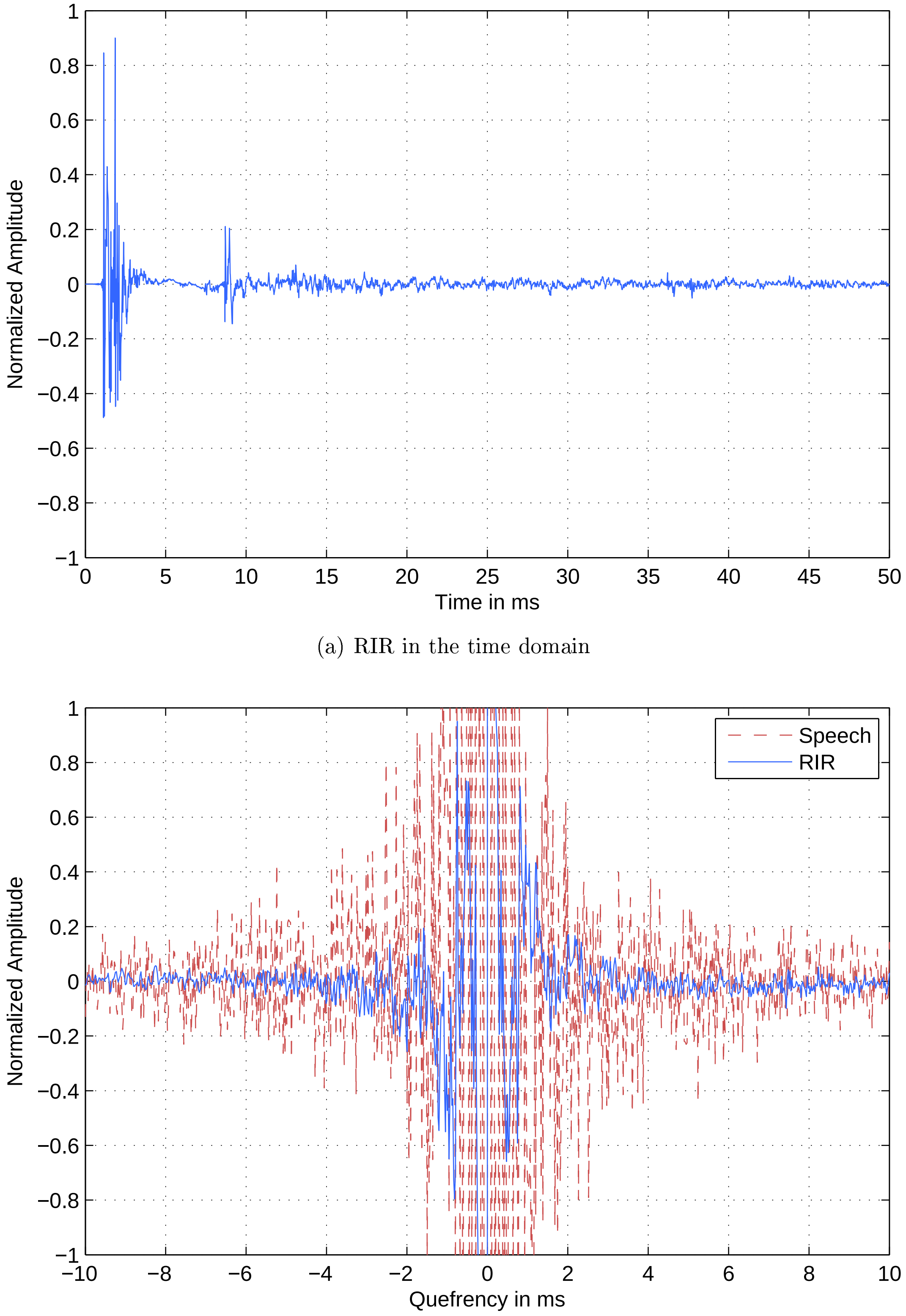


Figure 2.4 – Complex cepstrum of a single speech segment $\hat{x}(q)$ convolved with a simple RIR and corresponding cepstral peaks $\mathbb{H}_p(q)$

Further, peaks in the complex cepstrum can also correspond to the speech itself especially in the range around $q = 0$ as shown in Figure 2.5. These problems make it impossible to detect the RIR. However, in terms of a frame-based signal processing, statistical independences of speech can be used to estimate the RIR which takes us a step toward to the so called cepstral mean subtraction.



(a) RIR in the time domain

(b) Complex Cepstrum of RIR and speech

Figure 2.5 – Meeting room [14] RIR and the corresponding complex cepstrum and the complex cepstrum of a single speech segment

2.2.2 Cepstral Mean Subtraction

By frame-based signal processing, a short time Fourier transformed segment of reverberated speech $X(k, l)$ is defined, where k is the frequency bin index and l the frame index. By computing the complex cepstrum of $X(k, l)$ using Equation (2.10), $\hat{x}(q, l)$ is obtained. As speech is a non-stationary stochastic process, it is assumed that values in the frequency and vice-versa in the quefrency domain of the signal change rapidly in time and may be uncorrelated to other frames in a certain range. On the other hand a RIR is relatively time-invariant, if the speaker and the microphone are fixed or just slowly changing in time (e.g. by a moving speaker or microphone in the room). Thus, the RIR frames are highly correlated to others.

By computing the mean of several reverberated signal frames, the low correlation of speech will approximately average to a zero mean while the highly correlated RIR has its peaks in the cepstrum at the same quefrency bins. Under these circumstances, the following assumption for the cepstral mean $\hat{h}(q, \tau)$ for a time instance τ can be made:

$$\hat{h}(q, \tau) = \frac{1}{L} \cdot \sum_{l=\tau}^{\tau+L-1} \hat{x}(q, l), \tag{2.25}$$

where L is the number of frames used for the cepstral mean. The higher L, the higher is the probability that the cepstral mean of a reverberated signal converges to the complex cepstrum of the RIR. But under the assumption of a time-varying RIR also the cepstral coefficients of the RIR are uncorrelated and undergo the risk of a zero mean. The goal is to find an appropriate number of frames which yield to the best approximation of the RIR complex cepstrum whereas

$$\hat{h}(q, \tau) \approx \hat{h}(q, l) \tag{2.26}$$

holds. According to the above Equation, an estimation of the impulse response in the cepstral domain is given and the peak picking technique described in Section 2.2.1 can be used for frame-wise subtraction of the cepstral peaks with Equation (2.22) caused by the RIR. This leads to an estimation of the complex cepstrum of the dereverberated speech $\hat{y}(q, l)$. By inverse transformation of $\hat{y}(q, l)$ back into the time domain with Equation (2.16), the dereverberated speech signal can be reconstructed.

2.2.3 Phase Unwrapping

For the complex cepstrum and frame-based processing, there are a few problems that have to be considered. First, any computation of the complex cepstrum requires an efficient phase unwrapping algorithm for a correct computation of the phase information within the complex cepstrum [15]. The phase of a signal spectrum might jump on the bounds of π and $-\pi$. By computing the complex cepstrum without phase unwrapping, the data is corrupted by these jumps in the spectral domain. While exact phase unwrapping algorithms are of high computational cost, efficient ones cannot compute the exact unwrapped phase for the inverse transformation. Because of that, there are already small errors due to the computational procedure.

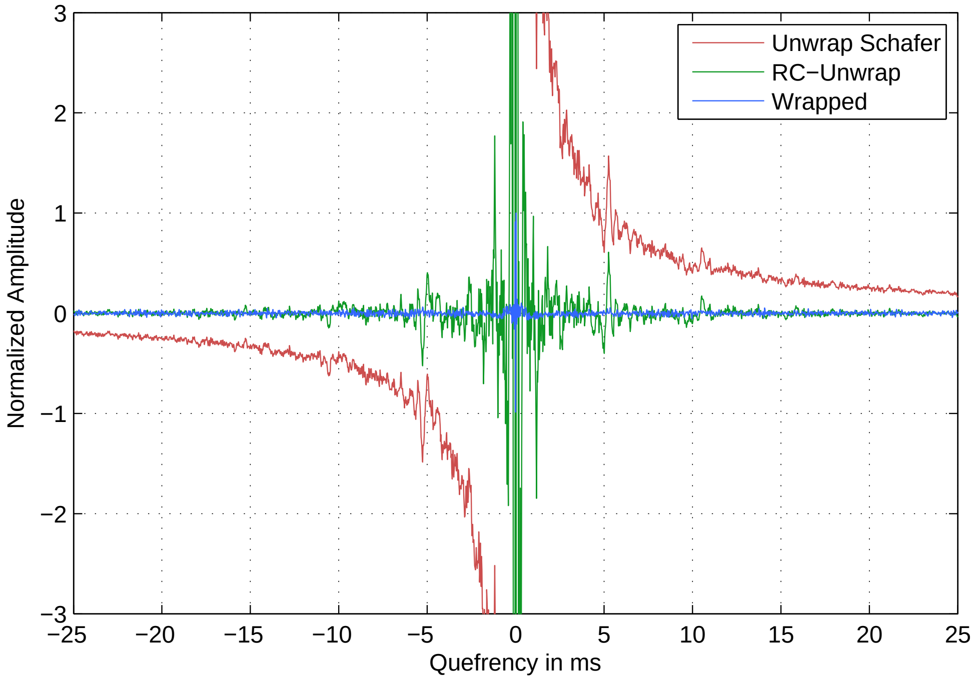


Figure 2.6 – Complex cepstrum of a single speech segment with the wrapped phase compared to different phase unwrapping algorithms

Figure 2.6 shows different efficient algorithms [16] for phase unwrapping compared to wrapped phase complex cepstrum with a speech segment as input. The Schafer algorithm [16] constructs the unwrapped phase by adding and subtracting 2π when the difference between adjacent phase spectrum values increases beyond a given threshold. RC-unwrap in the above code segment is a special version of the Schafer algorithm that subtracts a straight line from the phase. It can be seen, that the cepstral peaks fluctuate in bin index and amplitude wherefore the RIR peaks by the mean computation are more difficult to detect.

2.2.4 Cross-Talking due to STFT Based Compuation

There is another, more important problem that can corrupt the cepstral coefficients to non-causal effects. By computing the signal frame-wise, the windowing and overlap technique is employed to get the frames. Assuming a constant signal segment subdivided into frames with no pauses in it, a reflection cross-talk between two frames is observed. This means for example that the first samples of frame l contain reflections of the last samples of frame $l - 1$ Figure 2.7 visualizes this effect.

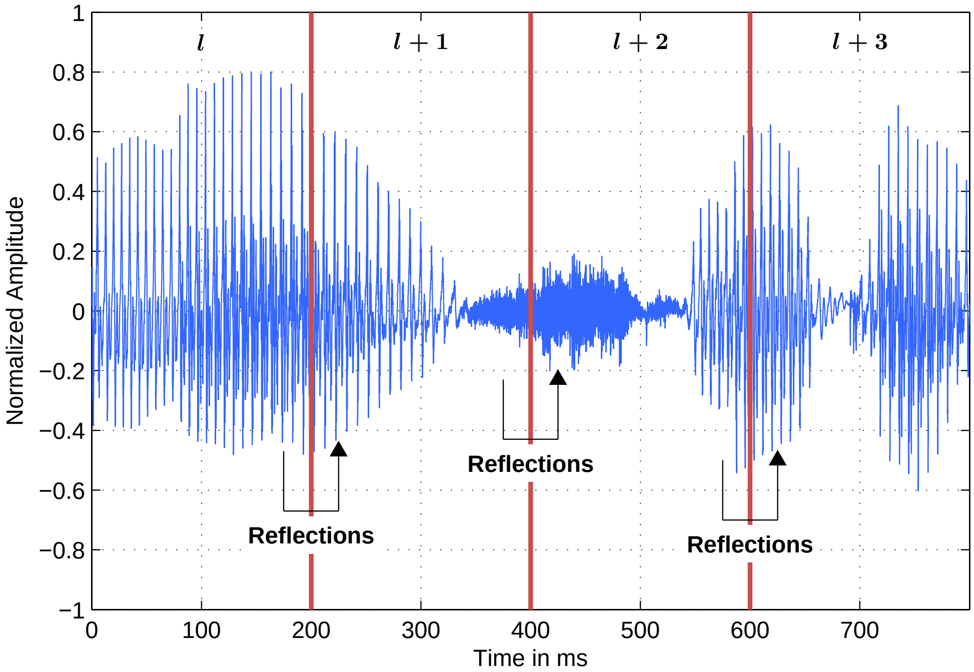


Figure 2.7 – Illustration of cross-talking reflections of the previous frame

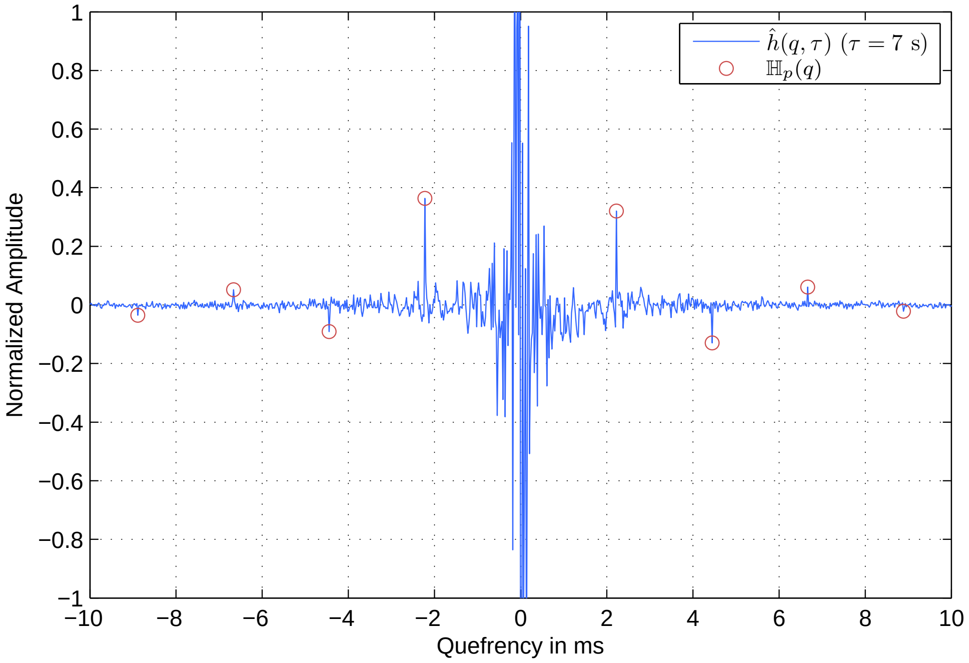


Figure 2.8 – Cepstral mean $\hat{h}(q,\tau)$ over $\tau = 7$ s of speech with corresponding cepstral peaks $\mathbb{H}_p(q)$

These reflections appear non-causal because they occur although their excitation lies in the previous frame. And they can be lowly correlated to the signal of the current frame. This effect also appears in the cepstral domain and causes non-symmetric peaks in the anti-causal complex cepstrum with quefrency bins $-q$ as shown in Figure 2.8.

Additionally causal peaks which correspond to the RIR are corrupted by the overlapping reflections. By subtracting these peaks in the cepstrum and transforming it back into the time domain, non-causal time shifts within the signal are produced. Thus it might be interesting to find appropriate frames where no overlapping reflections take place.

2.2.5 Cepstral Mean by Speech Segmentation

An approach for finding appropriate signal frames determines a segment-based analysis of the speech by detecting its voice activity [3]. Thereby, the problem of overlapping reflections as described before can be reduced. One idiosyncrasy of speech is that it contains many pauses where a speaker has to draw breath. These parts with no or insignificantly small signal amplitudes are used for the segmentation. If the algorithm only uses segments as frames following a pause, there are no reflections of a previous frame which can corrupt an actual frame. Figure 2.9 shows an example for such a voice activity segmentation of speech.

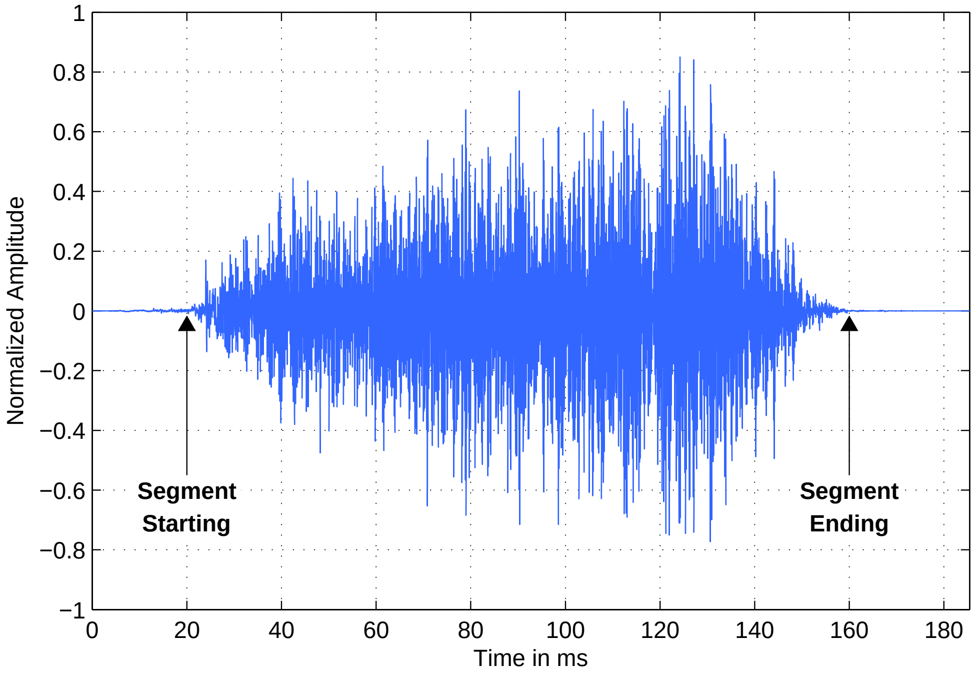


Figure 2.9 – Segmentation of speech

With such a segment, the mathematical model of the previous Section fits, and only causal cepstral peaks appear in the complex cepstrum of the signal frame.

Therefore, it is possible to detect a RIR by computing the cepstral mean

$$\hat{h}_s(q,\tau) = \frac{1}{|\mathbb{L}|} \cdot \sum_{l \epsilon \mathbb{L}} \hat{y}(q,l), \tag{2.27}$$

where $\mathbb{L}$ is a set of frames around the time period τ where the segmentation model fits. One problem of these technique is that the frame-wise subtraction is not possible anymore because the overlapping reflections corrupt a frame as we have seen and thus the subtraction does only fit to the segment frames which are used for calculating the cepstral mean. Alternatively it is possible to design an adaptive digital filter $g(n,\tau)$ consisting of linear-phase all-pass components

$$g(n,\tau) = -\hat{h}_s(q,\tau). \tag{2.28}$$

The accuracy of such a segmental cepstral mean also depends on the number of frames used. Obviously the active speech is much more present then pauses, so the number of appropriate segments decreases compared to the cepstral mean subtraction. If the number of usable frames within a time period τ is too small, the correlation between frames may be too high for an appropriate estimation of the RIR. It is therefore not robust enough for time-varying systems due to the very low amount of relevant frames.

2.3 Discussion

Unfortunately the results of this approach and all its techniques are not easily applicable even if the theory fits for appropriate reverberation cancellation. Computational complexity and the high sensitivity of the phase to wrapping, segmentation and noise make it difficult to use the complex cepstrum as a dereverberation method. By contrast, the real cepstrum might lead to an audible dereverberation. Still, its effect is too small to achieve reasonable results it in practical systems. Additionally, more complicated impulse responses than a simple comb filter have many reflections with low amplitude and some of them may be in the region around cepstral speech coefficients, some of them in the anti-causal part of the cepstrum. The sensitivity of the homomorphic deconvolution led to an investigation of another approach.

Chapter 3

Spectral Subtraction

In the previous Chapter, the underlying assumption was that reverberated signals can be expressed as a convolution of the impulse response of a system and the source signal plus noise. In this Chapter reverberation will be treated like a noisy part of the recorded speech. Thus it can be seen as an additive component so that the reverberated signal $x(n)$ can be expressed as

$$x(n) = s(n) + r(n) + v(n), \tag{3.1}$$

where $r(n)$ is the so called residual reverberation and $s(n)$ and $v(n)$ the given source signal and noise. As an advantage, the influence of the generally time-varying RIR can be estimated for each frame instead of estimating one impulse response over several frames. In the spectral domain, in terms of frame-based processing, it can be calculated as

$$X(k,l) = S(k,l) + R(k,l) + V(k,l). \tag{3.2}$$

The main idea of spectral subtraction is to jointly reduce reverberation and noise with a real valued gain function $G(k,l)$ in the spectral domain. Hence, the clean speech spectrum is estimated as

$$\hat{S}(k,l) = G(k,l) \cdot X(k,l). \tag{3.3}$$

Transforming $\hat{S}(k,l)$ back into the time domain leads to the dereverberated speech signal

$$\hat{s}(n) = \mathrm{iDFT}(\hat{S}(k,l)). \tag{3.4}$$

This Chapter introduces an iterative reverberation suppression scheme based on spectral subtraction, according to [10].

3.1 System Overview

The major challenge is to find an appropriate gain function to jointly achieve dereverberation and noise reduction. The gain function has to minimize the error between the source signal and the dereverberated signal according to Section 3.4. Therefore, an estimation of the power spectral densities (PSD) of the interfering reverberation and noise has to be done. Figure 3.1 shows an overview of the proposed spectral subtraction algorithm according to [10].

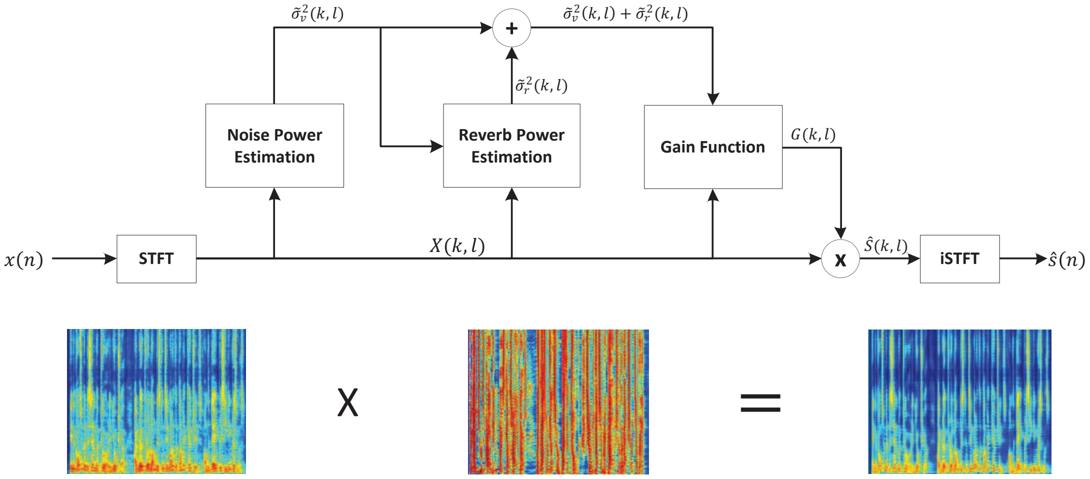


Figure 3.1 – Schematic overview of the proposed spectral subtraction system

In general there are three main stages of the algorithm. First, an estimate of the noise power spectral density (PSD) $\hat{\sigma}_v^2$ of a reverberated input short-time spectrum $X(k,l)$ is calculated by a technique proposed in Section 3.2. The second stage computes the reverberant PSD $\hat{\sigma}_r^2$ of $X(k,l)$. These two PSDs are then used to obtain the real valued gain function $G(k,l)$ in the last stage. The gain function $G(k,l)$ is multiplied with $X(k,l)$ to obtain the dereverberated output short-time spectrum $\hat{S}(k,l)$. An inverse short Fourier transformation reconstructs the dereverberated signal $\hat{s}(n)$. The following sections describe the main stages in detail.

3.2 Noise Power Estimation

The noise power estimation is based on the well-known method of minimum statistics as proposed in [17]. Instead of using a voice activity detector as in other approaches, this algorithm tracks spectral minima at each frequency bin even during speech activity. Additionally, an optimal smoothing parameter is determined in each frame by minimizing a mean square estimation error criterion. Based on the statistics of the spectral minima, a noise power estimator is defined.

3.2.1 Minimum Statistics

Minimum statistics (MS) is based on the observation that noise exhibits a more stationary and lower amplitude in frequency and time compared to speech signals. Further, the power level of a noisy signal frequently decays to the power level of noise during speech activity whereby reverberation follows another model and is highly correlated to speech. Thus we can derive an estimation of the noise PSD by tracking the minimum of a noisy signal PSD. Segments of speech presence can be bypassed because of areas of high energy. In Figure 3.1 we can see a short schematic overview of the proposed MS-System.

