AUTOMATIC RESTORATION OF AUDIO SIGNALS IN MEDIA ARCHIVES

Von der Fakultät für Medizin und Gesundheitswissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)
angenommene Dissertation

von

Matthias Brandt
geboren am 30. Juni 1980
in Bremen, Deutschland

Matthias Brandt: *Automatic Restoration of Audio Signals in Media Archives*

BETREUER:
Prof. Dr. ir. Simon Doclo, Prof. Dr.-Ing. Jörg Bitzer

ERSTGUTACHTER:
Prof. Dr. ir. Simon Doclo

WEITERE GUTACHTER:
Prof. Dr.-Ing. Jörg Bitzer
Prof. Dr. Joshua D. Reiss

TAG DER DISPUTATION:
8. Mai 2018

ABSTRACT

A large number of historically relevant audio recordings are stored in media archives around the globe and represent an important part of mankind's cultural heritage. These recordings are stored on a variety of carriers, which usually underlie an aging process that results in the physical decay of the carrier material and ultimately the loss of the stored information. Therefore, large efforts have been made in recent years to digitize the inventory of media archives and prevent further deterioration. However, disturbances that have already been caused by the aging process remain in the digitized version of the recording. To reduce these disturbances, efficient audio restoration algorithms have been proposed, which typically require the manual adjustment of one or more algorithm parameters for each individual recording to achieve optimal restoration results. While manual operation is not a problem for a selected number of particularly valuable recordings, supervised restoration of complete archives is usually infeasible due to the sheer number of recordings.

The main topic of this thesis is *automatic restoration* of audio signals in media archives. More specifically, we propose algorithms that allow for an unsupervised restoration of a large number of audio recordings that show a great diversity with regard to the desired signal type, the disturbance type and the intensity of the disturbance. In doing so, we address three important disturbance types, i.e., impulsive disturbances, hum disturbances and broadband noise. Impulsive disturbances frequently occur with grooved recording media, e.g., wax cylinders, shellac and vinyl discs, and are caused by dirt and mechanical deformations of the media. Hum disturbances are often caused by power line interference with the audio signal during the recording process. Broadband noise is mainly caused by restrictions of the recording medium, e.g., the size of the magnetic particles for tape media. A key element in the design of the proposed algorithms is the desire to keep the degradation of the desired signal as low as possible while achieving a substantial improvement of the audio quality.

Firstly, based on the observation that state-of-the-art impulsive disturbance restoration algorithms often reduce the audio quality for undisturbed signals, which prohibits their unsupervised application, we propose a classification algorithm to classify frames of the input signal as either clean or disturbed. The algorithm is based on supervised learning and uses a logistic regression model with features that are computed from the appropriately prewhitened input signal. Evaluation results with a large number of test signals show that applying an existing impulsive disturbance restoration algorithm only on those frames that have been classified as disturbed leads to an improvement of the perceptual quality for a large range of signal-to-noise ratios (SNRs). In doing so, especially undisturbed signals are protected from detrimental processing with an impulsive disturbance restoration algorithm.

Secondly, we propose an algorithm to detect hum disturbances in audio recordings and to estimate all required hum disturbance parameters, i.e., the frequencies of the hum partial tones and their start and end times. The algorithm uses a quantile-based statistical analysis of the short-time power spectral density (PSD) estimates of the input signal in order to detect the presence of stable hum tones. The accuracy of the frequency estimation is increased by means of adaptive notch filters that converge towards the true frequencies of the hum partial tones. Evaluation results with real and artificial test signals show that most perceivable hum disturbances are detected with a low false alarm rate and with low estimation errors. Furthermore, we compare the performance of three state-of-the-art hum reduction algorithms, i.e., comb filters, subband comb filters and notch filters. Evaluation results show that the performance of the three considered algorithms differs significantly with regard to the amount of hum reduction and signal degradation. The results suggest that notch filters yield a high amount of hum reduction with a low amount of signal degradation if the frequencies of the hum partial tones are known.