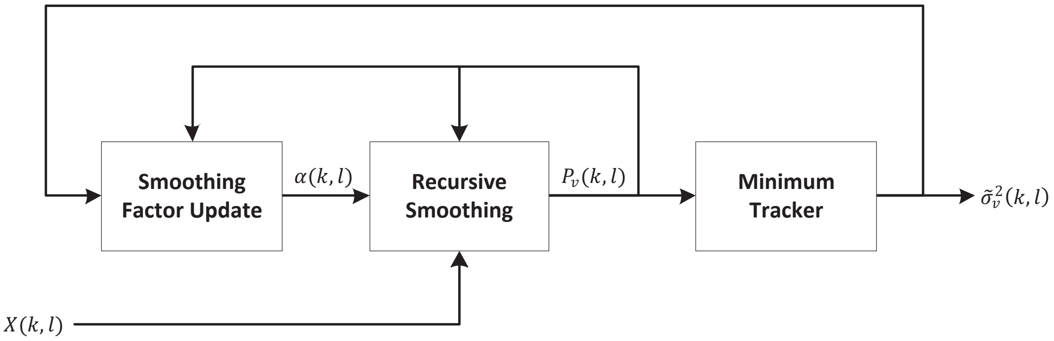


Figure 3.2 – Schematic overview of the MS noise estimation approach

The smoothing factor $\alpha(k,l)$ is calculated based on a previous noise estimation $\hat{\sigma}_v^2(k,l-1)$ as well as a previous instance of the smoothed periodogram $P_v(k,l-1)$. This smoothing factor is then used to obtain $P_v(k,l)$. Subsequently an estimation of the noise power spectrum $\sigma_v^2(k,l)$ is performed by minimum tracking of $P_v(k,l)$ within a sliding window. If we use Equation (3.2) and assume that the algorithm is robust to reverberation, we can write

$$X(k,l) = S(k,l) + V(k,l). \tag{3.5}$$

Taking into account that the speech energy is approximately or equal to zero in between words or during speech pauses, it is possible to track the minimum power with a sufficiently large sliding window. To avoid hard decision noise estimation, we first consider the smoothed periodogram

$$P_v(k,l) = \alpha(k,l)P_v(k,l-1) + (1-\alpha(k,l))|X(k,l)|^2, \tag{3.6}$$

where $\alpha(k,l)$ is the already mentioned smoothing factor. Figure 3.3 shows the periodogram of $|X(k,l)|^2$ as well as the smoothed periodogram $P_v(k,l)$ and the noise estimation $\hat{\sigma}_v^2(k,l)$ within a sliding window of length $T_{SL} = 100$ ms.

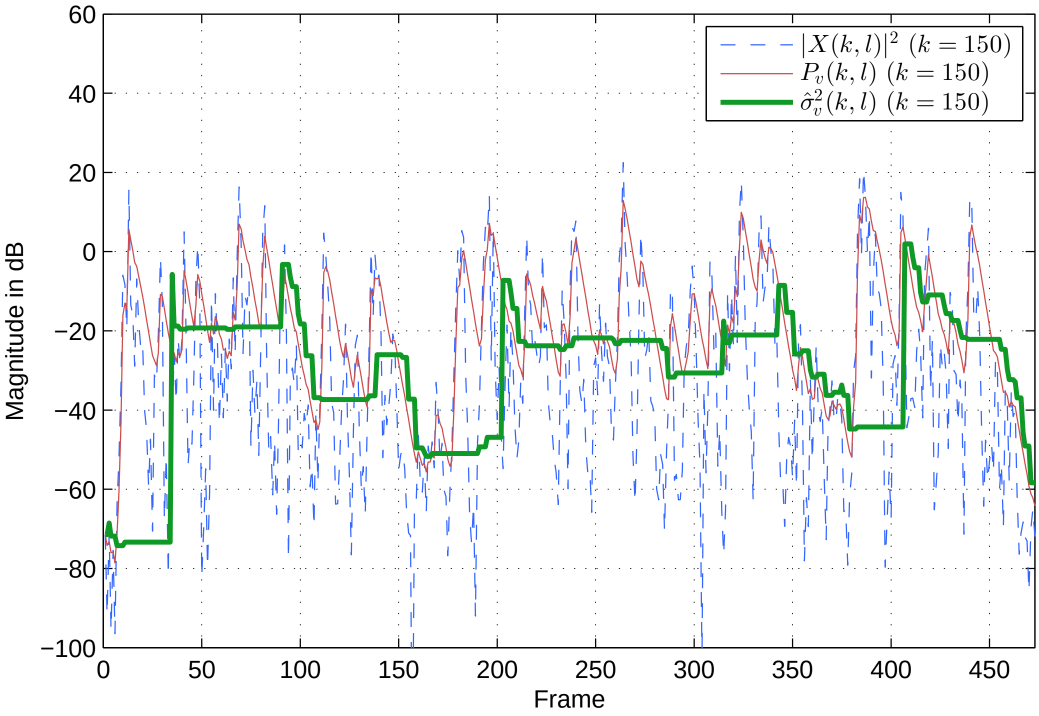


Figure 3.3 – Periodogram $|X(k,l)|^2$, smoothed periodogram $P_v(k,l)$ and noise estimate $\hat{\sigma}_v^2(k,l)$ for a single frequency bin $k = 150$

The length of the sliding window T_{SL} is crucial for an adequate noise PSD estimation. While too short windows might not include speech pauses to find a representative minimum that describes the noise level, very long windows are not robust to changes of the noise level. The length of the reverberation added to the speech might require larger periods T_{SL} to capture minimum levels representative for the noise level. For unknown reverberation times, longer tracking windows are more robust. However, as described in Section 3.3, only rooms with a reverberation time in the range of $0.3 \leq T_{60} \leq 0.7$ s are considered. Therefore, the value of the sliding window is set in the range of i.e. $1.5 \leq T_{SL} \leq 3.0$ s.

3.2.2 Optimal Smoothing

Figure 3.3 used a constant smoothing parameter α which yields smearing the peaks if speech is present. This can lead to an incorrect detection of the minimum inside the sliding window and therefore to an inaccurate estimation of noise. Further, if the noise power increases the estimated minimum is delayed due to lower discrepancy between noise and signal level. To circumvent this issue, a time-frequency dependent smoothing value $\alpha(k,l)$ is introduced which can be derived by minimizing the conditional mean square error

$$E\{(P_v(k,l) - \sigma_v^2(k,l))^2 | P_v(k,l-1)\}. \tag{3.7}$$

A solution for minimizing the error criterion to get an optimal smoothing value $\alpha_{opt}(k,l)$ is outlined in [17] and can be expressed as

$$\alpha_{opt}(k,l) = \frac{1}{1 + \left(\frac{P_v(k,l-1)}{\sigma_v^2(k,l)} - 1\right)^2}. \tag{3.8}$$

While $\sigma_v^2(k,l)$ is not available in practice, the true noise PSD is replaced by its previous value $\sigma_v^2(k,l-1)$. If we look at Equation (3.6) the smoothed periodogram will run into a deadlock $P_v(k,l) = P_v(k,l-1)$ for $\alpha(k,l) = 1$. Therefore, a maximum allowable smoothing value α_{max} is introduced. Further, due to the delay of the estimation of $\sigma_v^2(k,l-1)$, the true PSD in an actual frame might be either smaller or lager than the estimated PSD. To circumvent this issue, $\alpha(k,l)$ is corrected by monitoring the average short-term PSD estimate of the previous frame compared to the average periodogram as proposed in [17]. The correction is calculated by

$$\alpha_c(l) = 0.7\alpha_c(l-1) + 0.3\max(\tilde{\alpha_c}(l), 0.7), \tag{3.9}$$

$$\tilde{\alpha_c}(l) = \frac{1}{1 + \left(\frac{\sum_{k=0}^{N-1} P(k,l-1)}{\sum_{k=0}^{N-1} |X(k,l)|^2} - 1\right)^2}, \tag{3.10}$$

where the correction factor $\tilde{\alpha_c}(l)$ is limited by an empirically estimated value $\alpha_c(l)$. The optimal smoothing value $\alpha_{opt}(k,l)$ is corrected as

$$\alpha(k,l) = \frac{\alpha_{max}\alpha_c(l)}{1 + \left(\frac{P_v(k,l-1)}{\sigma_v^2(k,l-1)} - 1\right)^2}. \tag{3.11}$$

As a last step, a lower limit α_{min} was also found in [17] and is applied to increase the performance in higher levels of non-stationary noise. The noise PSD $\sigma_v^2(k,l-1)$ is estimated by using the corrected optimal smoothing value $\alpha_{opt}(k,l)$ from Equation (3.11) for the minimum tracking in the smoothed periodogram.

Figure 3.4 shows a noise estimation with the adaptive smoothing parameter $\alpha(k,l)$ with the same scheme as in Figure 3.3.

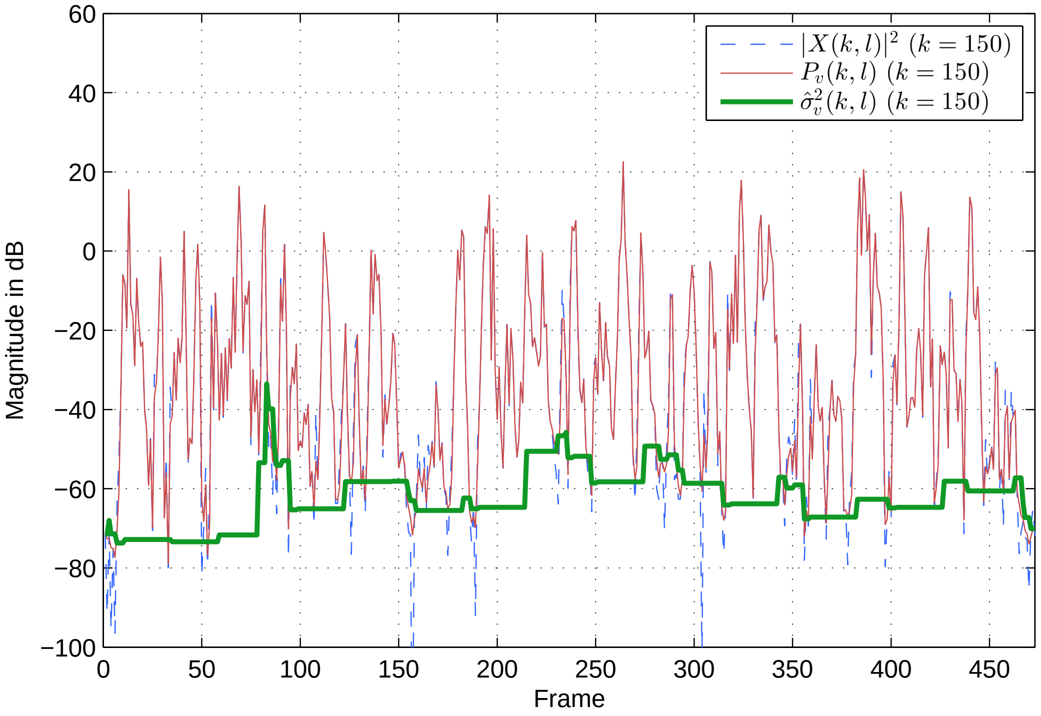


Figure 3.4 – Periodogram $|X(k,l)|^2$, smoothed periodogram $P_v(k,l)$ and noise estimate $\hat{\sigma}_v^2(k,l)$ for a single frequency bin $k = 150$ with an adaptive smoothing constant $\alpha(k,l)$

Compared to Figure 3.3, the peaks are not as often false detected and the estimator reacts more promptly to minima. The estimated noise power level $\hat{\sigma}_v^2(k,l)$ does not exhibit strong jumps and stays approximately constant over time as expected for the true noise power level $\sigma_v^2(k,l)$.

There is another approach called *minima controlled recursive averaging* (IMCRA) [18] that extends the MS approach with an additional rough voice activity detector (VAD) to provide better results in estimating the noise. But the IMCRA approach implies a higher complexity and this thesis is mainly focused on reducing reverberation. Therefore, the noise estimation results of the MS approach are good enough for our purposes.

3.3 Reverberation Power Estimation

In this Section the reverberant part of the speech is estimated from the levels of previous frames in the short-term power spectrum of reverberated speech. By so called cepstro-temporal smoothing [19] we can calculate a robust estimation of the noise-suppressed (but still reverberated) speech. This section describes the necessary steps to obtain a reverberation power estimation.

3.3.1 Polack's Reverberation Model

As outlined in Section 1.2, the RIR consists of direct sound, early reflections, and late reverberation. In general the time response of reverberation can be modeled by a decaying envelope. Polack [11] considers a time-domain model that treats the RIR as a Gaussian stationary noise signal $w_g(t)$ multiplied by an exponential decay rate δ

$$h(t) = w_g(t) \cdot e^{-\delta t} \quad \text{for } t \geq 0, \tag{3.12}$$

where the decay rate δ is related to the reverberation time T_{60} as

$$\delta = \frac{3 \ln 10}{T_{60} f_s}. \tag{3.13}$$

Due the convolution of the clean speech signal and the RIR in the time domain, the reverberated signal $x(n)$ also exhibits the exponentially decaying envelope. Thus, if we assume

$$\hat{\sigma}_z^2 = \hat{\sigma}_r^2(k,l) + \hat{\sigma}_s^2(k,l). \tag{3.14}$$

as the estimation of the reverberant speech PSD, reverberation can be seen as a multiplication of previous reverberant speech PSD frames and the exponential decay. $\hat{\sigma}_z^2$ can be obtained according to Section 3.3.3 by taking the estimated noise PSD $\hat{\sigma}_v^2$ into account to exclude noisy components when calculating reverberation along this model. According to the spectral subtraction method described in [8], an estimation $\hat{\sigma}_r^2(k,l)$ of the reverberant PSD $\sigma_r^2(k,l)$ can be estimated by

$$\hat{\sigma}_r^2(k,l) = e^{-2\delta T_d f_s} \cdot \hat{\sigma}_z^2(k, l - \tfrac{T_d}{T_s}), \tag{3.15}$$

where $T_s = \frac{M}{f_s}$ denotes the frame shift depending on the frame length M while T_d represents the time duration between the direct sound and early reflections. While Polack's model for a RIR fits properly to late reverberation where the diffuse sound field offers an equally spaced decay, it is hard to predict the decay level of early reflections or the exact time and amplitude since it depends on the room properties like dimensions and absorption coefficients.

Depending on the room dimensions T_d can be set between 20 ms and 80 ms. As we can see there is an uncertainty between the frame shift T_s and T_d. With a short frame length T_s, the minimum delay for the model $T_d = T_s$ can estimate the levels of the earlies part of the reverberation accurately. However, short frame shift T_s might cause an insufficiently resolved reverberated STFT frequency.

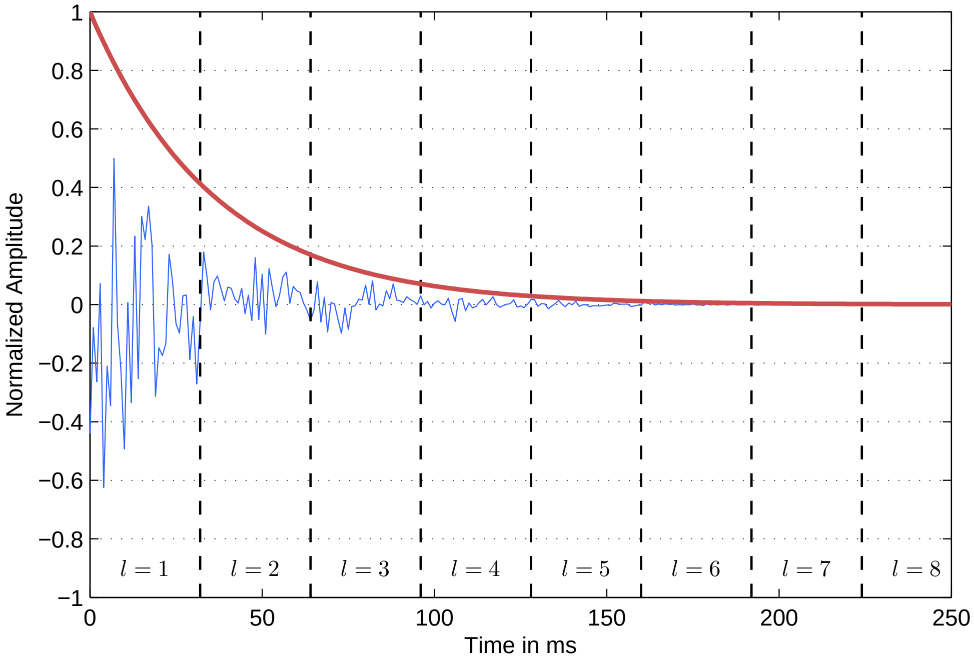
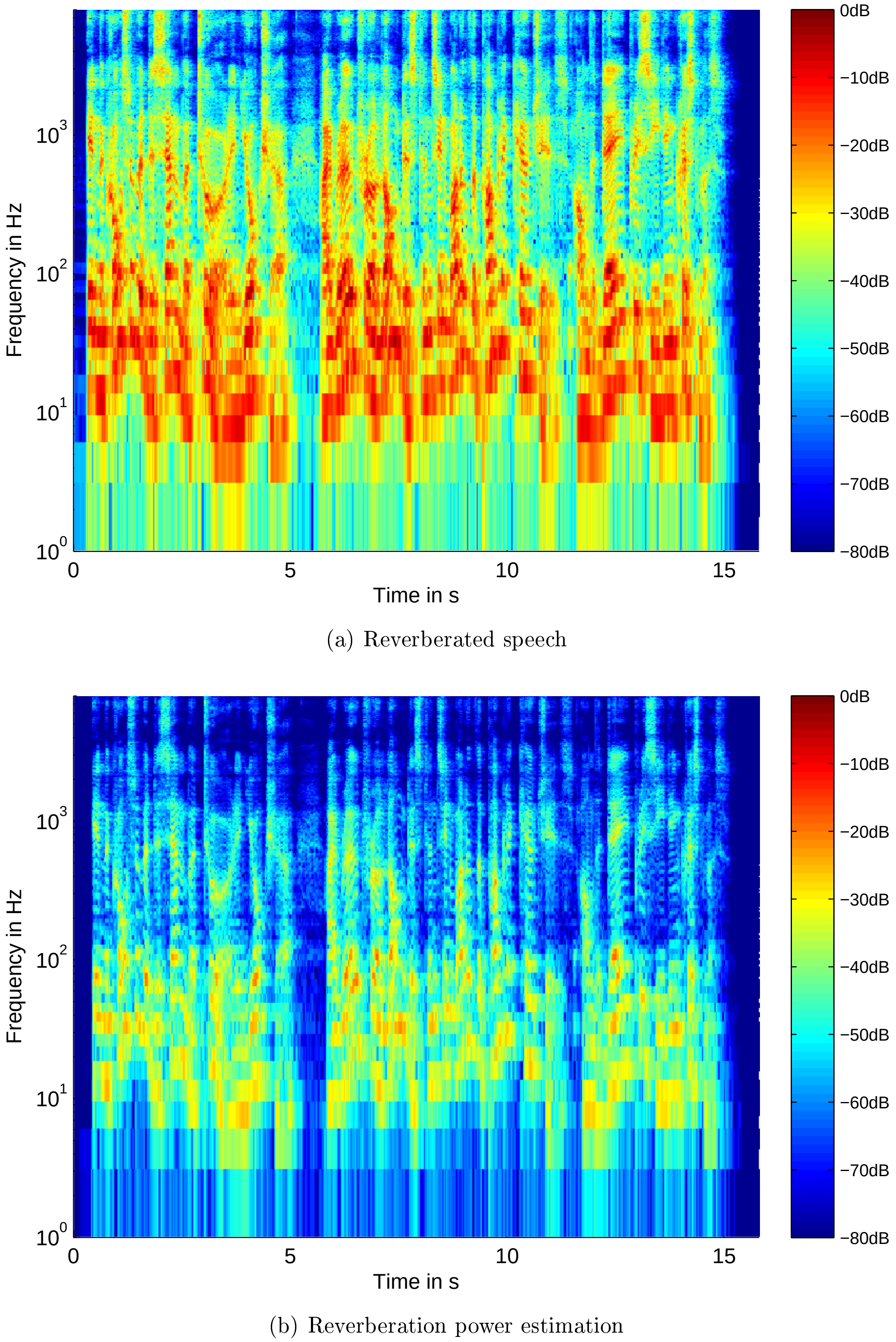


Figure 3.5 – Impulse response along Polack's Reverberation Model with exponential decay function

Assuming we have chosen a value for $T_d = 32$ ms with a frame shift of $T_s = 16$ ms, the frame index for reverberation estimation becomes $l - \frac{T_d}{T_s} = l - 2$. However, the reverberation model does not only predict reflections within one specific previous frame. As we can see in Figure 3.5, several frame exhibit parts of the reverberation. Thus, it might be suitable to take also neighboring frames like $l - 1$ and $l - 3$ into account. According to the decay model, only the value of T_d has to be adopted in advance. Because reverberation is treated as an additive part to the noisy speech power density, the estimation of early reflection levels can also be additive so that Equation (3.15) is extended to

$$\hat{\sigma}_r^2(k,l) = \sum_{i=1}^{I} e^{-2\delta T_{di} f_s} \cdot \hat{\sigma}_z^2(k, l - \tfrac{T_{di}}{T_s}), \tag{3.16}$$

where T_{di} is an element of a set $\mathbb{T}_d$ of I assumed early reflection i.e. $\mathbb{T}_d = \{32$ ms, 64 ms, 128 ms$\}$. The estimation of $\hat{\sigma}_r^2(k,l)$ can be better adjusted to the reverberation level than with a single frame shift for one specified early reflection. Figure 3.6 shows the spectrogram of a speech sample convolved with the office room RIR [14] and the corresponding reverberation estimation.



(a) Reverberated speech

(b) Reverberation power estimation

Figure 3.6 – Spectrograms of a reverberant speech segment and the corresponding reverberation power level estimation

The initial value of the early reflection duration time is set to $T_{d1} = 64$ ms. The frame length is set to $M = 32$ ms which means, that the reverberation power level estimate starts at frame $l - 2$. Thus, the spectrogram of the reverberation power level estimation shows an attenuated version of the $l - \frac{T_{di}}{T_s}$ previous reverberated speech frames, while number of previous frames used to estimate the reverberation power level is set to $|\mathbb{T}_d| = 3$.

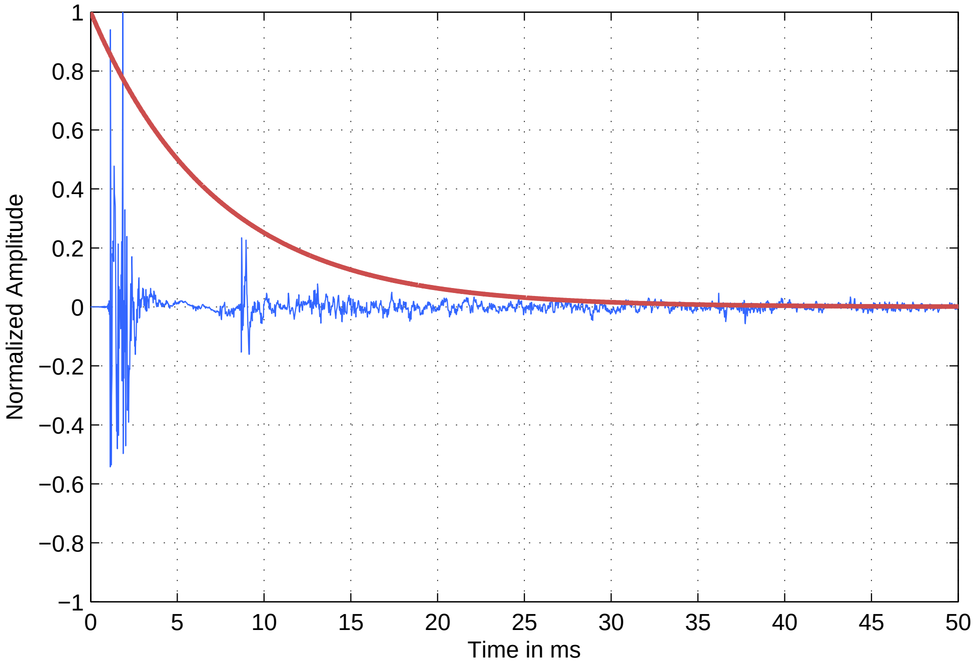


Figure 3.7 – Meeting room RIR and exponential decay function

Figure 3.7 shows a real RIR recorded in a meeting room [14] and the exponential decay function of Polack's reverberation model. The illustrated RIR does not consequently follow Polack's reverberation model but exhibits early reflections at about 9 ms with a higher amplitude than than previous samples of the RIR. Only for the first 5 ms, the RIR decays rapidly after the direct sound sample. To achieve suppression of early reflections in that range, the frame length M had to be at least 9 ms or lower which is not practical because of the low frequency resolution during the STFT computation. Thus, in rooms with a low reverberation time such as the meeting room with $T_{60} = 0.23$ s, the earliest reflections cannot be suppressed in under practical conditions. Depending on the considered room, the reverberation model might fit better or worse.

3.3.2 Reverberation Time Estimation

As we can see from Equation (3.13), the estimated T_{60} determines a big part of the reverberation power estimation. There are several approaches, for example the maximum-likelihood approach [20], or a noise-robust estimation based on spectral decay distributions [21]. As announced in Section 1.3, only early reflections or in other words small room acoustics are interesting. Therefore, only speech samples which were recorded in small rooms with a reverberation time in the range of $0.3 \leq T_{60} \leq 0.7$ s s are analyzed.

To investigate the influence of T_{60} on the dereverberation procedure, the exponential decay function $f_e(T_d)$ from Polack's reverberation model is considered. According to Equation (3.12), $f_e(T_d)$ is given by

$$f_e(T_d) = e^{-\frac{6 \ln 10}{T_{60}} \cdot T_d}. \tag{3.17}$$

Figure 3.8 shows these exponential decay functions for different T_{60} in the range of $0.3 \leq T_{60} \leq 0.7$ s s as well as the maximum error function

$$e_{max}(T_d) = e^{-\frac{6 \ln 10}{0.7 \text{ s}} \cdot T_d} - e^{-\frac{6 \ln 10}{0.3 \text{ s}} \cdot T_d}. \tag{3.18}$$

The maximum of $e_{max}(T_d)$ is located at around 32 ms ans its value peaks at $e_{max}(32 \text{ ms}) = 0.3$, only. So if an estimation of T_{60} exhibits a discrepancy of ± 0.3, the maximum decay value difference during the calculation of $\hat{\sigma}_r^2(k,l)$ is 30%, if we choose $T_{60} = 0.3$ s while the true reverberation time is $T_{60} = 0.7$ s or vice versa.

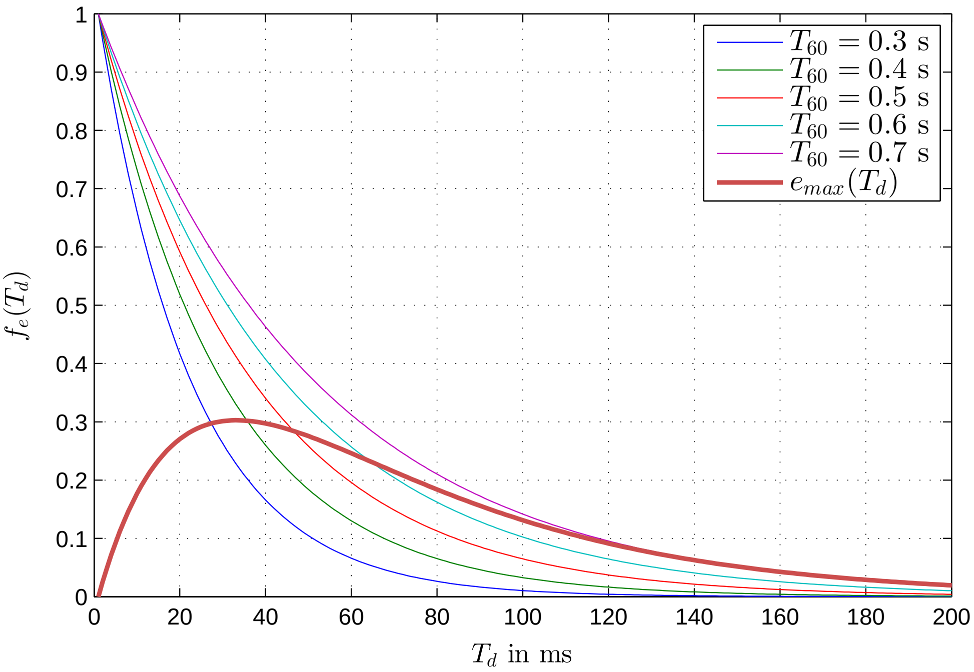


Figure 3.8 – Exponential decay functions with different T_{60} and corresponding maximum error function

By choosing $T_{60} = 0.5$ s, the maximum of $e_{max}(T_d)$ is just 18% at 29 ms. During the parameter optimization it has been detected that this discrepancy is not perceivable along our dereverberation procedure.

Because of this observation and the fact, that even the T_{60} estimation algorithms exhibit an error about ±0.2 s, it is for our purpose suitable to suppose in general a fixed reverberation time T_{60} in the range of $0.3 \leq T_{60} \leq 0.7$ s s so that the maximum discrepancy can be minimized.

3.3.3 Cepstro-Temporal Smoothing

In addition to the reverberation level estimation $\hat{\sigma}_r^2(k,l)$, Equation (3.16) also requires the estimation of the reverberated speech PSD $\hat{\sigma}_z^2(k,l)$. In order to achieve noise-robustness, the estimated noise PSD $\hat{\sigma}_v^2(k,l)$ of Section 3.2 is used to calculate a maximum-likelihood (ML) estimator ξ_z^{ml} of the *a priori* signal to noise ratio (SNR)

$$\xi_z^{ml} = \frac{|X(k,l)|^2}{\hat{\sigma}_v^2(k,l)} - 1. \tag{3.19}$$

The reverberated speech power $P_z(k,l)$ suppressing segments dominated by noise can be calculated as

$$P_z(k,l) = \hat{\sigma}_v^2(k,l) \cdot \max(\xi_z^{ml}, \xi_{min}^{ml}), \tag{3.20}$$

where $\xi_{min}^{ml} > 0$ is used as a lower bound for ξ_z^{ml} to prevent negative or zero values. Based on the idea described in Section 2.1, cepstro-temporal smoothing is applied to $P_z(k,l)$ as introduced in [19]. The cepstrum of $P_z(k,l)$ is used to design an appropriate, adoptively smoothed lifter suppressing parts of the signal that are dominated by noise. Along Equation (2.3), the real cepstrum $\hat{c}_P(q,l)$ of $P_z(k,l)$ is

$$\hat{c}_P(q,l) = \mathrm{iDFT}\left[\ln[|P_z(k,l)|]\right]. \tag{3.21}$$

Because of the symmetric real cepstrum, the following operations only need to be applied to the first half $q = 0, ..., \frac{K}{2}$, with K denoting the DFT-length assuming K even. For the inverse transformation back into the power spectral domain, the second half uses the mirrored first-half information to achieve a symmetric cepstrum. While coefficients representing the spectral envelope of the vocal tract are always at around $q = 0$, the fundamental frequency f_0 yields to a cepstral peak at $q = \frac{f_s}{f_0}$ that is time-varying and constantly changing in quefrency. Therefore, liftering needs to be adaptive in the cepstral domain. To suppress musical artifacts of rapidly changing steep lifter slopes, an efficient smoothing is applied to $\hat{c}_P(q,l)$,

$$\hat{c}_P(q,l) = \alpha(q,l)\hat{c}_P(q,l-1) + (1-\alpha(q,l))\hat{c}_P(q,l), \tag{3.22}$$

with an adaptive smoothing factor $\alpha(q,l)$ that is quefrency- and frame-depended. The smoothing factor represents an adaptive lifter in the cepstral domain in a similar way as outlined in Chapter 2.

It should be chosen so that little smoothing is applied to the low cepstral coefficients as well as the range around the quefrency $q = \frac{f_s}{f_0}$. The smoothing value $\alpha(q,l)$ can be adaptively updated by detecting f_0:

$$\alpha(q,l) = \begin{cases} \alpha_{pitch} & \text{if } q \in \mathbb{Q}_{pitch} \\ \overline{\alpha}(q,l) & \text{if } q \notin \mathbb{Q}_{pitch} \end{cases}, \tag{3.23}$$

where $\mathbb{Q}_{pitch}$ is a set of adjacent cepstral bins that represent $\frac{f_s}{f_0}$, while α_{pitch} denotes the smoothing constant for these bins. $\overline{\alpha}(q,l)$ represents previous estimations of f_0 and the constant values $\overline{\alpha}_{const}(q)$ for weighting the cepstral bins at the lower quefrencies corresponding to the spectral characteristics of the vocal tract. The cepstral peak of f_0 may jump between two frames, and therefore an estimation error of f_0 would lead to a strong smoothing of the true f_0 bin. To suppress the effect of such an error, also $\overline{\alpha}(q,l)$ is smoothed as

$$\overline{\alpha}(q,l) = \beta\alpha(q,l-1) + (1-\beta)\overline{\alpha}_{const}(q), \tag{3.24}$$

where the smoothing constant β is a factor that determines the speed of adapting α_{pitch} to $\overline{\alpha}_{const}(q)$, if it became lower in a previous frame. If f_0 is detected in frame $l = l_0$, the smoothing factor $\alpha(q,l)$ at α_{pitch} of the frame $l = l_0 - 1$ is slowly decreasing frame by frame and not getting zero immediately.

The fundamental frequency can be estimated by a simple maximum detection at the typical quefrencies $q = \frac{f_s}{f_0}$ for $80 \leq f_0 \leq 400$ Hz. For improving the f_0 estimation algorithm, the log-spectrum is low-pass filtered since f_0 does not appear in higher frequencies. Convolving the power cepstrum with a short Hamming window $w_H(q)$ of length 8 converts into a multiplication in the log-spectrum domain and is therefore low-pass filtered

$$\hat{c}_P^{lp}(q,l) = \hat{c}_P(q,l) * w_H(q). \tag{3.25}$$

The quefrency bin $q_{pitch}(l)$ representing f_0 is calculated as the maximum value within the interval $q_{low} \leq q \leq q_{high}$

$$q_{pitch}(l) = \arg\max_q \{\hat{c}_P^{lp}(q,l) | q_{low} \leq q \leq q_{high}\}, \tag{3.26}$$

where $q_{low} = \frac{f_s}{f_{0,high}}$ and $q_{high} = \frac{f_s}{f_{0,low}}$ are relevant bounds for searching $q = \frac{f_s}{f_0}$. Since voiced sounds are characterized by relatively high energy, the cepstral is typically clearly dominant. On the other hand, unvoiced sounds often do not produce a distinct peak. It is reasonable to introduce a threshold Λ^{thr} as a minimum value that has to be exceeded. Usually it is set to $\Lambda^{thr} = 20\%$ of the normalized cepstral amplitude. To further compensate errors during the fundamental frequency detection, the set $\mathbb{Q}'_{pitch}$ is introduced to search for peaks around the probable quefrency bin in the cepstrum.

From this $\mathbb{Q}_{pitch}$ is calculated as

$$\mathbb{Q}_{pitch} = \begin{cases} \mathbb{Q}'_{pitch} & \text{if } \hat{c}_P^{lp}(q_{pitch}, l) \geq \Lambda^{thr} \\ 0 & \text{otherwise} \end{cases} . \tag{3.27}$$

The set of bins around $q = \frac{f_s}{f_0}$ are defined by $\mathbb{Q}'_{pitch} = \{q_{pitch} - \Delta q_{pitch}, ..., q_{pitch} + \Delta q_{pitch}\}$, where Δq_{pitch} denotes the margin which produces the cepstral f_0. After calculating an appropriate smoothing value $\alpha(k, l)$, we are able to compute the cepstral smoothing along Equation (3.22). Finally, an estimation of the reverberated speech power $\hat{\sigma}_z^2(k, l)$ can be done by a computation of an inverse complex cepstrum

$$\hat{\sigma}_z^2(k, l) = \exp\left(\kappa + \text{DFT}\left\{\hat{c}_P(q, l)\right\}\right), \tag{3.28}$$

where κ [22] is a constant coefficient for bias compensation due to cepstral smoothing. Figure 3.9 shows the low-pass filtered real power cepstrum $\hat{c}_P^{lp}(q, l)$ of a single reverberated speech frame.

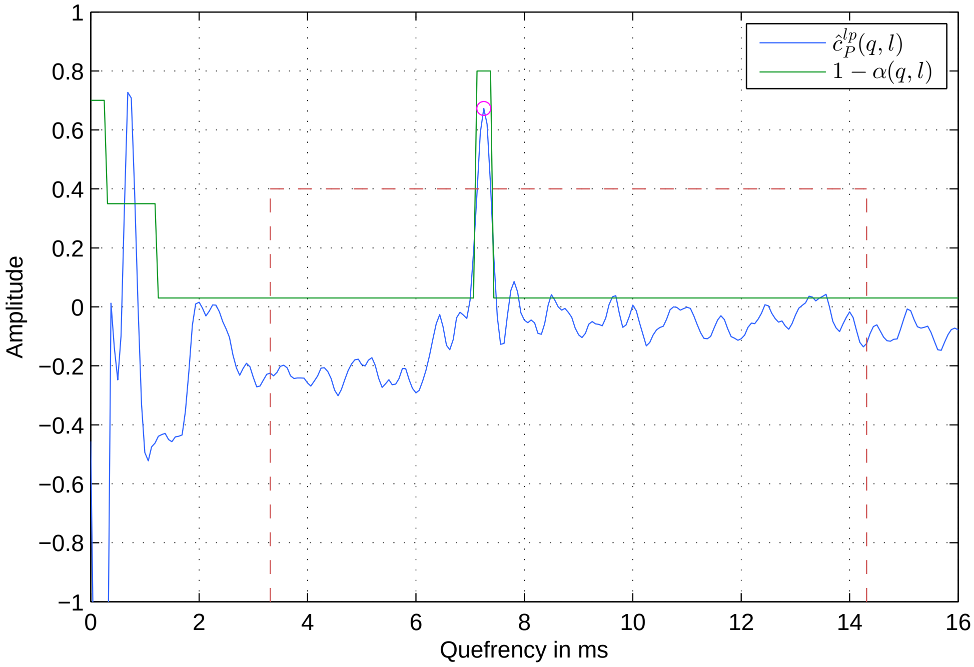


Figure 3.9 – Low-pass filtered real power cepstrum $\hat{c}_P^{lp}(q, l)$ of a speech frame with corresponding actual $1 - \alpha(q, l)$ with a f_0 peak at 7.5 ms without influence of previous frame

The green line displays the actual smoothing factor $1 - \alpha(q, l)$. At 7.5 ms, the magenta colored peak represents an estimated f_0 around which the interval contains the value $1 - \alpha_{pitch}$ due to $\mathbb{Q}'_{pitch}$. According to (3.23) only the values for $\mathbb{Q}_{pitch}$ vary over time, $\alpha(q, l)$ will be initialized as $\overline{\alpha}_{const}$ to achieve constant smoothing to lower cepstral values.

Figure 3.10 shows the effect of smoothing $\overline{\alpha}(q,l)$. In the displayed frame, f_0 is estimated at 6.3 ms. The interval of $1-\alpha(q,l)$ around 6.3 ms is again set to $1-\alpha_{pitch}$. The range $q_{low} \leq q \leq q_{high}$, where a f_0 estimation occurs is illustrated with the dotted red line. Further, the values of $1-\alpha(q,l)$ around the previous interval where $f_0 = 7.5$ ms was detected, are attenuated but still present because of the smoothing in Equation (3.24). In the quefrency range $0 \leq q \leq 1.5$ ms, $1-\alpha(q,l)$ contains the constant values $\overline{\alpha}_{const}(q)$ that are not affected by the smoothing of $\overline{\alpha}(q,l)$. It consists of three or more steps to weight quefrencies around $q=0$ with decreasing values by increasing quefrency q. Only the interval around the f_0 estimate is smoothed over time and set to $1-\alpha_{pitch}$ if a quefrency bin corresponding to f_0 was found.

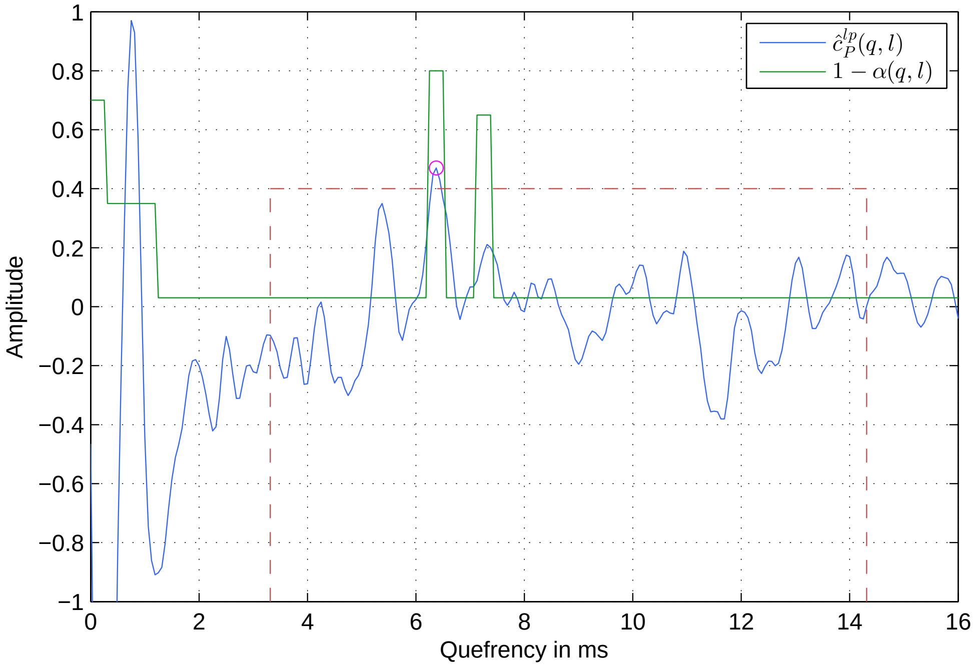
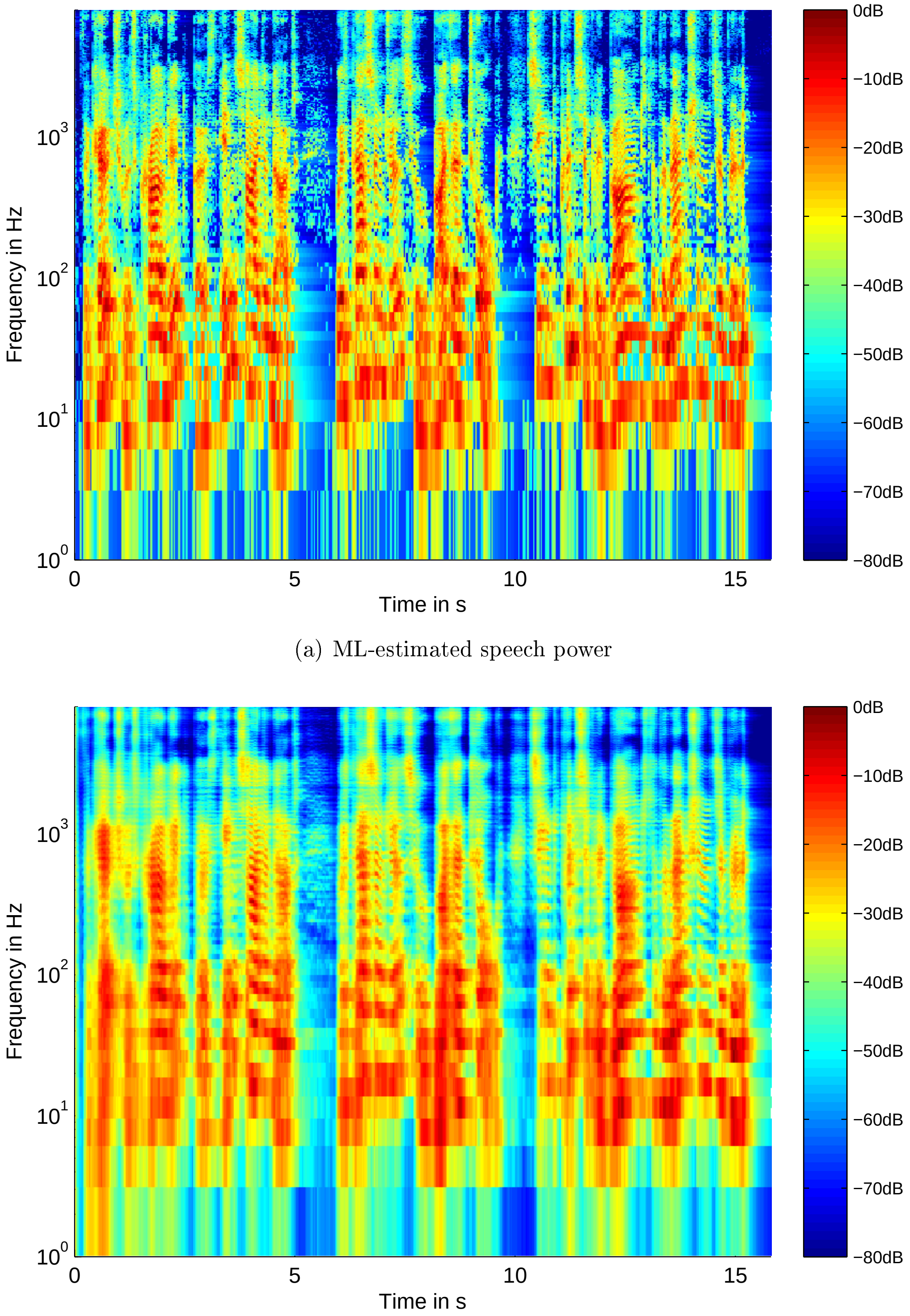


Figure 3.10 – Low-pass filtered real power cepstrum $\hat{c}_P^{lp}(q,l)$ of a speech frame with corresponding actual $1-\alpha(q,l)$ with a f_0 peak at 6.2 ms with the influence of $\alpha(k,l-1)$

Figure 3.11 shows the difference between a cepstro-temporal smoothed and the unprocessed ML-estimated reverberant speech power spectrogram. The smoothing of the real power cepstrum can be clearly seen by the smeared spectrogram over time. Since only the noise estimate $\hat{\sigma}_v^2(k,l)$ is taken into account, a discrepancy in the frequency domain is not detectable yet. However, in Section 3.4.2, also the reverberation estimate $\hat{\sigma}_r^2(k,l)$ is applied to calculate the maximum-likelihood of the *a priori* signal to interference ratio (SIR). Therefore, the cepstro-temporal smoothing can be applied to calculate the dereverberated speech power estimate with $\hat{\sigma}_r^2(k,l)+\hat{\sigma}_v^2(k,l)$ as interference in the denominator of Equation 3.19.



(a) ML-estimated speech power

(b) Cepstro-temporal smoothed estimated speech power

Figure 3.11 – Spectrograms of a speech power estimation

3.4 Spectral Gain Function

In the last block of the dereverberation system, a real-valued spectral gain function $G(k,l)$ is calculated to suppress reverberant parts of $X(k,l)$ as given in Equation (3.3). Assuming reverberation and noise to be additive interferences to the source signal as described in Equation (3.1), the main goal is to find an appropriate $G(k,l)$ to minimize the error between the source signal $s(n)$ and the dereverberated signal $\hat{s}(n)$. The derivation of the typical MMSE approach is explained in the next section.

3.4.1 MMSE Approach

If we assume a simple Wiener filter as shown in Figure 3.12, the optimal filter $h(n)$ can be found as soon as the error function $e(n)$ becomes zero.