Finally, based on the observation that state-of-the-art noise PSD estimation algorithms lead to significant estimation errors when applied to a large diversity of signals, especially at high SNRs and for music signals, we propose a noise PSD estimation algorithm which assumes that the noise in many archive recordings is stationary. The proposed algorithm estimates the noise PSD as the mean value of an exponential distribution that corresponds to the empirical distribution of the truncated short-time periodogram coefficients of the input signal. In addition, the algorithm provides a confidence measure that reflects the reliability of the noise PSD estimate, which can be used to decide whether restoration should be applied or not in a certain frequency band. Evaluation results with a large number of speech and music signals and a large range of SNRs show that the proposed estimation algorithm achieves significantly lower PSD estimation errors than a state-of-the-art algorithm based on minimum statistics. The evaluation results also show that the combination of the proposed noise PSD estimation algorithm with a state-of-the-art broadband noise reduction algorithm, rejecting noise PSD estimates with low confidence, leads to a quality improvement for a wide range of SNRs and only a small amount of signal degradation for practically noise-free signals.

The proposed algorithms constitute important steps for automatic audio restoration, over a wide range of SNRs and input signals, which are typically encountered in large media archives.

ZUSAMMENFASSUNG

Eine große Anzahl von historisch relevanten Tonaufnahmen wird in einer Vielzahl von Medienarchiven aufbewahrt und stellt einen wichtigen Teil des Kulturerbes der Menschheit dar. Diese Tonaufnahmen sind auf unterschiedlichen Trägerformaten gespeichert, die typischerweise einem Alterungsprozess unterliegen, der einen physikalischen Zerfall des Trägermaterials mit sich bringt. Auf lange Sicht ist eine Zerstörung der auf dem Träger gespeicherten Informationen unvermeidbar. Aus diesem Grund wird seit einiger Zeit großer Aufwand betrieben, um die Bestände der Archive zu digitalisieren und einem weiteren Verfall zuvorzukommen. Die bereits durch den Alterungsprozess verursachten Störungen verbleiben allerdings auch in der digitalen Version der Aufnahme. Um diese Störungen zu reduzieren, können effiziente Restaurationsalgorithmen verwendet werden, wobei typischerweise eine manuelle Einstellung von einem oder mehreren Parametern erforderlich ist um optimale Ergebnisse zu erzielen. Diese manuelle Bedienung stellt für eine ausgewählte Anzahl besonders wertvoller Aufnahmen kein Problem dar, macht allerdings die Restauration ganzer Archive aufgrund der großen Anzahl von Aufnahmen unmöglich.

Der Schwerpunkt dieser Dissertation ist die *automatische Restauration* von Tonaufnahmen in Medienarchiven. Genauer gesagt werden Algorithmen vorgestellt, die eine unbeaufsichtigte Restauration einer Vielzahl von Tonaufnahmen möglich machen, wobei sich die Tonaufnahmen durch eine große Vielfalt hinsichtlich des Nutzsignals, der Störungsart und der Intensität der Störung auszeichnen. Dabei werden drei wichtige Störungsarten behandelt: Impulsstörungen, Brummstörungen und breitbandiges Rauschen. Impulsstörungen treten häufig bei Trägern mit Tiefen- oder Seitenschrift auf, z.B. bei Wachszylindern, Schellack- und Vinyl-Schallplatten, und werden durch Schmutz und mechanische Verformung des Trägers verursacht. Brummstörungen werden häufig durch Einstreuungen aus dem Stromnetz in tonsignalführende Leitungen während des Aufnahmeprozesses verursacht. Breitbandiges Rauschen wird hauptsächlich durch Grenzen des Aufnahmenmediums verursacht, z.B. die Größe der Magnetpartikel bei Bandmedien. Eine wichtige Anforderung für die Entwicklung der vorgeschlagenen Algorithmen war, die Verfälschung des Nutzsignals so gering wie möglich zu halten und eine hörbare Verbesserung der Klangqualität zu erzielen.