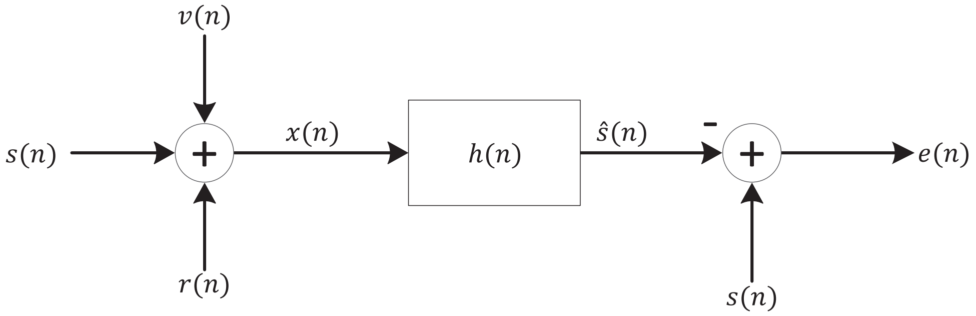


Figure 3.12 – Schematic overview of the Wiener filter

Thus, by minimizing the expected mean squared error between the source signal $s(n)$ and the dereverberated speech $\hat{s}(n)$ [23]

$$\min E\left\{e^2(n)\right\} = \min E\left\{(\hat{s}(n) - s(n))^2\right\} \quad (3.29)$$

$$= \min E\left\{(h(n) * x(n) - s(n))^2\right\} \quad (3.30)$$

where Equation (3.29) is called the MMSE criterion, we can calculate an optimal filter by equating its derivate $h(n)$ to zero [23]

$$h(n) * \psi_{xx}(n) = \psi_{xs}(n). \quad (3.31)$$

$\psi_{xx}(n)$ denotes the autocorrelation of the noisy reverberated signal $x(n)$ and $\psi_{xs}(n)$ the cross-correlation between the source signal $s(n)$ and $x(n)$. After transforming Equation (3.31) into the STFT domain, we obtain

$$H(k,l) \cdot \sigma_{xx}^2(k,l) = \sigma_{xs}^2(k,l), \quad (3.32)$$

where $\sigma_{xx}^2(k,l)$ represents the true PSD of the reverberated signal and $\sigma_{xs}^2(k,l)$ the true cross-PSD between the noisy reverberated and the source signal. Assuming orthogonality between $x(n)$ and $s(n)$, we may write $\sigma_{xs}^2(k,l) = \sigma_{ss}^2(k,l)$ and $\sigma_{xx}^2(k,l) = \sigma_{ss}^2(k,l) + \sigma_{vv}^2(k,l) + \sigma_{rr}^2(k,l)$, where $\sigma_{vv}^2(k,l)$ and $\sigma_{rr}^2(k,l)$ are the true PSDs of noise and reverberation. Thus we can write a solution for $H(k,l)$ with the true SIR $\xi(k,l)$,

$$H(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)}, \quad \text{with} \quad \xi(k,l) = \frac{\sigma_{ss}^2(k,l)}{\sigma_{vv}^2(k,l) + \sigma_{rr}^2(k,l)}, \tag{3.33}$$

where $H(k,l)$ is used as a spectral gain function $H(k,l) = G_{Wiener}(k,l)$ to reconstruct the dereverberated signal. Since we have no knowledge about any of the true PSDs in Equation (3.33), we have to calculate an estimation of the noise PSD $\hat{\sigma}_v^2(k,l)$, the reverberation PSD $\hat{\sigma}_r^2(k,l)$, and the source signal PSD $\hat{\sigma}_s^2$.

3.4.2 Dereverberated Speech Power Estimation

No matter if we use a Wiener filter alone or another variation for solving the MMSE criterion, we have to estimate at least an *a priori* SIR defined by

$$\hat{\xi}(k,l) = \frac{\hat{\sigma}_s^2(k,l)}{\hat{\sigma}_r^2(k,l) + \hat{\sigma}_v^2(k,l)}. \tag{3.34}$$

As described in Section 3.2 and 3.3, we are already able to calculate $\hat{\sigma}_r^2(k,l)$ and $\hat{\sigma}_v^2(k,l)$. The source signal estimation $\hat{\sigma}_s^2(k,l)$ is feasible by the afore-mentioned cepstro-temporal smoothing. According to Equation (3.19), we can already specify a ML-Estimation of the dereverberated speech power $P_s(k,l)$ with

$$\xi_s^{ml}(k,l) = \frac{|X(k,l)|^2}{\hat{\sigma}_v^2(k,l) + \hat{\sigma}_r^2(k,l)} - 1. \tag{3.35}$$

$$P_s(k,l) = \left[\hat{\sigma}_v^2(k,l) + \hat{\sigma}_r^2(k,l)\right] \max\left(\xi_s^{ml}(k,l), \xi_{min}^{ml}\right) \tag{3.36}$$

Subsequently the cepstro-temporal smoothing is applied as described in Section 3.3.3. Using the *a posteriori* SIR

$$\hat{\gamma}(k,l) = \frac{|X(k,l)|^2}{\hat{\sigma}_v^2(k,l) + \hat{\sigma}_r^2(k,l)}, \tag{3.37}$$

any spectral gain function can be developed for jointly suppressing both noise and reverberation. The next sections introduce some approaches for calculating a suitable spectral gain function $G(k,l)$ based on the *a priori* and *a posteriori* SIRs for solving the MMSE criterion.

3.4.3 Ephraim and Malah Spectral Subtraction Rule

Since the simple Wiener filter for solving the MMSE criterion produces disturbing artifacts known as *musical noise*, the so called Ephraim and Malah spectral subtraction rule (EMSR) was developed. It employs real-valued, frequency- and time-dependent gain function $G(k, l)$ to reconstruct the spectral amplitude of the source signal. The EMSR assumes that both the source signal and the interference stem from a process with complex Gaussian distribution, and that adjacent STFT-bands are orthogonal to each other

$$X(k,l) \perp X(k,l \pm 1). \tag{3.38}$$

Ephraim and Malah developed their *minimum mean-square error short-time spectral amplitude estimator* (MMSE-STSA) based on the assumption that speech and noise are random variables and statistically independent (for better readability we shortly use the acronym SA for MMSE-STSA). Thus on the basis of the Gaussian statistical model for spectral components, a gain function $G_{SA}(k, l)$ was derived in [24] which is defined by the *a priori* and *a posteriori* SIR and can be calculated by minimizing the MMSE of the spectra.

$$G_{SA}(k,l) = \frac{\sqrt{\pi}}{2}\sqrt{\frac{1}{\hat{\gamma}(k,l)}\frac{\hat{\xi}(k,l)}{1+\hat{\xi}(k,l)}}M\left[\hat{\gamma}(k,l)\frac{\hat{\xi}(k,l)}{1+\hat{\xi}(k,l)}\right], \tag{3.39}$$

$$\text{where } M(\mu) = e^{-\mu/2}\left[(1+\mu)I_0\left(\frac{\mu}{2}\right) + \mu I_1\left(\frac{\mu}{2}\right)\right] \tag{3.40}$$

denotes the confluent hypergeometric function of the first kind composed of the Bessel functions of 0. and 1. order. An extended approach of Ephraim and Malah was developed in [25]. Compared to the MMSE-STSA method the so called *minimum mean-square error log-spectral amplitude estimator* (MMSE LOG-STSA or LSA) minimizes the error of the log-spectra instead of the unprocessed spectra. It assumes that the mean-square error of the log-spectra is subjectively more convenient because of the logarithmic perception of sound volume of human hearing. Therefore, the spectral gain function $G_{LSA}(k, l)$ can be calculated by [26]

$$G_{LSA}(k,l) = \frac{\hat{\xi}(k,l)}{1+\hat{\xi}(k,l)} \exp\left(\frac{1}{2}\int_{\nu(k,l)}^{\infty}\frac{e^{-t}}{t}dt\right) \tag{3.41}$$

$$\text{with } \nu(k,l) = \frac{\hat{\xi}(k,l)}{1+\hat{\xi}(k,l)} \cdot \hat{\gamma}(k,l). \tag{3.42}$$

As we can see from Equation (3.42), the value $\nu(k, l)$ can be expressed by the Wiener gain function as $\nu(k, l) = G_{Wiener}(k, l) \cdot \gamma(k, l)$.

Furthermore, the exponential integral function can be written as an expansion in series of Bronstein [27], as

$$E_1(\nu(k,l)) = \int_{\nu(k,l)}^{\infty} \frac{e^{-t}}{t} dt = -C - \ln(\nu(k,l)) - \sum_{r=1}^{\infty} \frac{(-\nu(k,l))^r}{r \cdot r!}, \tag{3.43}$$

with $C = 0,5772156649...$(Euler Constant).

By limiting the sum of Equation 3.43, this expression can be used for a practical implementation of the LSA algorithm given by

$$G_{LSA}(k,l) = G_{Wiener}(k,l) \cdot \exp\left(\frac{1}{2} \cdot E_1(\nu(k,l))\right). \tag{3.44}$$

The range of $\nu(k,l)$ is limited to $0 < \nu(k,l) < \pi$ for numerical stability, the exponential integral will converge towards zero as the denominator $r \cdot r!$ of the sum increases rapidly with increasing r. Thus, it is suitable to stop the infinite sum at low values of r. For example for $r = 10$ the influence of the 10th element is at least 10 times smaller than the influence of $r = 9$ and even approximately $32 \cdot 10^6$ smaller than the ratio at $r = 1$. Therefore, appropriate accuracy may be achieved already with a maximum value of $r = 6$ as proposed in [26]. However, the computation of such an exponential integral requires high computational complexity. Thus, it is practical to search for simple alternatives such as the approximated LSA algorithm which can lead to a similar gain function.

3.4.4 MMSE-LSA Approximation

A simple approximation of the exponential integral function $E_1(\nu(k,l))$ can be found in [26]. If we look at Equation (3.43) and take Figure 3.13 into account, the logarithmic function $-\ln(\nu(k,l))$ increases for small arguments $\nu(k,l) \ll 1$ while the sum produces lower values for decreasing arguments $\nu(k,l)$. Therefore, it is suitable to neglect the sum for $\nu(k,l) \ll 1$ and the Equation simplifies to

$$E_1(\nu(k,l)) = -\ln(\nu(k,l)) - C = -\ln(\nu \cdot e^C). \tag{3.45}$$

Hence, the range of values for this definition holds for $\nu(k,l) \ll e^C$. For values $\nu(k,l) \gg e^C$, the exponential integral $E_1(\nu(k,l))$ converges towards zero. According to [26], a rough approximation is given by

$$E_1(\nu(k,l)) = -\ln\left(\frac{\nu}{e^{-C} + \nu}\right). \tag{3.46}$$

By means of Equation 3.46, the approximated LSA spectral gain function $G_{Approx} \approx G_{LSA}$ can be calculated as

$$\begin{aligned} G_{Approx}(k,l) &= G_{Wiener}(k,l) \cdot \exp\left(-\frac{1}{2}\ln\left[\frac{\nu(k,l)}{e^{-C}+\nu(k,l)}\right]\right) \\ &= G_{Wiener}(k,l) \cdot \sqrt{\frac{e^{-C}+\nu(k,l)}{\nu(k,l)}} \\ &= \sqrt{G_{Wiener}(k,l) \cdot \left(G_{Wiener}(k,l) + \frac{e^{-C}}{\hat{\gamma}(k,l)}\right)}, \end{aligned} \tag{3.47}$$

which shows a simple and computationally efficient way for calculating the spectral gain function. In Section 3.4.6, a short comparison between the introduced approaches is discussed. A disadvantage of the gain functions discussed so far is that all of them cannot be adjusted by compression factors or dumping constants. The computation of the spectral amplitude estimators is bound to the estimation of the *a priori* and *a posteriori* SIRs. Therefore, also the so called parameterized gain function [28] is introduced in the next section.

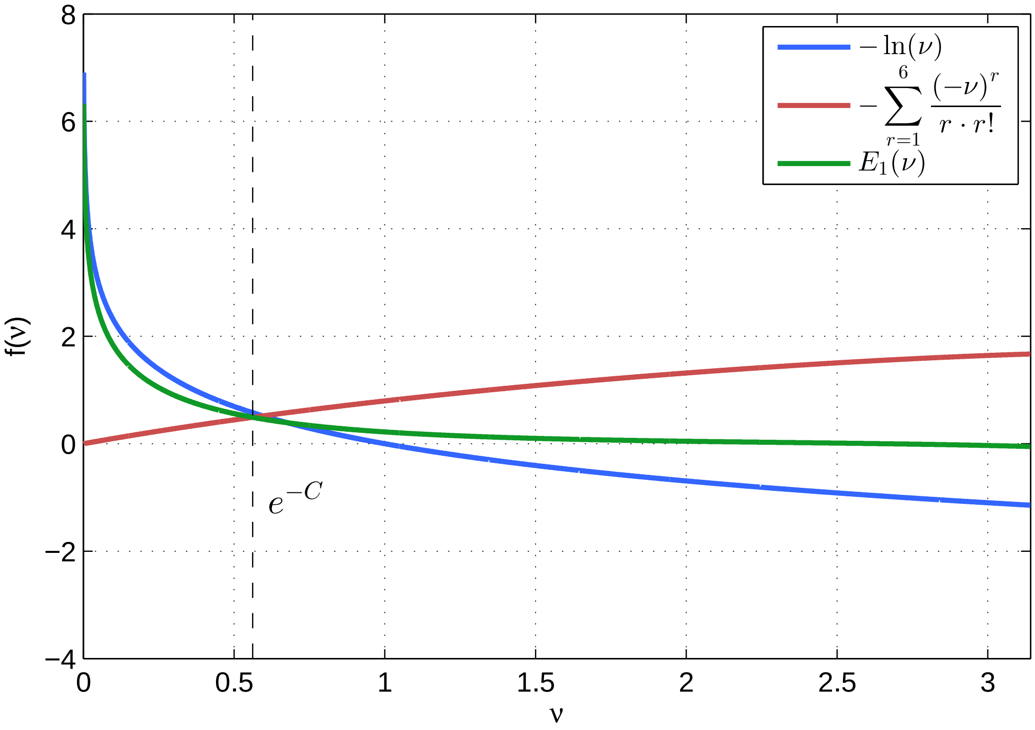


Figure 3.13 – Exponential function $E_1(\nu)$ and its components

3.4.5 Parameterized Gain Function

While most of the algorithms for solving the MMSE criterion are using either compressive weighting functions [25], [29] or statistical models of clean speech, the parameterized Gain function as proposed in [28] combines both of these approaches by calculating a spectral gain in a highly parameterized version. Thus it is possible to simulate many of the common spectral gain functions by choosing the corresponding parameters. Searching a solution for the MMSE criterion under the assumption of a chi-distribution of spectral speech magnitudes and a compression function $c(x) = x^{\beta}$ leads to the Equation given in [28]. With a compression factor β and a shaping factor μ the gain function can then be calculated by

$$G_{PAR}(k,l) = \sqrt{\frac{\hat{\xi}(k,l)}{\mu + \hat{\xi}(k,l)}} \left[\frac{\Gamma\left(\mu + \frac{\beta}{2}\right)}{\Gamma(\mu)} \frac{\Phi\left(1 - \mu - \frac{\beta}{2}, 1; -\nu(k,l)\right)}{\Phi(1-\mu, 1; -\nu(k,l))} \right] \left(\sqrt{\hat{\gamma}(k,l)}\right)^{-1} \tag{3.48}$$

$$\text{with } \nu(k,l) = \frac{\hat{\gamma}(k,l)\hat{\xi}(k,l)}{\mu + \hat{\xi}(k,l)},$$

where $\Gamma(\cdot)$ denotes the complete gamma function and $\Phi(\alpha, \gamma; z)$ the confluent hypergeometric function which is defined by [30]

$$\Phi(\alpha,\gamma;z) = 1 + \frac{\alpha}{\gamma}\frac{z}{1!} + \frac{\alpha(\alpha+1)}{\gamma(\gamma+1)}\frac{z^2}{2!} + \frac{\alpha(\alpha+1)(\alpha+2)}{\gamma(\gamma+1)(\gamma+2)}\frac{z^3}{3!} + \ldots \tag{3.49}$$

The parameterized spectral gain has its advantage in the possibility to simulate many different versions of gain functions and to perform a fine adjustment of the available parameters. For example, by choosing $\beta = 1$ and $\mu = 1$ the gain function simulates the STSA-estimator, whereas $\beta \rightarrow 0$ and $\mu = 1$ simulate the LSA-estimator. However, as we can see from Equation (3.48) it requires high computational complexity due to the computation of the confluent hypergeometric function. It is hard to decide at which instance the hypergeometric series of Equation 3.49 is stopped. It always depends on the input parameters γ and α. Therefore, if a more simple approach i.e. the approximation of Section 3.4.4 can lead to a similar dereverberation result without parameterizing the calculation, it is disputable if the high parameterized gain function is really necessary to compute.

3.4.6 Overview of Spectral Gain Functions

In the last few sections we have introduced a set of practical gain functions which can be used for dereverberation. Table 3.1 shows a summary of the introduced gain functions. All of these equations are based on solving the MMSE-criterion defined in Section 3.4.1 but consist of different probability models of speech and compressing or shaping functions. In Chapter 4 we will try to find out the most suitable gain function for the algorithm depending on computational complexity, possibility of fine-adjustment with compression or shaping factors and production of spectral artifacts or *musical noise* as a minor product.

Approach	**Gain function equation**
Wiener	$G_{Wiener}(k,l) = \frac{\hat{\xi}(k,l)}{1+\hat{\xi}(k,l)}$
MMSE-SA	$G_{SA}(k,l) = \frac{\sqrt{\pi}}{2}\sqrt{\frac{1}{\hat{\gamma}(k,l)}G_{Wiener}(k,l)} \cdot M\left[\hat{\gamma}(k,l)G_{Wiener}(k,l)\right]$
MMSE-LSA	$G_{LSA}(k,l) = G_{Wiener}(k,l) \cdot \exp\left(\frac{1}{2} \cdot E_1(\nu(k,l))\right)$
Approximated	$G_{Approx}(k,l) = \sqrt{G_{Wiener}(k,l) \cdot \left(G_{Wiener}(k,l) + \frac{e^{-C}}{\hat{\gamma}(k,l)}\right)}$
Parameterized	$G_{PAR}(k,l) = \sqrt{\frac{\hat{\xi}(k,l)}{\mu+\hat{\xi}(k,l)}}\left[\frac{\Gamma\left(\mu+\frac{\beta}{2}\right)}{\Gamma(\mu)}\frac{\Phi\left(1-\mu-\frac{\beta}{2},1;-\nu(k,l)\right)}{\Phi(1-\mu,1;-\nu(k,l))}\right]\frac{1}{\sqrt{\hat{\gamma}(k,l)}}$

Table 3.1 – Overview of spectral gain functions

3.4.7 Frequency Smoothed Gain Function

No matter which gain function of Table 3.1 is used or which compression factors are chosen, all of the gain functions produce strong spectral ripples (see Figure 3.14) at any time instance because the estimation of noise, reverberation and dereverberated speech occurs with a very high frequency resolution. These resonances and antiresonances in the spectral domain can produce *musical noise* or annoying distortion in speech reproduction. A sub-band based processing can find a remedy in suppressing such effects and is computationally efficient. However, an estimation of reverberated and dereverberated speech power level by cepstro-temporal smoothing and the MS noise estimation require a high frequency resolution for accurate results. Hence, a sub-band processing is neglected and the smoothing is applied as the last step of calculating the spectral gain function.

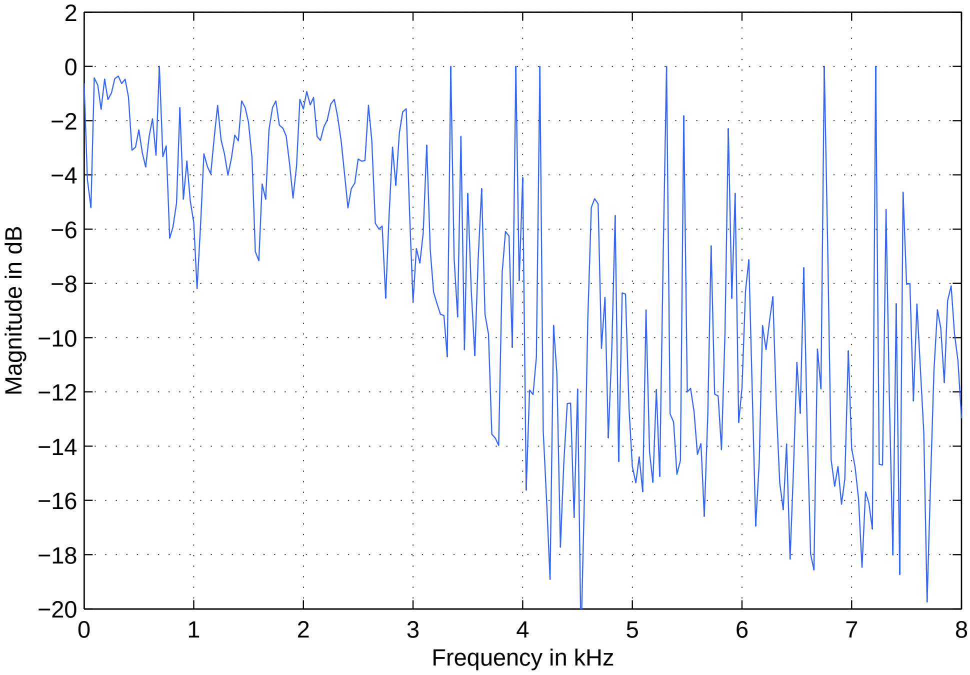


Figure 3.14 – Gain function $G(k, l)$ at a single frame $l = 9$ producing strong ripples in the spectral domain

According to the frequency selectivity of human hearing it is appropriate to perform spectral smoothing in bands. The bark scale [31] divides the human range of audibility into 24 sub-bands whose frequency width corresponds to a third-octave above 500 Hz. For the third-octave smoothing each magnitude value at frequency bin k is replaced by a arithmetic mean of the magnitude values of a half third-octave ($2^{1/6}$) underneath and a half third-octave above that frequency bin. For a higher resolution in the frequency domain, also the octave smoothing is a possible approach. the arithmetic mean then follows a half octave ($2^{1/16}$) underneath and above. For our purposes, the third-octave might overwhelm the frequency selectivity of the octave smoothing.

The third-octave smoothed spectral gain is given by

$$G_{ThSM}(k,l) = \frac{\sum\limits_{k}^{k\cdot 2^{-1/6}\leq k<k\cdot 2^{1/6}} G_{Ori}(k,l)}{\sum\limits_{k}^{k\cdot 2^{-1/6}\leq k<k\cdot 2^{1/6}} 1}, \tag{3.50}$$

where $G_{Ori}(k,l)$ denotes the chosen original gain function from Table 3.1. Whereas the computational complexity of such smoothing is high as we can see from Equation (3.50), the already outlined approach of the exponential smoothing exhibits a more efficient way of gain function smoothing since it is a simple recursive averaging method. Therefore, the exponentially smoothed gain function $G_{ExpSM}(k,l)$ with a constant smoothing parameter α is calculated by

$$G_{ExpSM}(k,l) = \alpha G_{Ori}(k,l) + (1-\alpha)G_{ExpSM}(k-1,l). \tag{3.51}$$

A low smoothing parameter i.e. $\alpha = 0.2$ can lead to an approximately third-octave smoothed spectral gain as we can see in Figure 3.15 for a single frame $l = 9$.

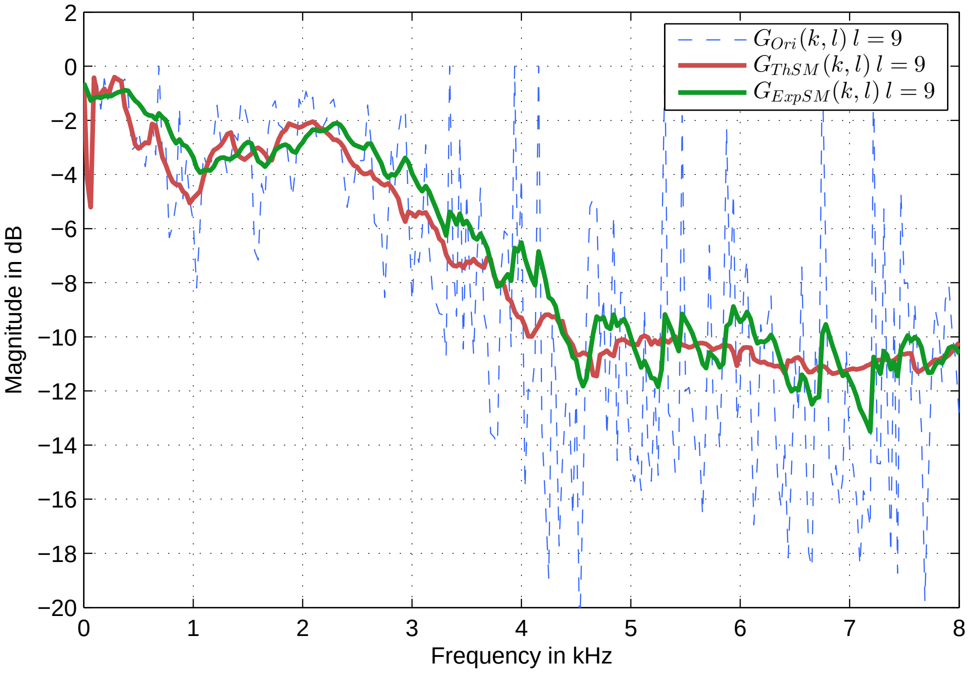


Figure 3.15 – Original gain function $G_{Ori}(k,l)$, third-octave smoothed gain function $G_{ThSM}(k,l)$ and exponentially smoothed gain function $G_{ExpSM}(k,l)$ with $\alpha = 0.1$ at a single frame $l = 9$

Since $G_{ExpSM}(k,l)$ applies a frequency-independent smoothing on the gain function, $G_{ThSM}(k,l)$ produces a frequency-dependent smoothing in which the strength of the smoothing is constant in each octave band which better matches the human hearing. However, the difference between the two smoothing variants appeared negligible in the parameter adjustment by ear and thus the computational efficiency of exponential smoothing was preferred to the accuracy of octave band smoothing.