Ausgehend von der Beobachtung, dass Algorithmen zur Impulsstörungsrestauration nach dem Stand der Technik häufig zu einer Verschlechtung der Klangqualität für ungestörte Aufnahmen führen, was eine unbeaufsichtigte Anwendung unmöglich macht, schlagen wir als erstes einen Klassifikationsalgorithmus vor, der Blöcke des Eingangssignals als entweder gestört oder störungsfrei klassifiziert. Der Algorithmus basiert auf überwachtem Lernen und nutzt ein logistisches Regressionsmodell mit Merkmalen, die aus dem auf geeignete Art geweißten Eingangssignal berechnet werden. Die Ergebnisse einer Evaluation mit einer großen Anzahl von Testsignalen zeigen, dass die Anwendung eines vorhandenen Algorithmus zur Impulsstörungsrestauration nur auf diejenigen Blöcke, die als gestört klassifiziert wurden, für einen großen Bereich von Signal-Rausch-Verhältnissen (SNRs) zu einer Erhöhung der wahrgenommenen Qualität führt. Dabei werden insbesondere ungestörte Signale vor einer nachteiligen Bearbeitung mit einem Impulsstörungsrestaurationsalgorithmus bewahrt.

Als zweites schlagen wir einen Algorithmus zur Detektion und Parameterschätzung von Brummstörungen in Tonaufnahmen vor. Die geschätzten Parameter sind die Frequenzen der Brumm-Partialtöne und deren Anfangs- und Endzeiten. Der Algorithmus verwendet eine quantilbasierte, statistische Analyse der geschätzten Kurzzeit-Leistungsdichtespektren (LDS) des Eingangssignals, um das Vorhandensein stabiler Brummtöne zu detektieren. Die Genauigkeit der Frequenzschätzung wird durch adaptive Kerbfilter erhöht, die gegen die wahren Frequenzen der Brumm-Partialtöne konvergieren. Die Ergebnisse einer Evaluation mit echten und künstlichen Testsignalen zeigen, dass der größte Teil der wahrnehmbaren Brummstörungen mit einer niedrigen Fehlalarmrate und mit kleinen Schätzfehlern detektiert wird. Des Weiteren wird die Effizienz von drei Brummreduktionsalgorithmen nach dem Stand der Technik verglichen, nämlich von Kammfiltern, Teilband-Kammfiltern und Kerbfiltern. Evaluationsergebnisse zeigen, dass die drei Algorithmen unterschiedliche Eigenschaften haben im Hinblick auf die Dämpfung der Brummstörung und den Grad der Verfälschung des Nutzsignals. Ausserdem deuten die Simulationsergebnisse darauf hin, dass eine starke Reduzierung der Brummstörung bei einem geringen Grad der Nutzsignalsverfälschung mit Kerbfiltern möglich ist, wenn die Frequenzen der Brumm-Partialtöne bekannt sind.

Schließlich, ausgehend von der Beobachtung, dass Rausch-LDS-Schätzalgorithmen nach dem Stand der Technik zu signifikanten Schätzfehlern führen, wenn diese auf sehr unterschiedliche Signale angewendet werden, insbesondere bei großem SNR und bei Musiksignalen, schlagen wir einen Rausch-LDS-Schätzalgorithmus vor, der von der Annahme ausgeht, dass das Rauschen in vielen Archivaufnahmen stationär ist. Der vorgeschlagene Algorithmus ermittelt das Rausch-LDS als den Erwartungswert derjenigen Exponentialverteilung, die zu den abgeschnittenen Kurzzeit-Periodogrammkoeffizienten des Eingangssignals passt. Des Weiteren ermittelt der Algorithmus ein Konfidenzmaß, dass die Zuverlässigkeit der Rausch-LDS-Schätzung anzeigt, und das verwendet werden kann um zu entscheiden, ob einzelne Frequenzbänder restauriert werden sollen oder nicht. Ergebnisse einer Evaluation mit einer großen Anzahl von Sprach- und Musikaufnahmen und über einen großen SNR-Bereich zeigen, dass der vorgeschlagene Algorithmus signifikant niedrigere LDS-Schätzfehler erzielt als ein Algorithmus nach dem Stand der Technik, der auf der Statistik des Minimums basiert. Die Evaluationsergebnisse zeigen außerdem, dass die Kombination des vorgeschlagenen Rausch-LDS-Schätzalgorithmus mit einem Reduktionsalgorithmus für breitbandiges Rauschen nach dem Stand der Technik und der Ablehung von LDS-Schätzwerten mit niedriger Konfidenz zu einer Qualitätsverbesserung über einen großen SNR-Bereich und nur zu einer marginalen Qualitätsreduktion für praktisch ungestörte Signale führt.