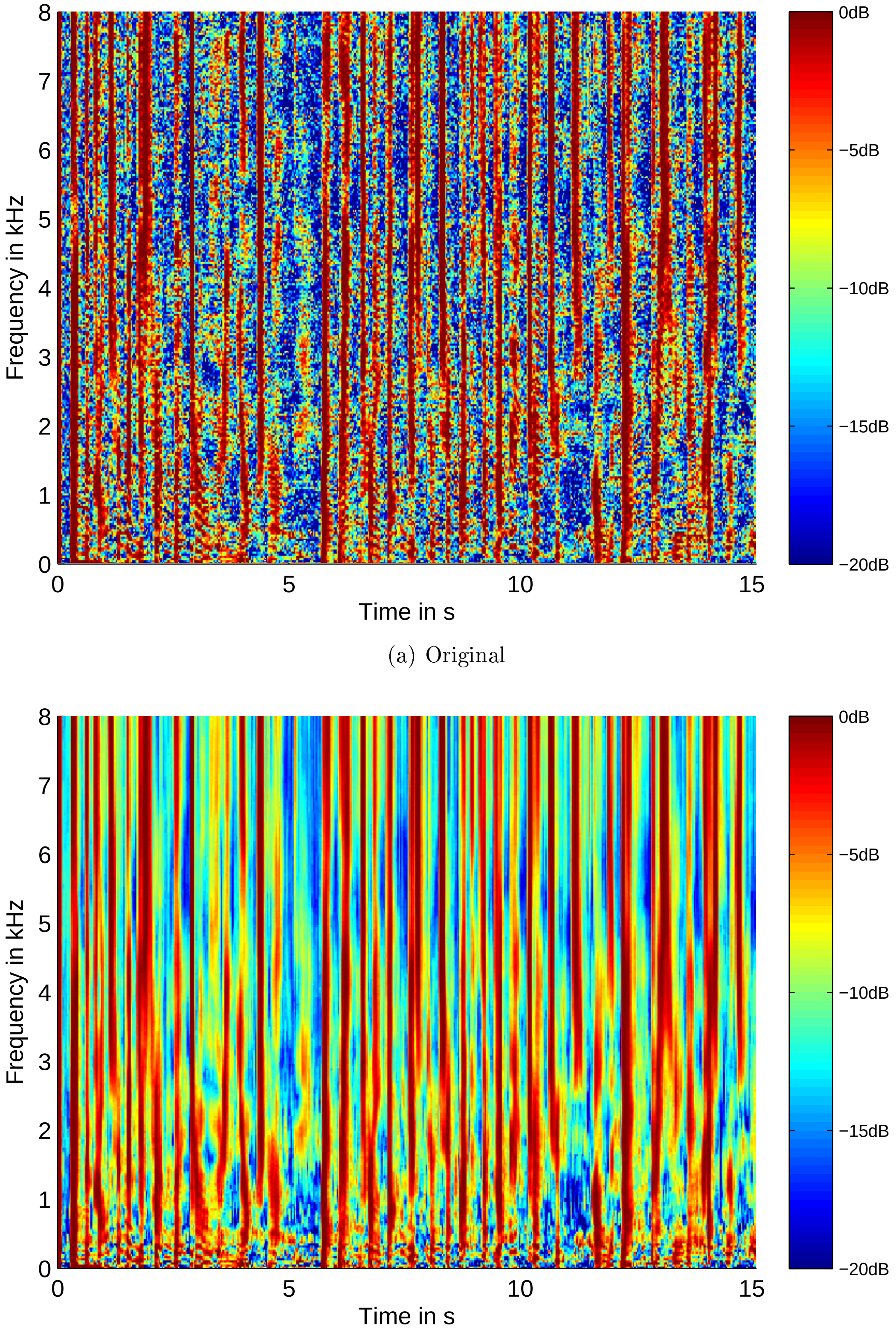
In Figure 3.16, an original approximated LSA gain function compared to the exponentially smoothed gain function is shown. Due to the frequency smoothing and the preceding cepstro-temporal smoothing done during the calculation of the gain function, there is no more hard decision suppression in $G_{ExpSM}(k,l)$ that can produce strong magnitude fluctuations. However, there are still parts in which $G_{ExpSM}(k,l)$ reaches -20 dB suppression and can therefore produce *musical noise*. In order to reduce such artifacts, $G_{ExpSM}(k,l)$ is floored by $\tilde{G}(k,l)$

$$\tilde{G}(k,l) = \max\left(G_{ExpSM}(k,l), G_{min}\right). \tag{3.52}$$

Typically $G_{min} = -10$ dB, but variations will also be evaluated in Chapter 4. Additionally to prevent computational errors which can occur during the computation of the *a priori* and *a posteriori* SIRs, also a ceiling of $\tilde{G}(k,l)$ to $G_{max} = 0$ dB as a maximum value is introduced

$$\hat{G}(k,l) = \min\left(\tilde{G}(k,l), G_{max}\right). \tag{3.53}$$

The final gain function $\hat{G}(k,l)$ is then used to obtain the dereverberated short-time spectrum $\hat{S}(k,l)$ by multiplying $\hat{G}(k,l)$ with the reverberated short-time spectrum $X(k,l)$.



(a) Original

(b) Exponential smoothed

Figure 3.16 – Comparison between original and exponentially smoothed gain function (approximated LSA)

3.5 Summary

In this Chapter, a system for spectral subtraction which jointly suppresses noise and reverberation was introduced. Furthermore, it is robust to noise at reverberation estimation and vice versa. As we have already announced in Section 3.2 the MS approach offers good enough noise estimation results for our purposes. As shown in Section 3.3.2, an estimation of the reverberation time is not crucial for an appropriate dereverberation, whereas the reverberation estimation is based on Equation (3.16). The cepstro-temporal smoothing was discussed as an estimation of the reverberated as well as the dereverberated speech.

Additionally an exponential smoothing in the frequency domain on the calculated gain function is applied. Since all gain functions are dependent on an estimation of the *a priori* and *a posteriori* SIR we do not determine a specific gain function yet. According to these definitions the complete spectral subtraction dereverberation system is schematically described in Figure 3.17. The spectral subtraction procedure in Figure 3.18 shows the clean speech signal spectrogram of a male speaker at 16 kHz as initial point. The kitchen RIR [14] was used for the reverberated speech as we can see in the middle spectrogram. The last spectrogram corresponds to the dereverberated speech signal. It can be noticed that a spectral suppression in specific frequency regions takes place and that there is a reasonable improvement in the dereverberated speech spectrogram.

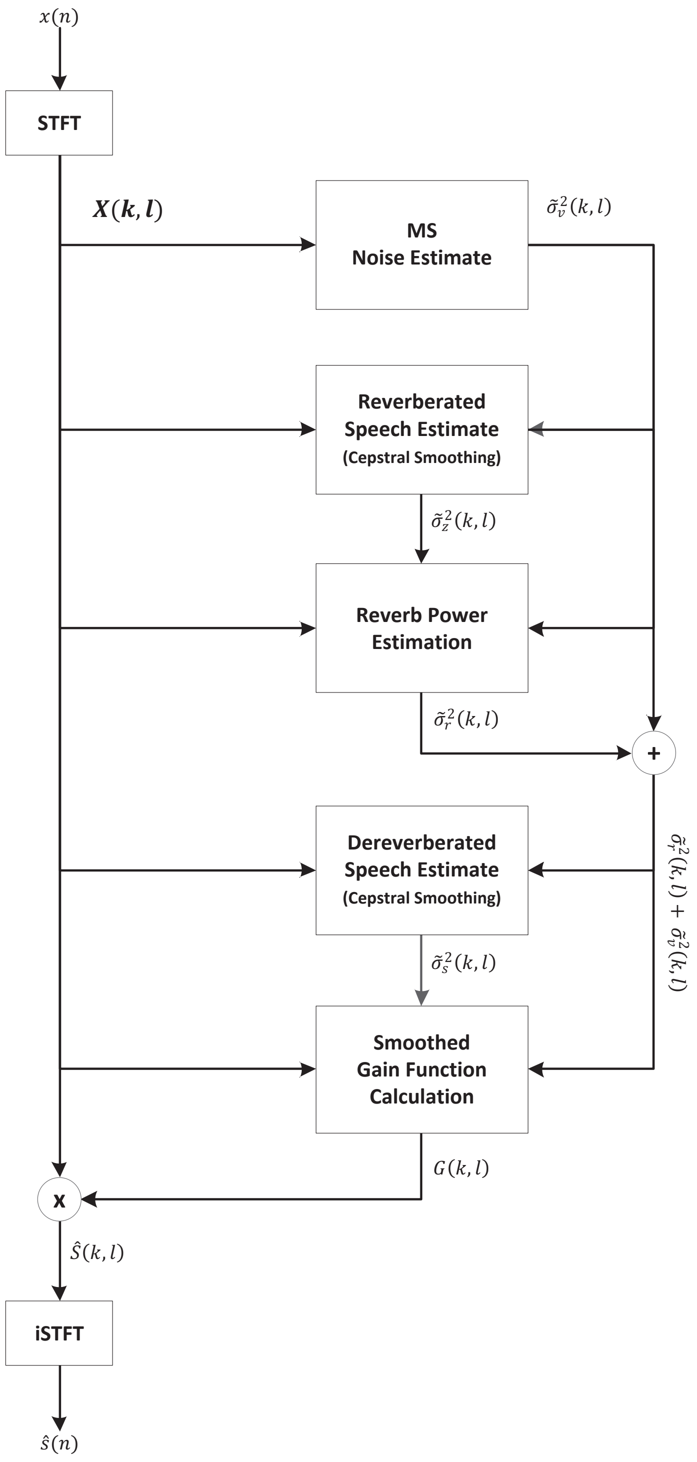
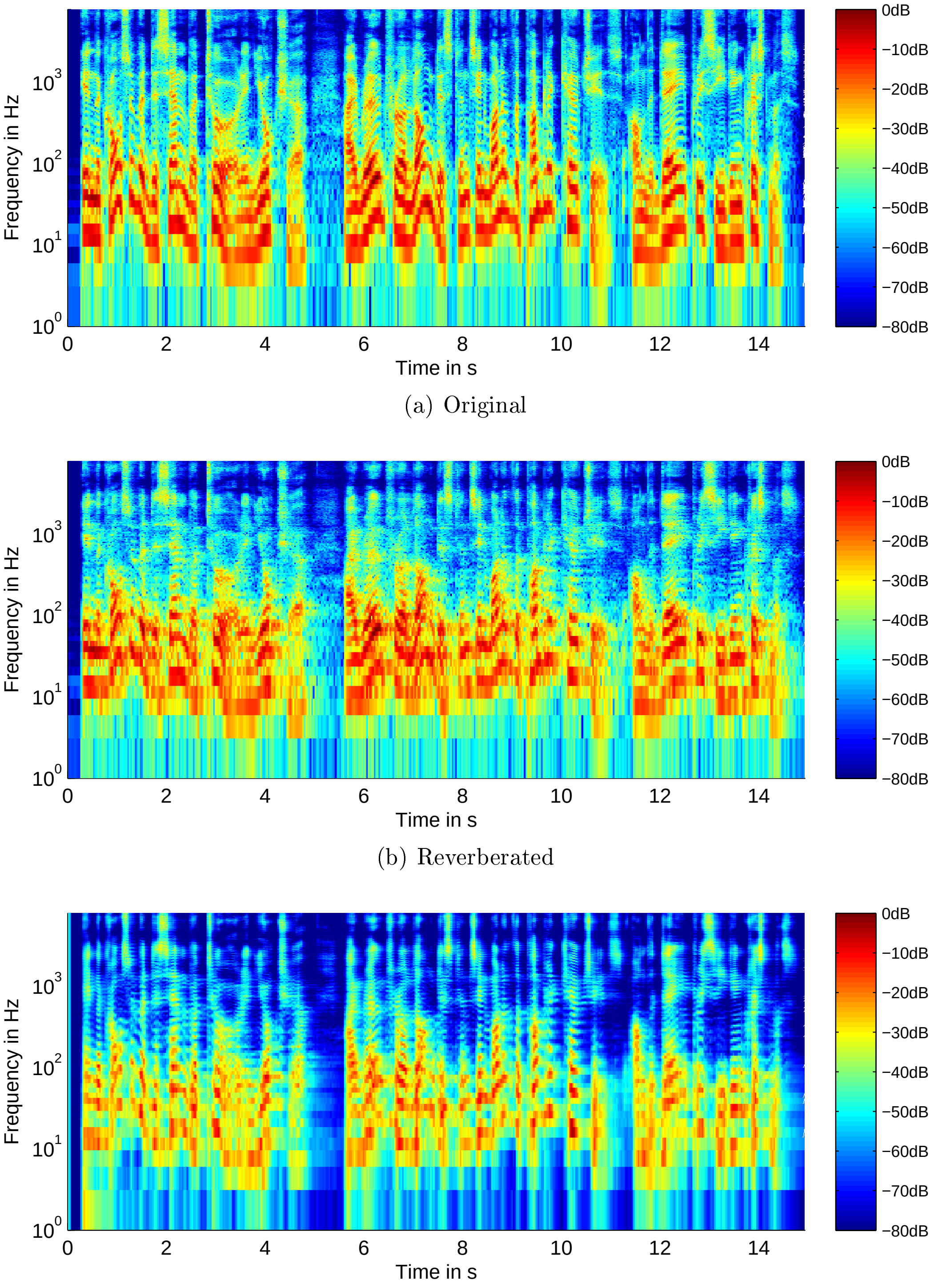

$x(n)$
STFT
$X(k,l)$
MS Noise Estimate
$\tilde{\sigma}_v^2(k,l)$
Reverberated Speech Estimate (Cepstral Smoothing)
$\tilde{\sigma}_z^2(k,l)$
Reverb Power Estimation
$\tilde{\sigma}_r^2(k,l)$
+
$\tilde{\sigma}_r^2(k,l)+\tilde{\sigma}_v^2(k,l)$
Dereverberated Speech Estimate (Cepstral Smoothing)
$\tilde{\sigma}_s^2(k,l)$
Smoothed Gain Function Calculation
$G(k,l)$
X
$\hat{S}(k,l)$
iSTFT
$\hat{s}(n)$

Figure 3.17 – Schematic overview of the final dereverberation system



(a) Original

(b) Reverberated

(c) Dereverberated

Figure 3.18 – Spectrograms of the dereverberation procedure of a male speaker @16 kHz recorded in the kitchen RIR [14]

Chapter 4

Algorithm Evaluation

This Chapter discusses the behavior of the algorithm using various parameter sets and acoustic environments. Therefore, technical quality metrics are introduced that provide a rough overview of how the algorithm performs. Furthermore, a multi-stimulus test was undertaken with several persons to provide a rough perceptual evaluation. For the evaluation, the source signal $s(n)$ as well as the corresponding interference $r(n) + v(n)$ defined by the RIR and noise are known.

4.1 Evaluation by Technical Quality Metrics

4.1.1 Signal-to-Interference Ratio

To evaluate the efficiency of the algorithm according to the relation between signal and containing amount of reverberation and noise, the signal-to-interference ratio (SIR) can be used. In general the SIR is defined as the ratio between the clean speech signal power P_s and the reverberating power P_r plus noise power P_v . In the logarithmic scale it is given by

$$\mathrm{SIR} = 10 \log_{10} \left(\frac{P_s}{P_r + P_v} \right) \mathrm{dB}. \tag{4.1}$$

The ratio between the power values can therefore be calculated by the quadratic Euclidean norm of the absolute signal values [32]. Under our assumptions of Chapter 3, reverberation is treated as a noisy part of the signal, where $r(n) + v(n) = x(n) - s(n)$ holds. The SIR is therefore given by

$$\mathrm{SIR} = 10 \log_{10} \frac{\sum\limits_{n=0}^{N-1} s^2(n)}{\sum\limits_{n=0}^{N-1} (x(n) - s(n))^2} \mathrm{dB}, \tag{4.2}$$

where N denotes the signal length. For a representative measurement, we first use Equation (4.2) to obtain the SIR for the reverberant an noisy signal $x(n)$ and then replace $x(n)$ with the dereverberated signal $\hat{s}(n)$ to obtain a second SIR. Comparing these two measures provides information about the quality of the joint dereverberation algorithm. In most cases, there are parts of the signal that contain speech pauses with low amplitude corresponding to noise. Therefore, the SIR tends towards infinite. It is suitable to use a so called segmental SIR which extracts defined frames $\mathbb{L}$ of the speech sample. The average of the SIR values is given by

$$\mathrm{SIR}_{seg} = \frac{1}{|\mathbb{L}|} \cdot \sum_{l \epsilon \mathbb{L}} \mathrm{SIR}(l), \tag{4.3}$$

where $\mathbb{L}$ is a set of appropriately chosen frames containing enough signal energy for an SIR computation. A proper choice considers frames in which the SIR is above a lower threshold (i.e. -10dB) and saturated at an upper threshold (i.e. $+30$dB). Averaging the SIR over frequency bands smoothes out the SIR to what might be relevant for human hearing. The frequency-weighted segmental SIR (FWSegSIR) [33] is calculated by

$$\mathrm{SIR}_{FWSeg} = \frac{1}{|\mathbb{L}|} \sum_{l \epsilon \mathbb{L}} \frac{1}{W_l} \sum_{k=1}^{K^*} 10 \log_{10} \left[\frac{w_{k,l} \cdot \sum s^2(n)}{\sum \left(x(n) - s(n)\right)^2} \right], \tag{4.4}$$

where $W_l = \sum_{k=1}^{K^*}(w_{k,l})$ is the sum of the interference-dependent weights $w_{k,l}$ over all frequency bands K^*. The FWSegSIR will be used as a suitable quality metric in Chapter 4 by taking into account that a higher FWSegSIR indicates a better quality of speech.

4.1.2 Speech-to-Reverberation Modulation Engery Ratio

This section presents an technical metric called speech-to-reverberation modulation energy ratio (SRMR). The SRMR is a so called non-intrusive measure which means that the quality rating does not depend on a distance measure between the clean source signal and the reverberant or dereverberated counterpart. Therefore, it is often used as a measurement utility for speech dereverberation methods. As suggested in [34], this metric includes standardized quality measurements such as estimated coloration, reverberation fail effects, and overall quality. Furthermore it is also used as a quality metric in the REVERB-Challenge [35], hence it is suitable to allow comparative evaluation.

The SRMR is based on the gammatone modulation spectral energy, which can be calculated by a discrete Fourier transform of the gammatone filtered temporal envelope. A gammatone filter bank with $J = 23$ channels is used to represent the frequency-space transformation in the cochlea [36]. The modulation frequencies are grouped to $K^* = 8$ bands to emulate an auditory modulation filter bank [37].

The gammatone filtered, band-limited modulation spectral energy $\varepsilon_{j,k}(l)$ is averaged over all frames to get the so called modulation spectrogram $\overline{\varepsilon}_{j,k}$. The SRMR is defined by [34]

$$\text{SRMR} = \frac{\sum\limits_{k=1}^{4}\sum\limits_{j=1}^{J}\overline{\varepsilon}_{j,k}}{\sum\limits_{k=5}^{K^*}\sum\limits_{j=1}^{J}\overline{\varepsilon}_{j,k}}, \tag{4.5}$$

where K^* depends on the signal and denotes the number of gammatone filters that contain approximately 90% of the total signal energy. For example, if the first 8 gammatone filters contain 90% of the total modulation energy, $K^* = 8$. This allows a signal-dependent calculation of the SRMR. The higher a SRMR value of a signal, the better the performance of the dereverberation algorithm.

4.1.3 Cepstral Distance

Since speech segments are mainly characterized by the cepstral coefficients at lower quefrencies [38], it is reasonable to measure the Euclidean distance $d_c(l)$ between the first cepstral coefficients of the clean speech real cepstrum $c_s(q,l)$ and the dereverberated real cepstrum $c_{\hat{s}}(q,l)$. The calculation of the cepstral distance (CD) for a single frame l is given by

$$d_c(l) = \sqrt{\sum_{q=1}^{Q_{max}} \left(c_s(q,l) - c_{\hat{s}}(q,l)\right)^2}, \tag{4.6}$$

where Q_{max} is the maximum number of cepstral coefficients i.e. $Q_{max} = \frac{Q}{4}$ used for the distance measure. For the evaluation it is necessary to compute a mean over all frames to obtain an overall evaluation result. The mean cepstral distance d_c^{mean} is calculated as

$$d_c^{mean} = \frac{1}{Q} \cdot \sum_{l=1}^{L} d_c(l). \tag{4.7}$$

The lower the distance, the more similar are the clean and the dereverberated cepstral coefficients and thus the signals themselves. This distance measure is therefore a powerful indicator for the performance of the algorithm.

4.2 Evaluation by Perceptual Multi Stimulus Test

The multi-stimulus test is based on the approach of the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test which is an subjective evaluation method for audio signals. It was primarily developed to evaluate the perceived quality of audio compression algorithms in telephone systems. In comparison to the older Mean Opinion Source (MOS) test, the MUSHRA requires fewer participants because all experiments are made out of just one sample in different situations which allows us to get statistically significant results [39].

4.2.1 Experimental Setup

The anchor of MUSHRA is generally defined as a low-passed version of the original signal, which is not suitable for our evaluation study, in which the perceived amount of reverberation and sound quality as the main attributes. Therefore, the anchor can be neglected in this test. The main goal of the multiple stimuli test is to identify how good the dereverberation algorithm works compared to the reverberated speech. Thus, in addition to the dereverberated signal, the reverberated speech is also evaluated by each listener.

Condition	T_d **and** T_{60}	**Smoothing**	**Gain Function**
Condition 1	$T_d = 32$ ms , $T_{60} = 0.3$ s	Yes	MMSE-LSA
Condition 2	$T_d = 32$ ms , $T_{60} = 0.3$ s	Yes	Approximated
Condition 3	$T_d = 32$ ms , $T_{60} = 0.3$ s	No	MMSE-LSA
Condition 4	$T_d = 32$ ms , $T_{60} = 0.3$ s	No	Approximated
Condition 5	$T_d = 64$ ms , $T_{60} = 0.5$ s	Yes	MMSE-LSA
Condition 6	$T_d = 64$ ms , $T_{60} = 0.5$ s	Yes	Approximated
Condition 7	$T_d = 64$ ms , $T_{60} = 0.5$ s	No	MMSE-LSA
Condition 8	$T_d = 64$ ms , $T_{60} = 0.5$ s	No	Approximated

Table 4.1 – Overview of multi stimulus conditions

The main test evaluates three different fundamental algorithm settings for which the parameter sets were predefined as good as possible to focus on practically relevant settings of the evaluation. The value of T_d in Equation (3.15) is crucial for the perceived amount of reverberation and depends on the size of the room in which the RIR was recorded.

This also involves the adaption of the parameter T_{60}. While we have seen that for dereverberation in small room acoustics, the estimation of the reverberation time is not crucial, the algorithm is also tested with larger rooms to test the robustness under marginal conditions. Thus T_{60} is adopted to T_d in those specific setups. Further, the smoothing of the real-valued gain function as described in Section 3.4.7 is used as a test value to screen if the listeners prefer quality improvements by suppressing *musical noise* rather than a lower reverberation, as it is an adjustable trade-off by using the smoothing function. As a last setting the difference between the most suitable gain functions in Table 3.1 are tested.The MMSE-LSA and the approximated MMSE-LSA were found to give the best dereverberation results in terms of reverberation and suppression at a low amount of *musical noise* in the parameter adjustment phase during the preparation of the experimental conditions. The goal is to find out if it is necessary to calculate the exponential integral of the LSA algorithm or if the simple calculation of the approximated LSA is sufficient for the dereverberation. These three testing parameters (T_d and T_{60}, Smoothing, Gain Function) led to the eight experimental conditions listed in Table 4.1.

These conditions are tested in three different rooms extracted from the RWTH Aachen RIR database [14]. The room size increases from small to large.

— **Meeting Room** defines a small room with short reverberation and reproduces the acoustic behavior of different speakers during a meeting.
The average reverberation time is $\overline{T}_{60} = 0.23s$

— **Office** defines a medium large room with medium reverberation and is a typical office room with standard office furniture.
The average reverberation time is $\overline{T}_{60} = 0.43s$

— **Lecture Room** defines the largest room with long reverberation and typical auditorium furniture such as desks and chairs.
The average reverberation time is $\overline{T}_{60} = 0.78s$

Basically the evaluation is based on comparisons of a clean reference sound and a number of dereverberated test sounds under the usage of the conditions described in Table 4.1 plus the reverberated sound. Two attributes are used for the evaluation:

— **Reverberation Level:** This attribute is related to the perceived amount of reverberation caused by a high ratio of reflected to direct sound energy depending on the room size which leads to the impression of a high diffusivity. Example: The perceived level of reverberation differs significantly between rather small and very large spaces, such as living rooms and churches.

— **Overall quality:** This attribute is used to judge any and all detected differences (in terms of characteristics of reverberation, additive noise, and processing distortion, timbrel characteristics, naturalness, and so on) between the reference and test sound excerpts. Typical of low overall quality: train station announcements. Impression of how well the words of a speaker can be understood. Typical for high overall quality: Newscaster.