Die vorgeschlagenen Algorithmen stellen wichtige Schritte für eine automatische Audiorestauration über einen großen SNR-Bereich und unterschiedliche Arten von Eingangssignalen dar, wie sie üblicherweise in großen Medienarchiven vorzufinden sind.

CONTENTS

1 INTRODUCTION 1
- 1.1 Short History of Sound Recording 1
- 1.2 Archive Audio 4
- 1.3 Prior Work 10
- 1.4 Contributions and Thesis Outline 23

2 CLASSIFICATION OF IMPULSIVE DISTURBANCES 29
- 2.1 Introduction 29
- 2.2 Signal Model 32
- 2.3 Classification Algorithm 37
- 2.4 Evaluation Method & Results 41
- 2.5 Conclusions 52
- 2.A Appendix 52

3 DETECTION OF HUM DISTURBANCES 55
- 3.1 Introduction 55
- 3.2 Problem Statement 56
- 3.3 The Detection Algorithm 60
- 3.4 Evaluation 66
- 3.5 Conclusions 79

4 REDUCTION OF HUM DISTURBANCES 81
- 4.1 Introduction 81
- 4.2 The Signal Model 82
- 4.3 Hum Removal Filter Algorithms 83
- 4.4 Evaluation 83
- 4.5 Results 85
- 4.6 Conclusions 88

5 BROADBAND NOISE PSD ESTIMATION 91
- 5.1 Introduction 91
- 5.2 Signal Model 95
- 5.3 Noise PSD Estimation Algorithm 96
- 5.4 Evaluation & Results 101
- 5.5 Summary & Conclusions 111

6 SUMMARY, CONCLUSIONS AND FURTHER RESEARCH 113
- 6.1 Summary and Conclusions 113
- 6.2 Suggestions for Further Research 117

REFERENCES 121

ACKNOWLEDGMENTS 135

LIST OF FIGURES

Figure 1.1 Ethnographer Frances Densmore with Blackfoot chief Mountain Chief in 1916, during a phonograph recording session. . 4
Figure 1.2 Photomicrographs of microgroove vinyl discs. 8
Figure 1.3 Impulse restoration with a standard AR-model-based algorithm. 14
Figure 1.4 Erroneous reduction of transients by inappropriate processing of a music signal with a standard AR-model-based impulse restoration algorithm. 15
Figure 1.5 Loss of high-frequency energy caused by inappropriate processing of a music signal with a standard AR-model-based impulse restoration algorithm. 16
Figure 1.6 Frequency-domain denoising. 18
Figure 1.7 Noise PSD estimates obtained with state-of-the-art algorithms and the algorithm proposed in Chapter 5 for a music signal. 20
Figure 1.8 Automatic audio restoration workflow. 21
Figure 1.9 Schematic overview of the thesis. 26

Figure 2.1 Schematic overview of the classification algorithm. 33
Figure 2.2 Flow diagram of the impulsive disturbance classification system. 38
Figure 2.3 Mean power spectral density of music signals from several decades of the 20th century. 40
Figure 2.4 Classification accuracy for all analyzed types of prewhitening with varying block lengths. 47
Figure 2.5 Results of the instrumental audio quality evaluation of the three impulse restoration algorithms. 49
Figure 2.6 Results of the instrumental audio quality evaluation of the full restoration processing chain. 51