In the next sections we will discuss the proposed multi stimulus test procedure on the basis of the MUSHRA approach which contains a training and evaluation phase.

4.2.2 Training Phase

The training phase is intended to allow the listener to obtain an idea of the whole range of the possible qualities, reverberation and noise that can be experienced in the grading phase. The listener is asked to listen to all sound excerpts that have been selected. The training phase requests the listener to adjust the volume of headphones, which should not be change during the grading phase to ensure consistency. As soon as the listener is aware of all test sounds in a certain experiment, the listening test continues with the grading phase.

4.2.3 Grading Phase

In the grading phase the listener is invited to evaluate all the sound files from the training phase divided into different experiments depending on the different reference signals. The rating should reflect the subjective judgment of the quality and reverberation level for each of the sound excerpts presented. As shown in Figure 4.1 the grading phase consists of two sliders for each sound item. One for the perceived amount of reverberation and one for overall quality. The listener is able to evaluate each sound file compared to the reference on top from 0 (=Bad) to 100 (=Excellent) for the overall quality and from 0 (=High) to 100 (=low) for the perceived reverberation level. The listener is able to listen to all sound files as often and long as required to provide a thorough response.

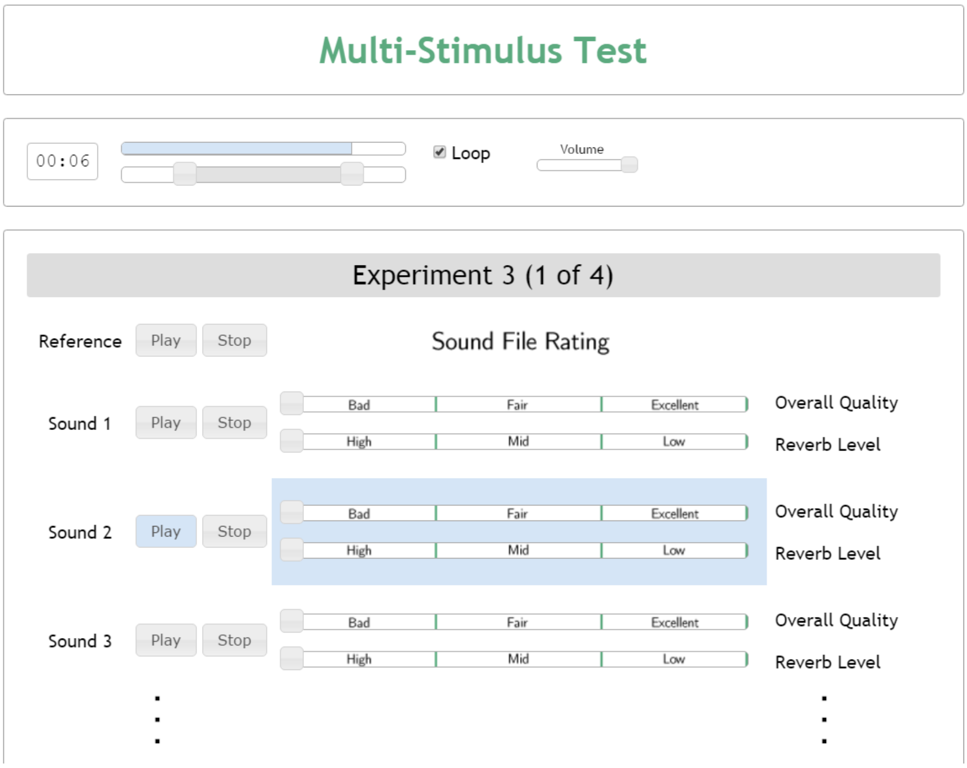


Figure 4.1 – Screenshot of the multi stimulus dereverberation test

4.2.4 Analysis of Results

A meaningful evaluation of the results is subject to a careful inspection of the data, first. Therefore, the procedure of correctly analyzing the scores given by the listeners is presented in this section. To verify if the listeners scores are reliable, the office room is tested twice during the multi-stimulus test without knowledge of the listener. If the discrepancy between the first and the second test score during the office room tests is higher than 20% at any instance, all of the listeners scores are excluded from the analysis.

The remaining results are analyzed using the method proposed in [40]. First, the mean score as an average of the scores of I listeners is calculated by

$$\overline{u}_{jk} = \frac{1}{I}\sum_{i=1}^{I} u_{ijk} \tag{4.8}$$

for a test condition j and audio sequence k. To present the overall results of the test, an associated, 95% confidence interval which is given by

$$[\overline{u}_{jk} - \delta_{jk}, \overline{u}_{jk} + \delta_{jk}] \tag{4.9}$$

is calculated. The value δ_{jk} is given by

$$\delta_{jk} = t_{0.05}\frac{S_{jk}}{\sqrt{I}}, \tag{4.10}$$

where S_{jk} denotes the standard deviation for each condition and audio sequence. It is usually calculated by

$$S_{jk} = \sqrt{\sum_{i=1}^{I} \frac{(\overline{u}_{jk} - u_{ijk})^2}{I-1}}. \tag{4.11}$$

The value $t_{0.05} = 0.05$ is used to obtain the usual significance level of 95% as proposed in [40]. The confidence interval of Equation 4.9 is then illustrated with box plots to show the upper and lower bounds $\pm\delta_{jk}$ of the confidence interval.

4.3 Evaluation Results

This section discusses the evaluation results obtained from different quality measures and the above mentioned multi stimulus test. These results help us to adjust the algorithm as its best and to find out if the dereverberation system is suitable for practical applications.

4.3.1 Algorithm Settings

First, an overview of the parameter settings used for the evaluation of the technical measures is given and the listening test in general. The STFT is based on a 32 ms root-Hann window with 50% overlap, wherefore the block length is also 32 ms long. The noise estimation is made by the MS algorithm from Section 3.2.1. The used sliding window length to $T_{SL} = 3s$ and the maximum allowable smoothing constant is set to $\alpha_{max} = 0.95$.

The reverberation power estimation is based on the extended reverberation model of Equation (3.16), where $T_{60} = 0.4$ s is assumed in general. The set of prediction delays is chosen as $\mathbb{T}_d = \{32 \text{ ms}, 64 \text{ ms}, 128 \text{ ms}\}$. For the *a priori* speech power and dereverberated speech power estimation, the cepstro-temporal smoothing from Section 3.3.3 is employed. The parameter settings are shown in Table 4.2

$f_{0,low} = 70$ Hz	$\alpha_{pitch} = 0.2$
$f_{0,high} = 300$ Hz	$\beta = 0.95$
$\Lambda^{thr} = 0.4$	$\kappa = 0.2886$
$\Delta q_{pitch} = 2$	$\xi^{ml}_{min} = 0.0022$

$$\overline{\alpha}_{const}(q) = \begin{cases} 0.3 & \text{if } q \,\epsilon\, \{0, ..., \hat{q}_{low}\} \\ 0.65 & \text{if } q \,\epsilon\, \{\hat{q}_{low}+1, ..., \hat{q}_{mid}\} \\ 0.97 & \text{if } q \,\epsilon\, \{\hat{q}_{mid}+1, ..., N\} \end{cases},$$

Table 4.2 – Parameter settings for the cepstro-temporal smoothing

where $\hat{q}_{low} = \frac{95 \text{ ms}\cdot f_s}{N}$ is the lower and $\hat{q}_{mid} = \frac{650 \text{ ms}\cdot f_s}{N}$ the middle bound for $\overline{\alpha}_{const}(q)$. The gain function employed is varied in the conditions of the evaluation itself. As a reference, the approximated MMSE-LSA is used to evaluate optimal parameters. The minimum Gain G_{min} is usually -10 dB but is also varied. The maximum gain G_{max} is 0 dB. The additional exponential smoothing in the frequency domain is also applied as a standard with a smoothing factor $\alpha = 0.2$. With these settings, first a technical quality evaluation is done for the algorithm with different RIRs and parameter sets. An overview can be seen in Table 4.3, where the cepstro-temporal smoothing is denoted as "Ceps Smooth" and the smoothing of the gain function as "Gain Smoothing". Third-octave and octave denote the band dependent filter while exponential corresponds to the smoothing from Equation (3.51).

Furthermore, the speech files are English and German men and women at a sample rate of 48kHz or 16kHz. The rooms are extracted from the RIR database [14] and the REVERB-Challenge [35].

Parameter	Test Value				
G_{min}	-15 dB	-12 dB	-9 dB	-6 dB	-3 dB
Gain Function	G_{Approx}	G_{SA}	G_{LSA}	G_{Wiener}	G_{PAR}
Gain Smoothing	None	Third-Octave	Octave	Exponen.	-
T_d	32 ms	64 ms	128 ms	-	-
	$@T_{60} =$ 0.3 s	$@T_{60} =$ 0.5 s	$@T_{60} =$ 0.7 s	-	-

File	Room Impulse Response	
M. Eng. @16kHz	Kitchen [14]	SimRoom 1 Far [35]
W. Eng. @16kHz	Office [14]	SimRoom 1 Near [35]
M. Ger. @48kHz	Meeting Room [14]	SimRoom 2 Far [35]
W. Ger. @48kHz	Lecture Room [14]	SimRoom 2 Near [35]
Various @48kHz	Bathroom [14]	-

Table 4.3 – Overview of the possible test setups for the technical quality metrics evaluation

4.3.2 Results Technical Quality Metrics

This section discusses the results of the quality measures according to the settings of Section 4.3.1. First, the most important parameter G_{min} which is able to set the strength of the dereverberation algorithm. For example if $G_{min} = 0$ dB, the algorithm will do nothing for $G_{max} = 0$ dB. If G_{min} is set gradually lower, the algorithm will suppress noise and reverberation more and more until the lowest gain function level is reached. Figure 4.2 shows the quality measure values for the male English speaker with kitchen as RIR as a function of the minimum gain.

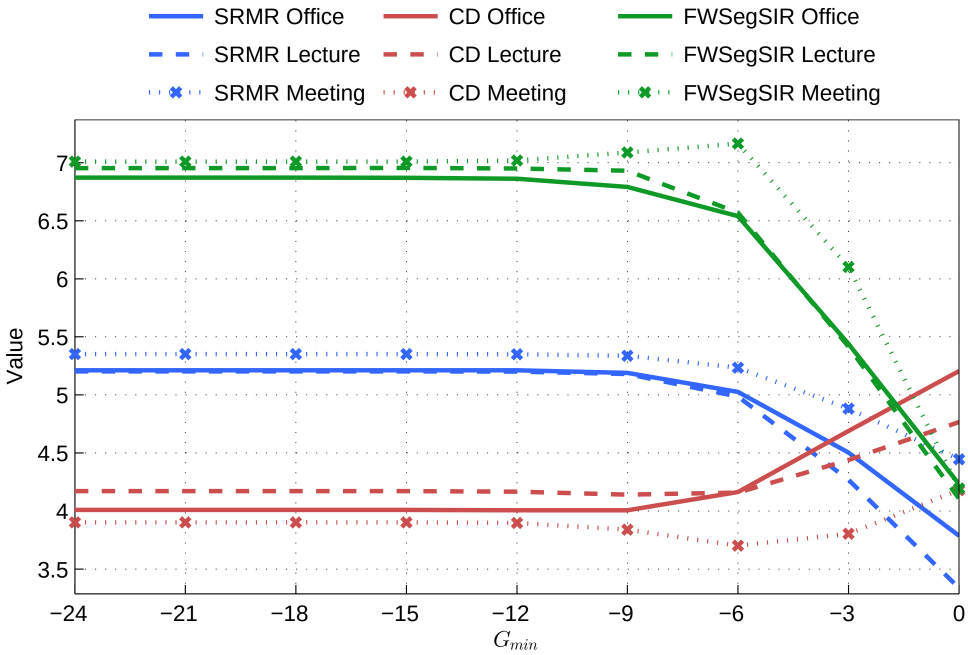


Figure 4.2 – Quality measure of G_{min} with M. Eng. @16kHz as sound file and different RIRs

As illustrated in Figure 4.2, an increase of G_{min} does not show any appreciable changes on the quality measures until −10 dB. After that, the SRMR as well as the FWSegSIR decrease with an increasing CD with lecture and office as RIR which is comprehensible because of the above explanation. In the first approach this could mean that a lower G_{min} is always better for dereverberation. However, by looking at the meeting RIR evaluation, the quality values do not always exhibit this behavior so that the best value for the meeting room is $G_{min} = -6$ dB because i.e. the higher *musical noise* at lower G_{min} can also cause worse quality values. However, it could be found from other measures (see Appendix A) that an optimal G_{min} can be around −10 dB. This is where all quality measures have their best values averaged over all RIRs and sound files of Table 4.3. This might be a good compromise between reverberation level and overall quality. Even if G_{min} can be manually set by the user in future, it is advantageous to set an initial value.

Another interesting measurement depends on the gain function employed. Since some of them are of high computational complexity while others do not offer a good dereverberation or too high *musical noise* artifacts, it is crucial to find an appropriate gain function that balances the overall quality and the reverberation level. Figure 4.3 shows the quality measure as a function of the gain function employed.

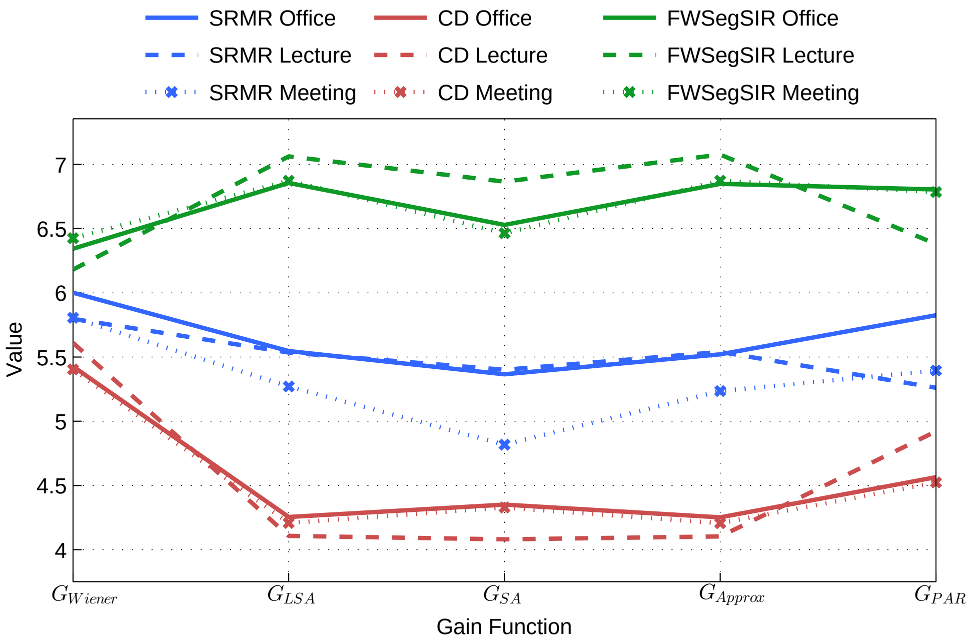


Figure 4.3 – Quality measure of gain functions with M. Eng. @16kHz as sound file and different RIRs

By comparing the different gain functions it can be seen that the Wiener gain function is not suitable for the displayed situations because of the high CD and the low FWSegSIR values. The SRMR might be higher because of the higher reverberation suppression but at cost of intense *musical noise* which is strongly perceivable. The MMSE-SA algorithm is slightly worse than the MMSE-LSA or the approximated MMSE-LSA, which is as expected if the descriptions in Section 3.4.3 are considered. An interesting observation is that the approximated MMSE-LSA exhibits nearly the same values as the MMSE-LSA but with much lower computational complexity. Additionally these two approaches offer the lowest values in general. The parameterized gain function is hard to evaluate because of the potentiality to simulate all other gain functions of Table 3.1. However, this evaluation used the parameter values which were determined properly as proposed in [10] ($\beta = 0.9$ and $\mu = 0.1$). In general the choice of the gain function doesn't us a better dereverberation result in terms of the quality measures. The results of Figure 4.3 are representative for most of the situations of Table 4.3 because the tendencies are approximately the same. For the multi-stimulus test it may be interesting to evaluate both, the MMSE-LSA and the approximated MMSE-LSA gain function to find out which is the most suitable one for the dereverberation algorithm.

For investigating the influence of the choice which part of the reverberation is suppressed, T_d is also evaluated. Since T_d is directly bounded to the frame length, it only makes sense to test T_d values as multiples of the frame length. Thus, if a frame length of 32 ms is defined, a set of $\mathbb{T}_d = \{32 \text{ ms}, 64 \text{ ms}, 128 \text{ ms}\}$ is investigated. While 32 ms and 64 ms correspond to early reflections, 128 ms merges into the late reverberation. However, if the extended reverberation model from Equation (3.16) is used, the set $\mathbb{T}_d$ may be included into the reverberation estimation anyway. It is meaningful to set only the initial value T_{d1} to one of $\mathbb{T}_d$. The arrival of early reflections also depends on the room size. By assuming $T_{d1} = 32$ ms, the assumed room is smaller than with $T_{d1} = 64$ ms. It is suitable to assume a higher T_{60} for higher T_{d1} values as in the multi stimulus test. With the settings from Table 4.3, the evaluation of T_d and T_{60} can be seen in Figure 4.4.

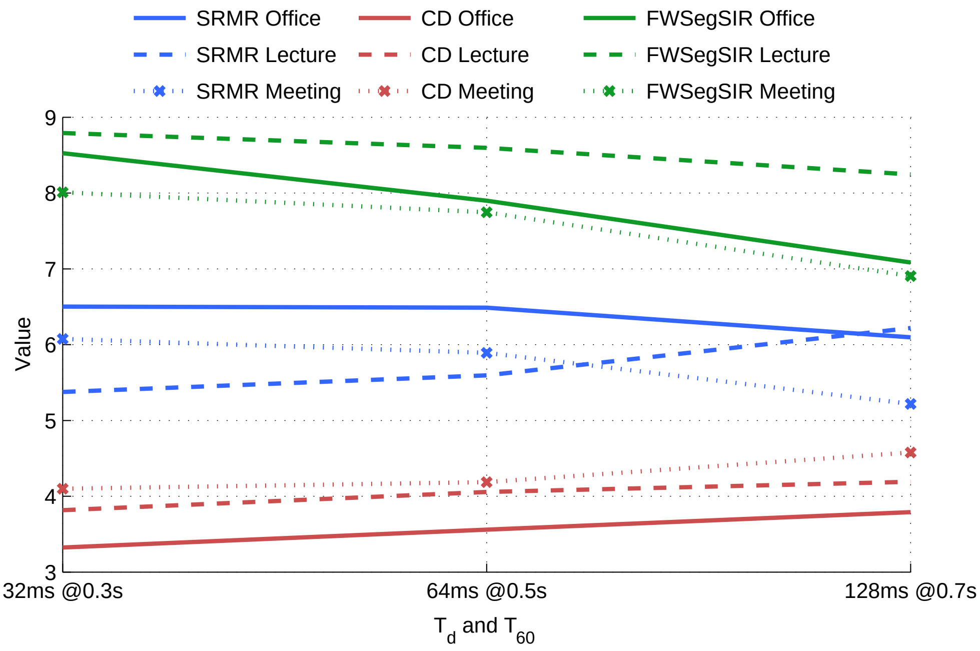


Figure 4.4 – Quality measure of T_d and T_{60} with M. Eng. @16kHz as sound file and different RIRs

Whereas the cepstral distance always rises room-independently with T_d and T_{60}, clear differences in the SRMRs can be seen. This behavior is also detectable in Appendix A. The best SRMR value for the small meeting room is considerably at $T_{60} = 0.3$ s and $T_d = 32$ ms while the best SRMR value for the large lecture room is at $T_{60} = 0.7$ s and $T_d = 128$ ms as expected. The middle value $T_{60} = 0.5$ s and $T_d = 64$ fits directly in between and completes the linearity. Since the dimensions of the room where the recording takes place are typically unknown, the reverberation time is set to $T_{60} = 0.4$ s as an average value as defined in Section 3.3.2. For the extended reverberation model, using $T_d = 32$ ms achieves dereverberation concentrated on early reflections.

The last technical quality evaluation demonstrates the effect of spectral gain function smoothing as described in Section 3.4.7. The setup consists of four different versions of gain functions. The unsmoothed gain function, the third-octave and octave band smoothed, and the exponential smoothed gain function. While the third-octave and octave processing is band dependent and requires high computational resources, the exponential smoothing is simple to apply and can approximate the effectiveness of the band dependent smoothing methods in a certain way. To evaluate the quality of the different approaches, they are considered in the quality evaluation as well as in the multi-stimulus listening test. Figure 4.4 shows the technical metrics for four different smoothing versions.

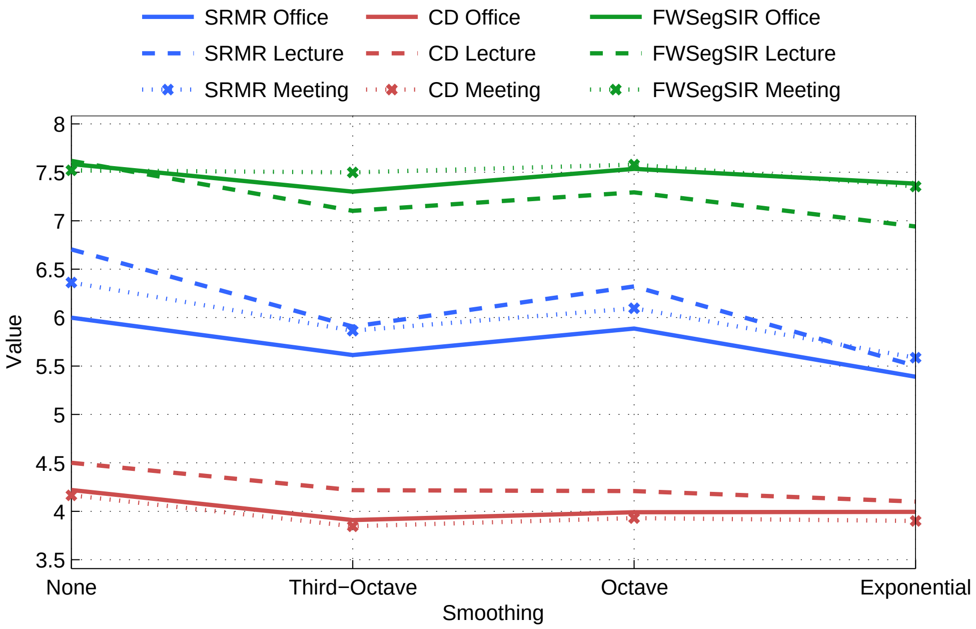
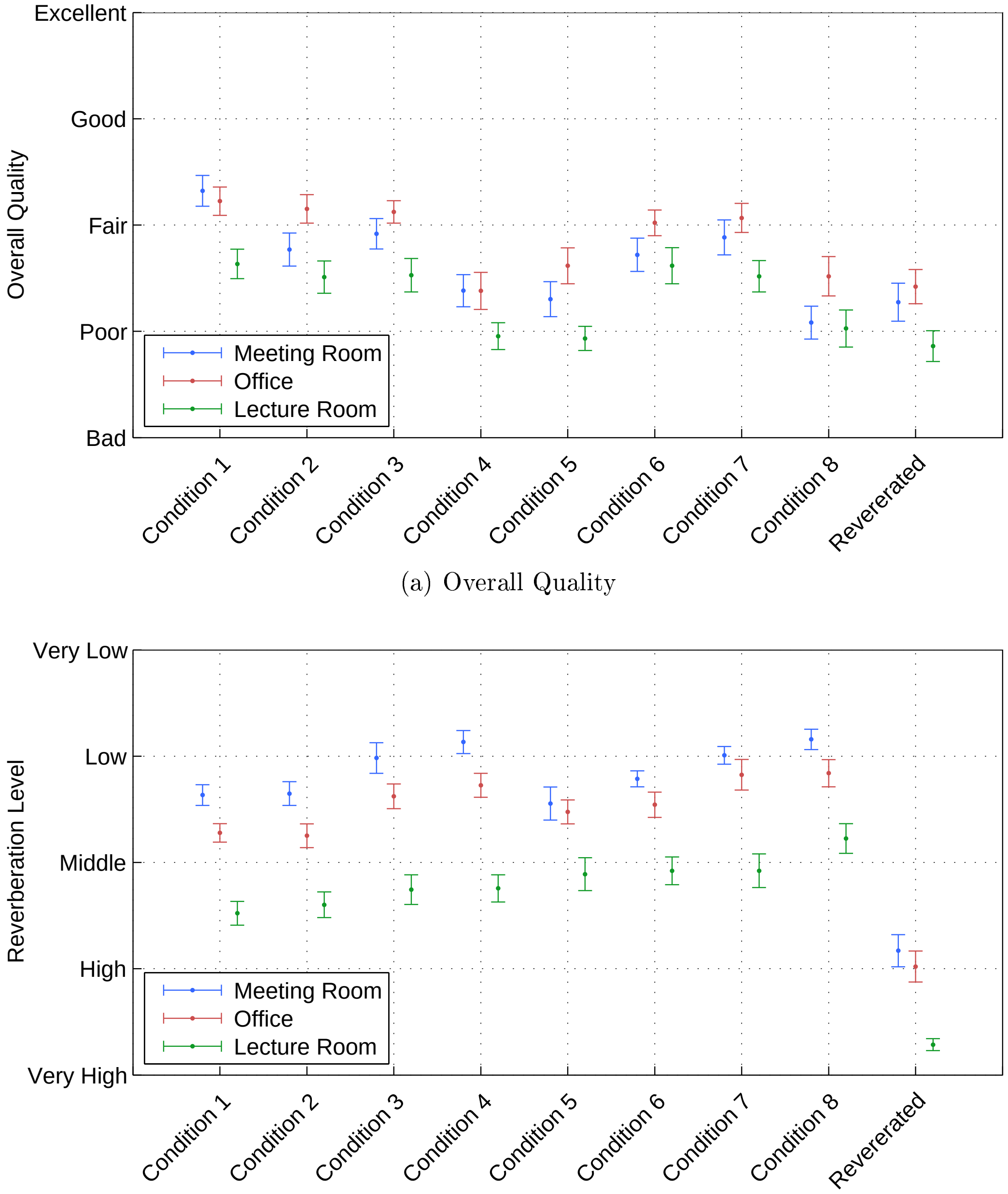


Figure 4.5 – Quality measure of gain function smoothing with M. Eng. @16kHz as sound file and different RIRs

While the FWSegSIR is approximately constant for every smoothing approach, a discrepancy in the SRMR and the CD can be detected. The cepstral distance decreases by using any of the smoothing methods, although they can better suppress *musical noise*. In fact also the higher SRMR without smoothing denotes a better dereverberation, but it can produce spectral artifacts which decrease the overall quality. By comparing the SRMR and the FWSegSIR of the different smoothing approaches, the octave smoothing seems to work best, but also the quality scores for the exponential smoothing with a smoothing constant $\alpha = 0.2$ are acceptable. The results of the multi-stimulus test are meant to reveal if the exponential smoothing is an appropriate and practical alternative to the more complex band-dependent approaches.