Figure 3.1 Time signal and spectrogram of a music recording containing a hum disturbance. 57
Figure 3.2 Power spectral density of hum disturbances. 59
Figure 3.3 Temporary decrease of the hum sinusoid power by destructive interference with elements of the clean signal. 61
Figure 3.4 Spectrogram and quantiles that are used for hum detection. 62
Figure 3.5 The tone detector stage. 65
Figure 3.6 Adaptive notch filter with bandpass constrained input signal. 66

Figure 3.7 Total error probability vs. threshold Θ. 71
Figure 3.8 The minimum error depending on the quantile combination. 72
Figure 3.9 Hit and false alarm rate for artificial test signals. 74
Figure 3.10 Frequency estimation error for artificial test signals. 75
Figure 3.11 Start time estimation errors for artificial test signals. 77
Figure 3.12 End time estimation errors for artificial test signals. 78

Figure 4.1 Transfer function and impulse response of the comb filter. . 84
Figure 4.2 Magnitude transfer functions of the complementary band splitting filters. 85
Figure 4.3 Transfer function and impulse response of the subband comb filter. 86
Figure 4.4 Transfer function and impulse response of the allpass notch filter. 87
Figure 4.5 Hum reduction properties of the allpass based notch filter in dependence on the notch bandwidth and the notch depth. . 89
Figure 4.6 Amount of desired signal distortion of the allpass based notch filter in dependence on the notch bandwidth and the notch depth. 89

Figure 5.1 Intermediate steps of the proposed noise PSD estimation algorithm, exemplarily shown with an artificially disturbed music signal as input. 97
Figure 5.2 Absolute difference (AD) and $\mathrm{AD}_{\mathrm{max}}$ for exemplary values of $\mathbf{P}_b$, μ and b. 101
Figure 5.3 Noise PSD estimation errors of the proposed algorithm for different SNRs for both distance measures. 106
Figure 5.4 Noise PSD estimation errors of the proposed algorithm using the normalized total AD distance measure and the minimum statistics algorithm for two search window lengths. 108
Figure 5.5 Confidence values of the proposed algorithm, averaged over all frequencies for each disturbed signal. 108
Figure 5.6 Instrumental audio quality evaluation of the processed signals obtained by combining the proposed noise PSD estimate with the MMSE STSA noise reduction algorithm for different MRC values. 109
Figure 5.7 Instrumental audio quality evaluation of the processed signals obtained by several algorithms. 111

LIST OF TABLES

Table 1.1 Main disturbance types in audio signals. 9

Table 2.1 Default parameters of the impulsive disturbance generation method from [125]. 36
Table 2.2 The ODG scale. 45
Table 2.3 Classification performance in dependence on the SNR. . . . 48
Table 2.4 Mean and standard deviation of the shape parameter of the gamma distribution for different SNRs. 53
Table 2.5 Features of the prewhitened signal that have been investigated. 54

Table 3.1 ODG values, perceived impairment, and median SNR values. 67
Table 3.2 Detection performance with real-world input data. 76

Table 4.1 Performance measures for the FIR comb filter. 88
Table 4.2 Performance measures for the subband FIR comb filter. . . 88

Table 5.1 The ODG scale. 104

1

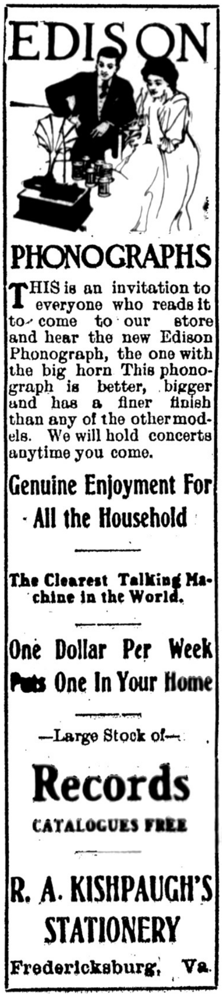
INTRODUCTION

1.1 Short History of Sound Recording

> "This is an invitation to everyone who reads it to come to our store and hear the new Edison Phonograph, the one with the big horn. This phonograph is better, bigger and has a finer finish than any of the other models. We will hold concerts anytime you come."
>
> —Advertisement for Edison Phonographs in The Free Lance, Vol. 22, No. 149, p. 3, January 11, 1908, Virginia, United States of America (a copy of the advertisement is on page 2).