4.3.3 Results Multi Stimulus Test

The results of the multi stimulus test according to Section 4.2 are analyzed subsequently. For the test 23 listeners were asked for several sound files and experiments along Table 4.1. The differential scores for submitted results which exhibit statistical representativeness are shown in Figure 4.6.



(a) Overall Quality

(b) Reverberation Level

Figure 4.6 – Mean scores of the multi-stimulus speech dereverberation test and the evaluation of a confidence interval of 95% for different rooms and conditions as described in Table 4.1

The main awareness from the results is, that the reverberation level for all setups decreases strongly compared to the reverberated signal. This is a good verification of our goal outlined in Chapter 1.3. Further,an uncertainty relation between overall quality and the reverberation level can be seen. For each setup and room, the overall quality decreases by an excessive decrease of the reverberation level.

If the different parameter settings are compared, it can be seen that for example the lecture room, higher T_d values outperform in terms of reverberation level while the overall quality is not affected by changing T_d. Further, for the meeting room and the office, the higher $T_d = 64$ ms seems to work at least as good as $T_d = 32$ ms if the reverberation level and the overall quality is considered. By comparing the conditions with active or inactive smoothing, is can be seen that the smoothed gain functions provide a better overall quality but lower reverberation suppression as expected from the technical quality measures. By taking the gain functions into account, the approximated MMSE-LSA seems to yield to a lower reverberation suppression and higher overall quality when compared to the MMSE-LSA.

The experiment may not provide a general answer, because of its limited number of participants, rooms, and sound samples. Moreover, the vast number of thinkable parametric and algorithmic variations make it difficult to find definite answers indicating the perfectly compiled and adjusted algorithm. Still, the experiment managed to highlight some relevant tendencies among the different conditions and suitable default settings that were desired.

It always depends on the recording and especially on the room where the recording took place. Some parameters can potentially be optimized by blindly calculating i.e. the room dimensions or reverberation level. On the other hand, parameters like the minimum Gain G_{min} should be controlled by the user itself because of the ability to control the relation between overall quality and the strength of reverberation suppression.

Although the multi-stimulus test had fewer participating listeners and different rooms compared to the MUSHRA test in the REVERB-Challenge [35], it is interesting to offer a compromise between the test results. In Appendix B, Figure B.1 shows the results of the MUSHRA tests of the REVERB-Challenge. Especially the scores of the reverberation level differ as the proposed algorithm exhibits a lower reverberation level for all conditions and rooms except the large lecture room. The overall quality is similar for most of the conditions compared to the REVERB-Challenge results. However, for the conditions 1 and 7, the overall quality exhibits better results which could be induced due to the exponentially smoothed gain functions.

Chapter 5

Conclusion

Reverberation and noise degrade the speech intelligibility under particular circumstances and affect the naturalness of recorded sounds. The main goal of this work was to develop an approach for suppressing such unwanted influences by means of state-of-the-art dereverberation methods. Two different methods based on spectral subtraction and homomorphic deconvolution were investigated and extended. It has been shown exemplary that the reverberation cancellation under the successful use of the homomorphic deconvolution is still a big challenge and unlikely to yield accurate results due to phase unwrapping errors and the problem of cross-talking reflections within a frame-based dereverberation.

On the other hand, the spectral subtraction offered a promising technique for reverberation suppression. Based on the usage of Polack's reverberation model, the MS noise estimator, and the EMSR it was possible to exploit the perception of human hearing for the benefit of an appropriate dereverberation. Further, it can be seen from the results of the REVERB-Challenge that this is one of the most powerful techniques for single-channel dereverberation at the current state of research. Compared to the method proposed in [10], this thesis extended the reverberation model with a simple summation procedure and used a constant T_{60} assumption for the reverberation time estimation to achieve dereverberation in small rooms. Furthermore, different spectral gain functions were investigated and the parameter settings optimized for small room acoustic purposes.

Still, it is an illusion to think that this powerful method would achieve a perfect reconstruction of the clean signal with spectral subtraction because of the problem of exact detection of interference and speech. Especially for reflections, Polack's reverberation model does not fit anymore for all sorts of recordings or rooms respectively. Since the properties of an arbitrary room are unknown, it is impossible to find generally optimal settings that also cover the estimation of the early reflections. To achieve a complete reverberation cancellation, the magnitude and phase of the processed signal have to be considered at every time instance. In theory this is only approximately achievable with a very high amount of data and computational complexity. Thus, it is impracticable for current as well as for future applications.

Bibliography

[1] D. Bees, M. Blostein, and P. Kabal, “Reverberant speech enhancement using cepstral processing,” in *ICASSP, Toronto, Ontario*, 1991.

[2] D. Naik, “Pole-filtered cepstral mean subtraction,” in *ICASSP, Detroit*, 1995.

[3] S. Xizhong and M. Guang, “Complex cepstrum based single channel speech dereverberation,” in *ICCSE, Nanning*, 2009.

[4] T. Nakatani and M. Miyoshi, “Blind dereverberation of single channel speech signal based on harmonic structure,” in *ICASSP, Hong Kong*, 2003.

[5] T. Nakatani, K. Kinoshita, and M. Miyoshi, “Harmonicity-based blind dereverberation for single-channel speech signals,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 80 – 95, Jan. 2007.

[6] T. Nakatani, K. Kinoshita, M. Miyoshi, and P. S. Zolfaghari, “Harmonicity-based blind dereverberation with time warping,” in *ITRW on Statistical and Perceptual Audio Processing ICC, Jeju*, 2004.

[7] M. Tonelli and M. Davies, “A blind multichannel dereverberation algorithm based on the natural gradient,” IDCOM and Joint Research Institute for Signal and Image Processing, Edinburgh, Tech. Rep., 2010.

[8] K. Lebart, J. M. Boucher, and P. N. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acoustica*, vol. 87, no. 3, pp. 359 – 366, May. 2001.

[9] E. A. P. Habets, “Single channel speech dereverberation based on spectral subtraction,” 2004.

[10] B. Cauchi, I. Kodrasi, and R. Rehr, “Joint dereverberation and noise reduction using beamforming and a single-channel speech enhancement scheme,” in *REVERB Challenge Workshop*, 2014.

[11] J. D. Polack, “La transmission de l’ énergie sonore dans les salles,” Ph.D. dissertation, Ph.D. dissertation, Université du Maine, Le Mans, 1988.

[12] E. J. Nemer, “Blind dereverberation of speech using complex adaptive kurtosis maximization in the subband domain,” in *AES 60th Conference DREAMS ITN, Leuven*, 2016.

[13] H. Padaki, K. Nathwani, and R. M. Hegde, “Single channel speech dereverberation using the lp-residual cepstrum,” in *National Conference on Communications (NCC)*.

[14] M. Jeub, M. Schäfer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *Proceedings of the 16th International Conference on Digital Signal Processing*, 2009.

[15] R. Maia, M. Akamine, and M. J. F. Gales, “Complex cepstrum as phase information in statistical parametric speech synthesis,” in *ICASSP, Kyoto*, 2012.

[16] T. Apel, “Real-time complex cepstral signal processing for musical applications,” Center for Research in Computing and the Arts University of California, San Diego, Tech. Rep., Nov. 2001.

[17] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504 – 512, Jul. 2001.

[18] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466 – 475, Sept. 2003.

[19] C. Breithaupt, T. Gerkmann, and R. Martin, “A novel a priori snr estimation approach based on selective cepstro-temporal smoothing,” in *ICASSP, Las Vegas*, 2008.

[20] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, “An improved algorithm for blind reverberation time estimation,” in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[21] J. Eaton, N. D. Gaubitch, and P. A. Naylor, “Noise-robust reverberation time estimation using spectral decay distributions with reduced computational cost,” in *ICASSP, Vancouver*, 2013.

[22] Y. Ephraim and M. Rahim, “On second-order statistics and linear estimation of cepstral coefficients,” *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 162 – 176, Mar. 1999.

[23] P. Vary, U. Heute, and W. Hess, *Digitale Sprachsignalverarbeitung.* Teubner.

[24] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109 – 1121, Dec. 1984.

[25] ——, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443 – 445, Apr. 1985.

[26] F. Zotter, “Unterdrückung hörbarer Störgeräusche in Echtzeitsystemen,” Diplomarbeit, Institut für Elektronische Musik und Akustik, 2004.

[27] I. N. Bronstein, K. A. Semendjajew, and G. Musiol, *Taschenbuch der Mathematik.* Deutsch (Harri), 2008.

[28] C. Breithaupt, M. Krawczyk, and R. Martin, “Parameterized mmse spectral magnitude estimator for the enhancement of noisy speech,” in *ICASSP, Las Vegas*, 2008.

[29] P. J. Wolfe and S. J. Godsill, “Simple alternatives to the ephraim and malah suppresion rule for speech enhancement,” in *IEEE Workshop on Statistical Signal Processing*, 2001.

[30] I. S. Gradshteyn, I. M. Ryzhik, A. Jeffrey, D. Zwillinger, and I. Scripta Technica, *Table of integrals, series, and products.* Elsevier, 2007.

[31] E. Zwicker, “Subdivision of the audible frequency range into critical bands,” *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248 – 248, Feb. 1961.

[32] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation.* Springer Science and Business Media, 2009.

[33] B. Grundlehner, J. Lecocq, R. Balan, and J. Rosca, “Performance assessment method for speech enhancement systems,” in *in Proc. 1st Annu. IEEE BENELUX/DSP Valley Signal Process. Symp*, 2005.

[34] T. H. Falk, C. Zheng, and W.-Y. Chan, “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766 – 1774, Aug. 2010.

[35] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, “A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research,” *EURASIP Journal on Advances in Signal Processing*, vol. 1, no. 7, pp. 1 – 19, Jan. 2016.

[36] M. Slaney, “An efficient implementation of the patterson-holdsworth auditory filter bank,” Apple Computer, Tech. Rep., Nov. 1993.

[37] T. Daua, D. Püschelb, and A. Kohlrausch, “A quantitative model of the effective signal processing in the auditory system. i. model structure.” *Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3615 – 3622, Feb. 1996.

[38] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Prentice-Hall, Inc., 1993.

[39] N. Jillings, B. D. Man, D. Moffat, and J. D. Reiss, “Web audio evaluation tool: A browser-based listening test environment,” in *The 12th Sound and Music Computing Conference, Music Technology Research Group*, 2015.

[40] ITU-R, “Method for the subjective assessment of intermediate quality level of coding systems,” International Telecommunication Union, Tech. Rep., 2003.

Appendices

Appendix A

Results of the Technical Quality Metrics

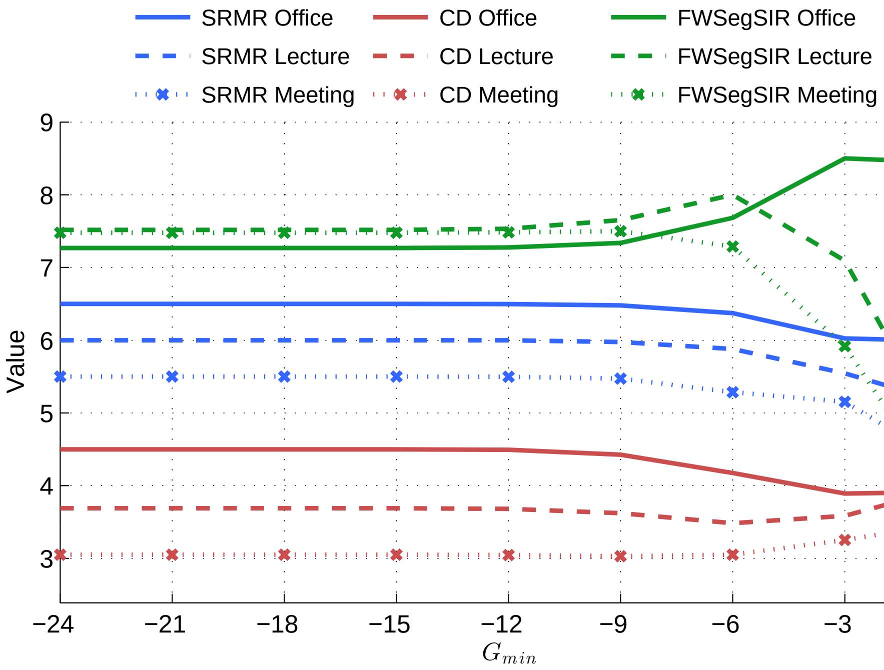


Figure A.1 – Quality Measure of G_{min} with W. Eng. @16kHz

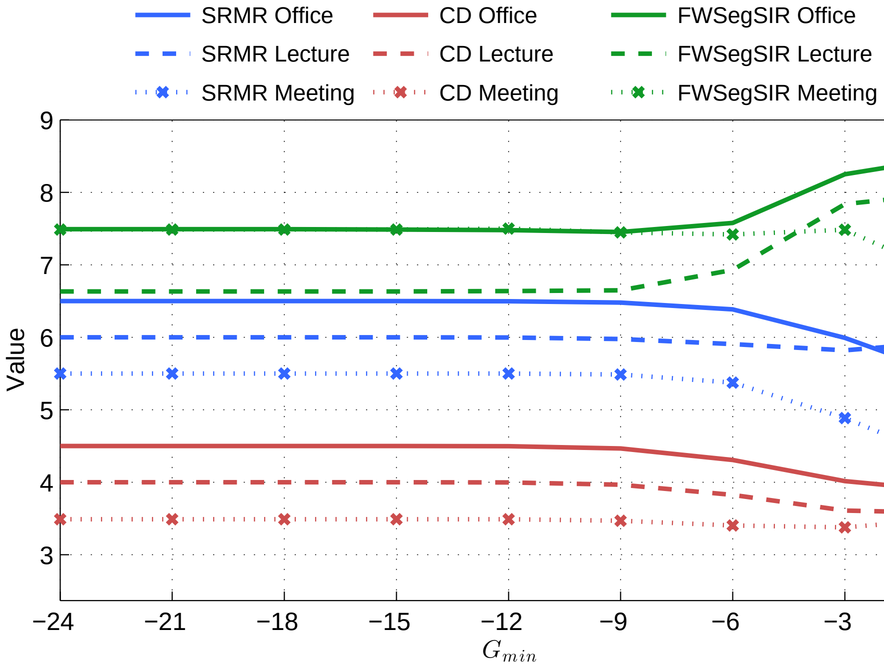


Figure A.2 – Quality Measure of G_{min} with W. Ger. @48kHz

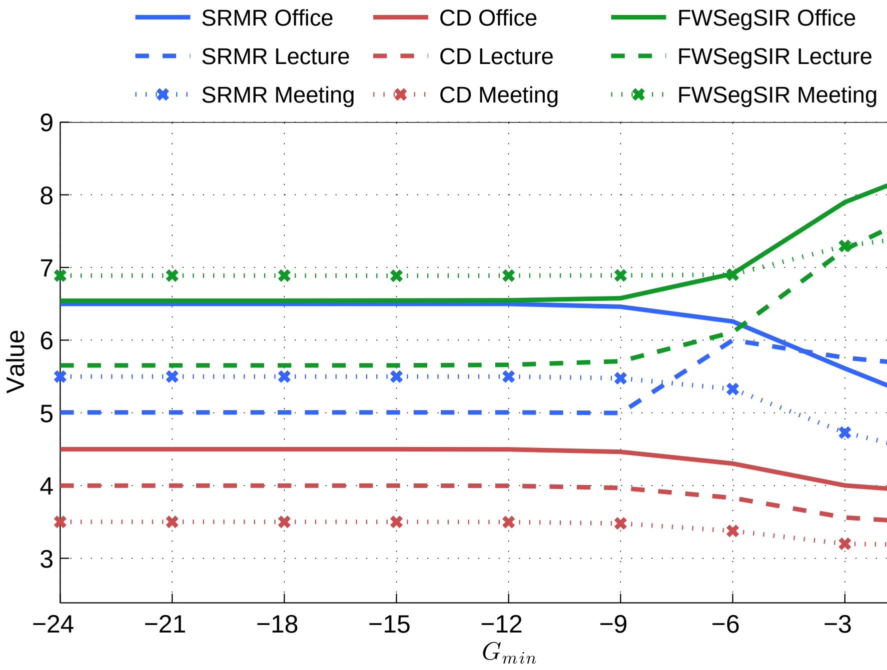


Figure A.3 – Quality Measure of G_{min} with M. Ger. @48kHz

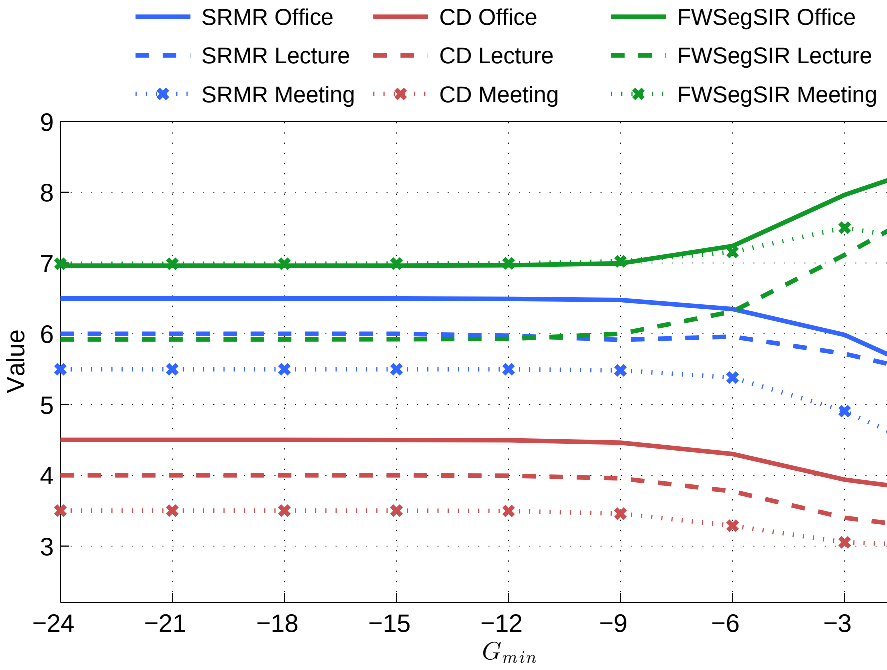


Figure A.4 – Quality Measure of G_{min} with Various @48kHz

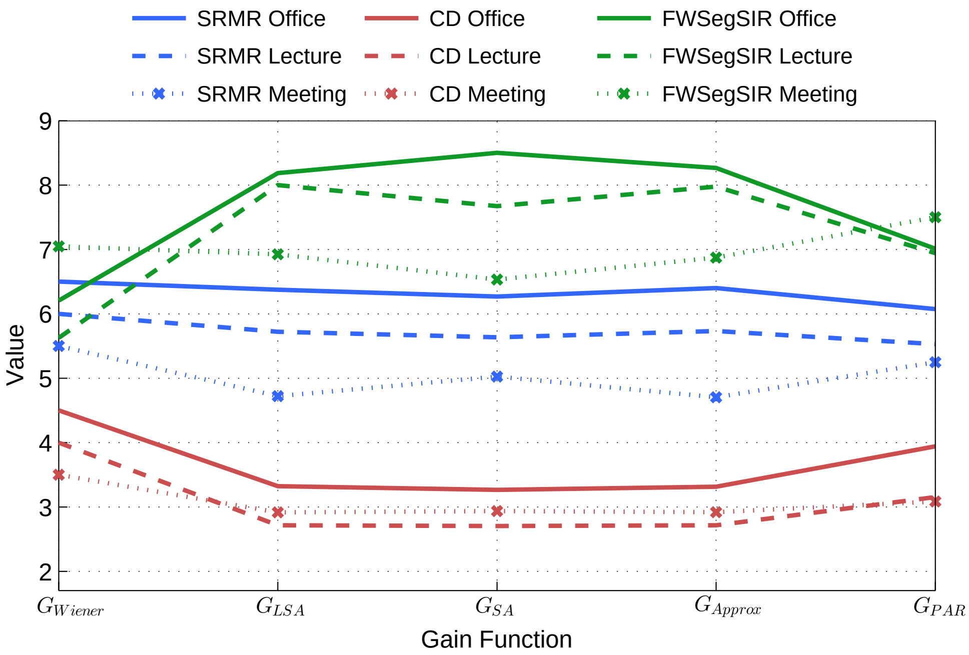


Figure A.5 – Quality Measure of gain functions with W. Eng. @16kHz

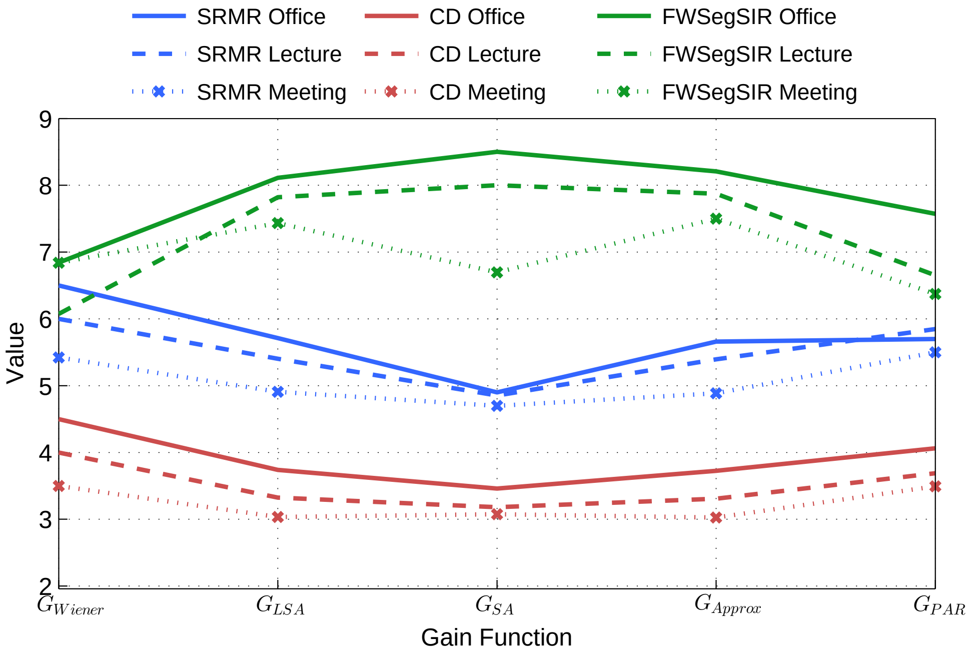


Figure A.6 – Quality Measure of gain functions with W. Ger. @48kHz

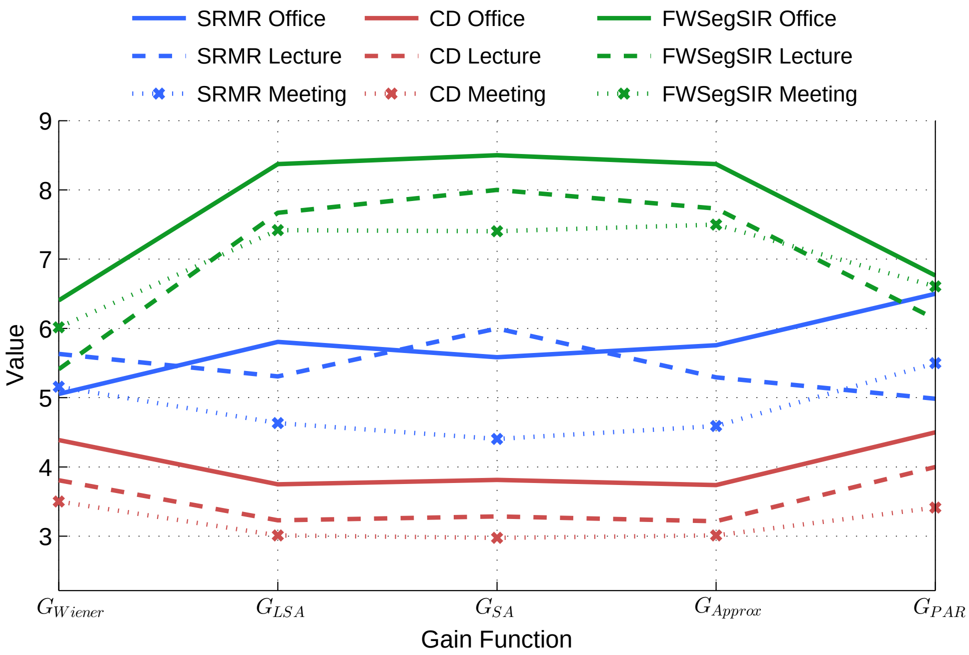


Figure A.7 – Quality Measure of gain functions with M. Ger. @48kHz

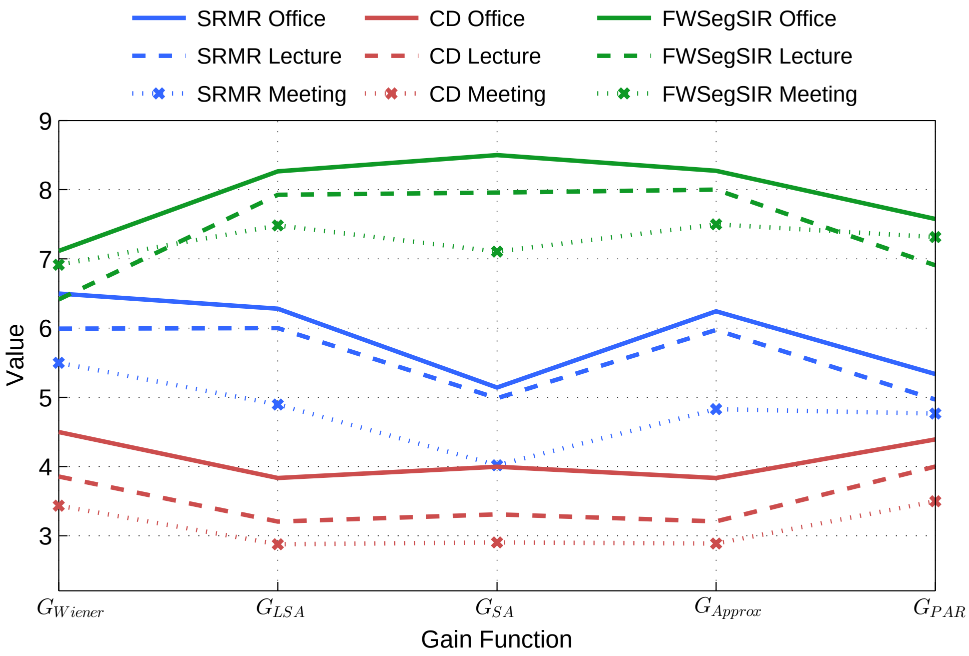


Figure A.8 – Quality Measure of gain functions with Various @48kHz

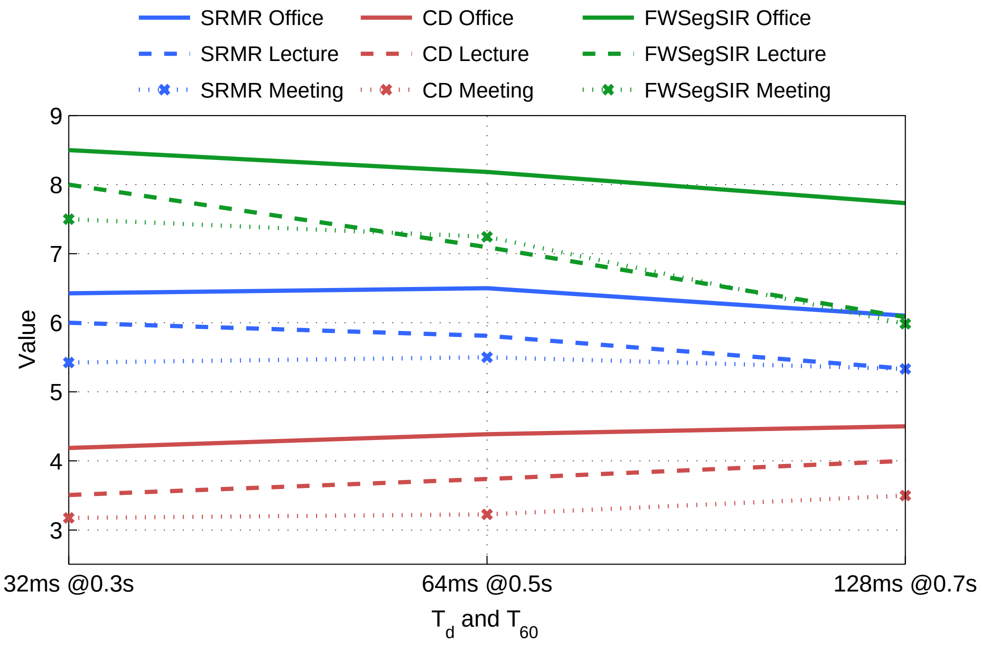


Figure A.9 – Quality Measure of T_d and T_{60} with W. Eng. @16kHz

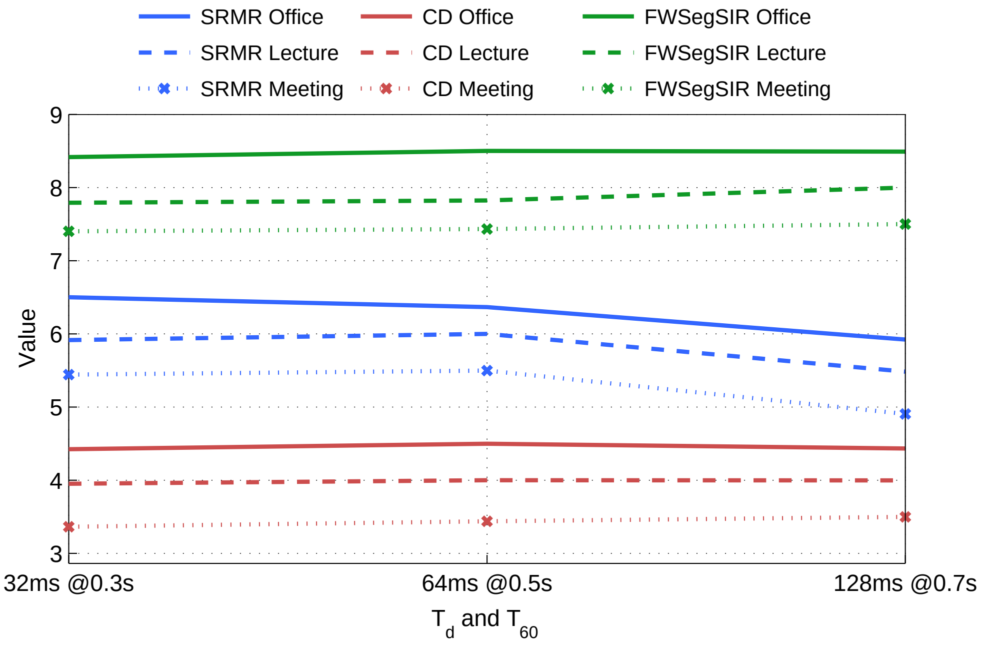


Figure A.10 – Quality Measure of T_d and T_{60} with W. Ger. @48kHz

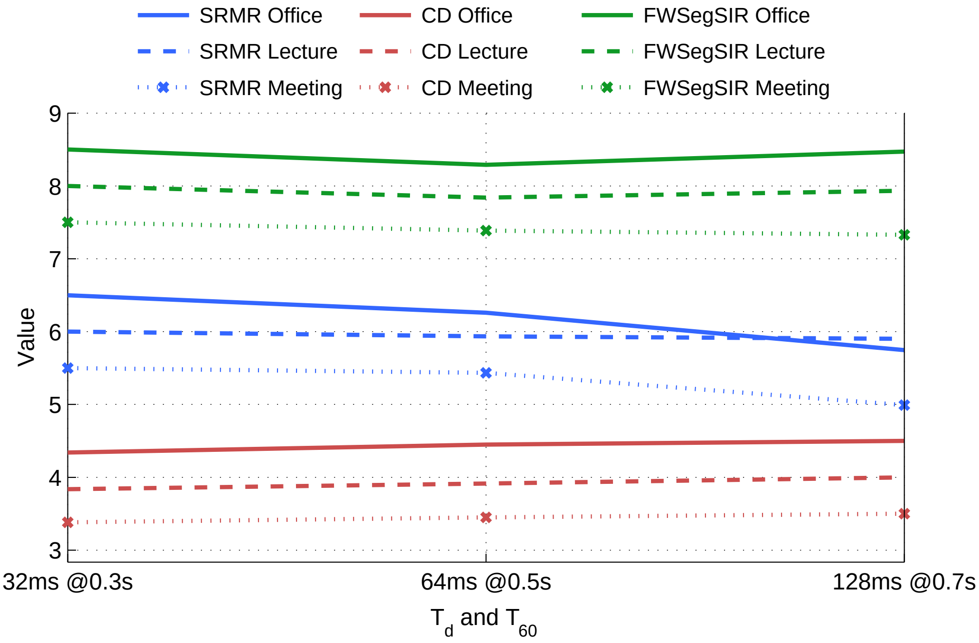


Figure A.11 – Quality Measure of T_d and T_{60} with M. Ger. @48kHz

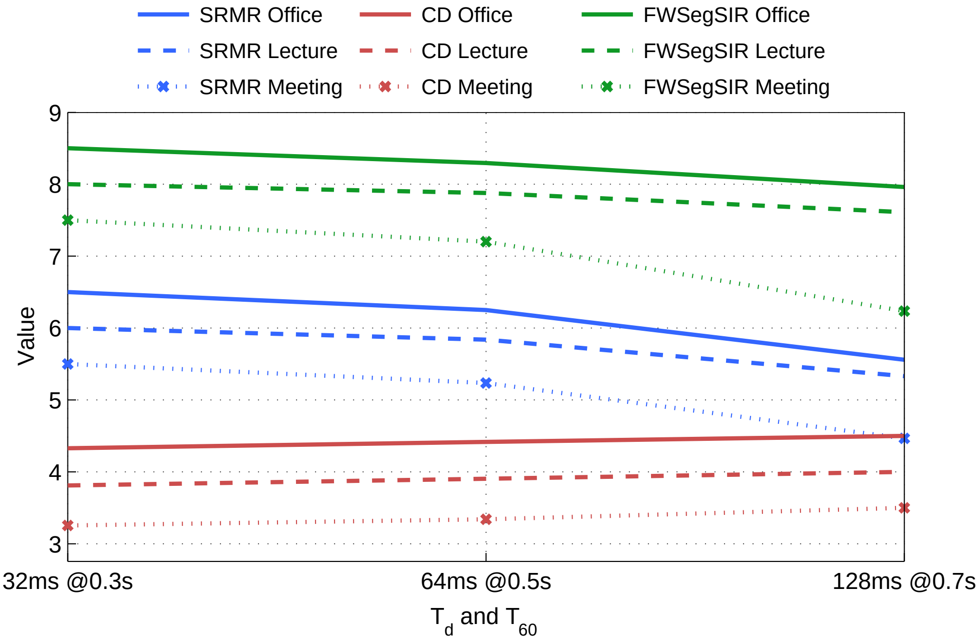


Figure A.12 – Quality Measure of T_d and T_{60} with Various @48kHz

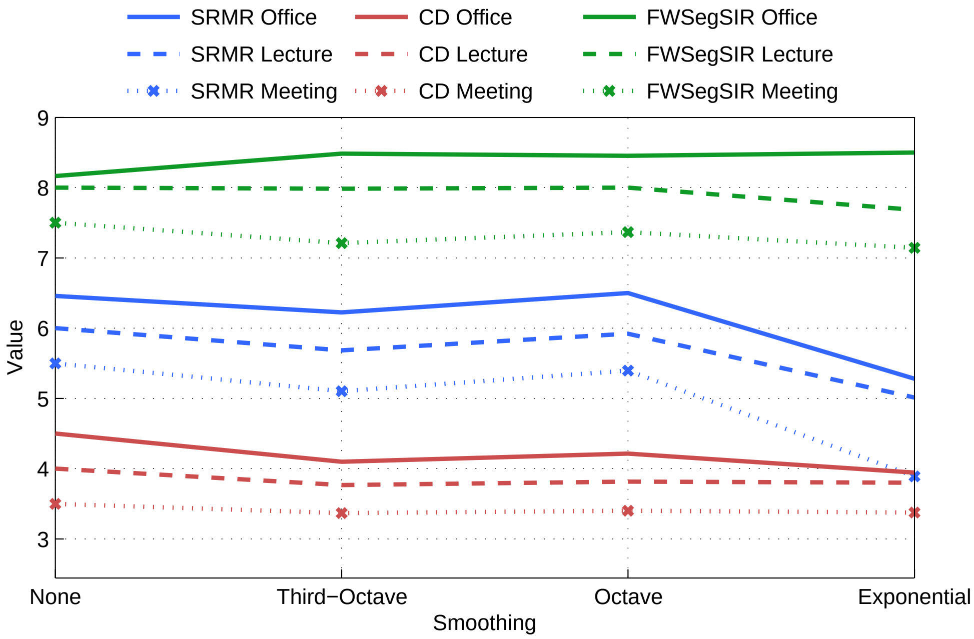


Figure A.13 – Quality Measure of gain function smoothing with W. Eng. @16kHz

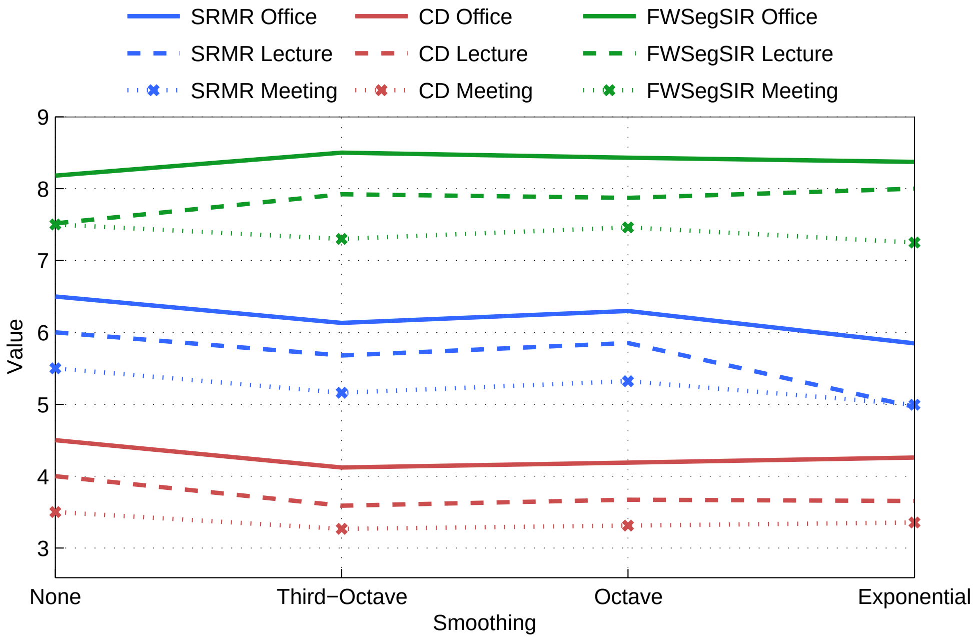


Figure A.14 – Quality Measure of gain function smoothing with W. Ger. @48kHz

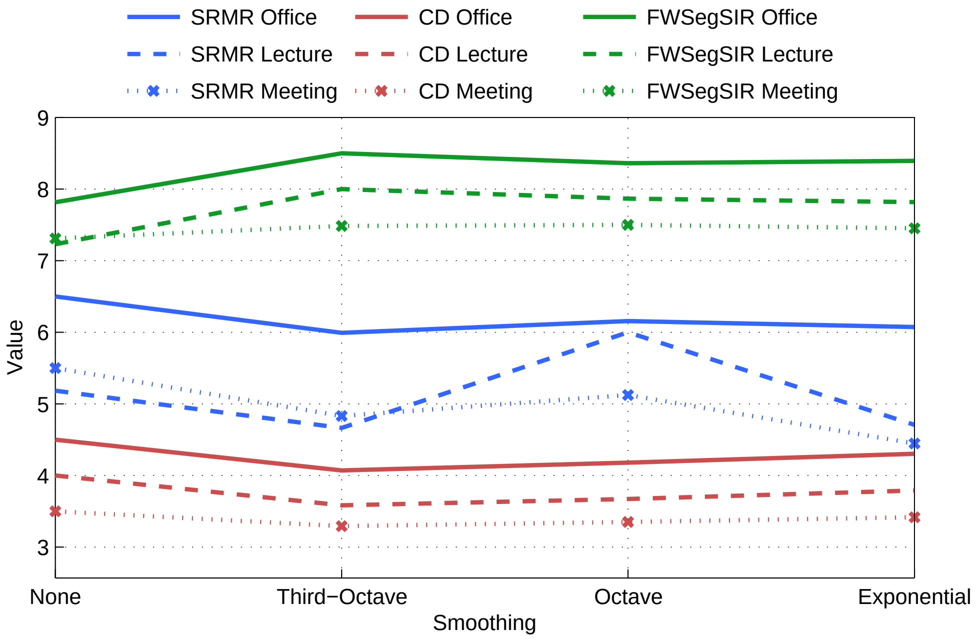


Figure A.15 – Quality Measure of gain function smoothing with M. Ger. @48kHz

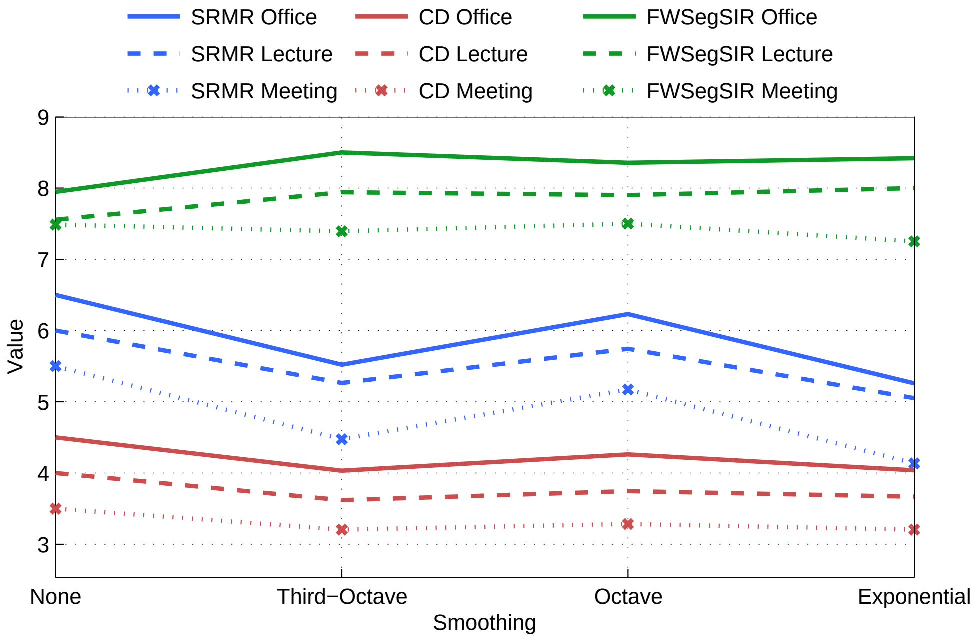
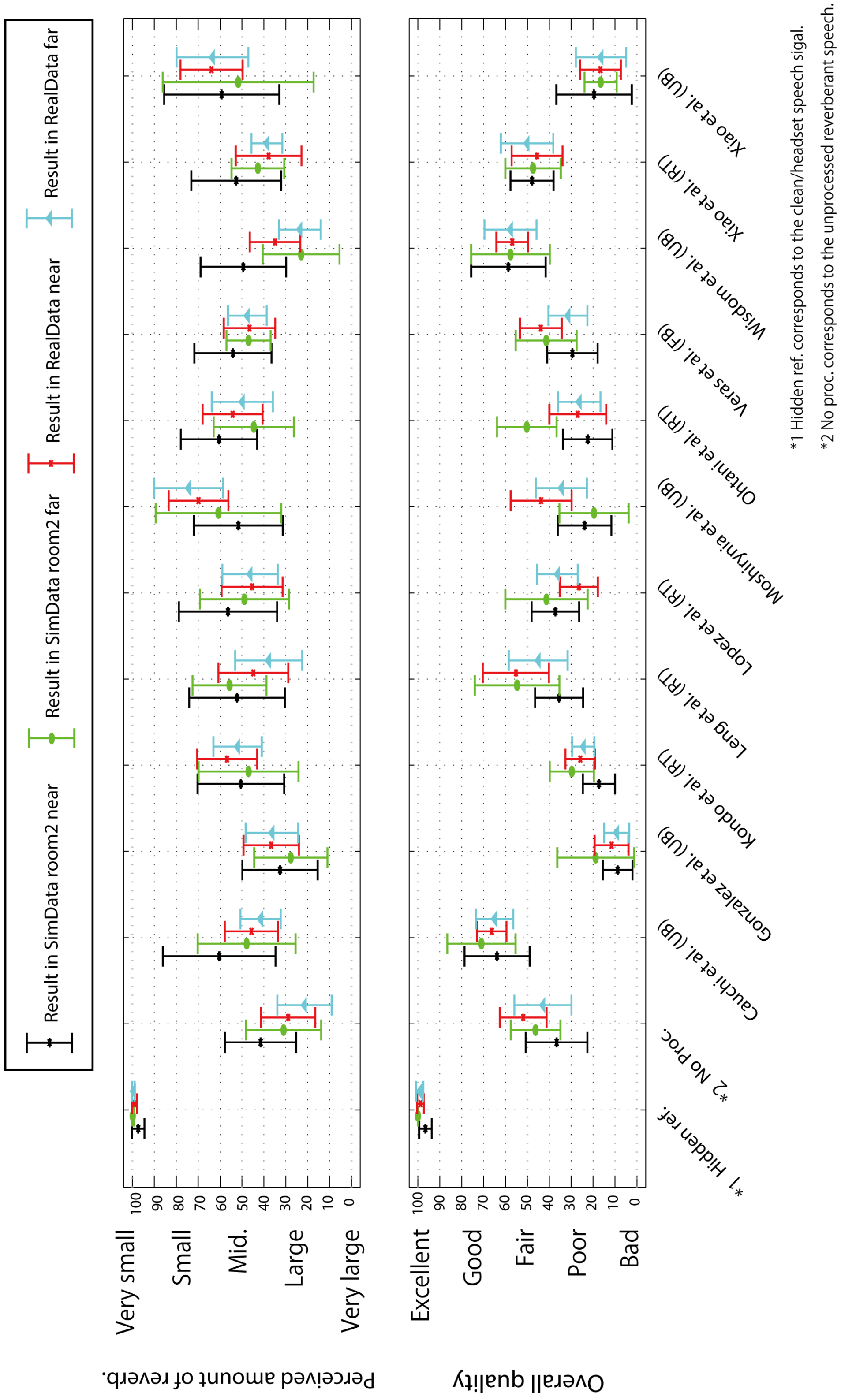


Figure A.16 – Quality Measure of gain function smoothing with Various @48kHz

Appendix B

MUSHRA Results of the REVERB-Challenge



*1 Hidden ref. corresponds to the clean/headset speech sigal.

*2 No proc. corresponds to the unprocessed reverberant speech.

Figure B.1 – Mean scores for the REVERB-Challenge [35] MUSHRA test plotted along with the associated 95% confidence interval