It was in 1877 when Thomas Alva Edison turned the handle of this new device, setting the tinfoil-covered cylinder in motion. Reciting with a stentorian voice the nursery rhyme "Mary Had a Little Lamb", it was the first time a sound was captured for later reproduction. Edison himself was shocked after playing back his just made recording, listening to what seemed like the "ghost of speech" [1]. After thousands of years of mankind's fascination for the reproduction of the sound of human voice [2], the *phonograph* represented a breakthrough in this long quest. Edison's invention changed the way we think about sound and paved the way for new branches in culture, industry, technology, art and more.

The night after the first successful demonstration of his phonograph, Edison began to improve the phonograph in order to enhance the obtained sound quality. This was the beginning of a long line of research activity, aiming for the best possible reproduction. While Edison started by trying to fit the tinfoil more properly onto the cylinder of the phonograph, it would not take long until others began to introduce more substantial changes to the machine. In 1881, Charles Sumner Tainter and Chichester Bell replaced Edison's tinfoil with a wax-covered cylinder at the Volta Laboratory [3, 4]. This change led to an improved sound and less background noise. Furthermore, the wax cylinders were much more durable than the early tinfoil and their handling was easier. The new, improved machine was called *graphophone*, and altogether it represented a severe improvement compared to the phonograph. At first, phonographs and graphophones were mainly rented to businesses for dictation [5]. However, first music recordings were made in 1888 [6] and Edison realized that money could be made by selling prerecorded cylinders [4]. A major problem with wax cylinders at that time, however, was mass production: To create copies of the original recording, a so-called pantographic process was used that allowed to produce about 25 copies of each master. By recording the original performance on up to twenty cylinders, about 500 copies of one performance could be made. To create more copies, the performance had to be redone and recorded again, resulting in a rather tedious and time-consuming process [6]. Therefore, Edison experimented with different materials and spent years developing a process that allowed mass production of cylinder copies. Finally, he developed a method to create negative gold molds from the original recording. In a complicated process, positive copies could then be created from the molds by dipping them into hot wax. Due to the geometric shape of the cylinders, this process was error-prone and deformations of the wax material and mistakes in the process led to deformation of the grooves, resulting in distortions of the sound [4]. The achieved dynamic range was approximately 50 dB [7] with a frequency response of approximately 200 Hz to 4 kHz [8].



EDISON

PHONOGRAPHS

THIS is an invitation to everyone who reads it to come to our store and hear the new Edison Phonograph, the one with the big horn This phonograph is better, bigger and has a finer finish than any of the other models. We will hold concerts anytime you come.

Genuine Enjoyment For All the Household

The Clearest Talking Machine in the World.

One Dollar Per Week Puts One In Your Home

—Large Stock of—

Records

CATALOGUES FREE

R. A. KISHPAUGH'S STATIONERY

Fredericksburg, Va.

Shortly before Edison started his venture into selling prerecorded cylinders in 1889, Emil Berliner had patented his *gramophone* in 1887 [4]. In contrast to vertical modulation, Berliner used lateral modulation and flat discs instead of cylinders as a recording medium. After receiving enthusiastic responses from an audience that was impressed with the quality and clarity of his gramophone, Berliner began to believe that this new medium of sound recording had great applications in the area of entertainment, rather than business [2]. The combination of this insight with Berliner's disc production process was the beginning of the recording industry. A crucial fact for this to happen was the easy mass production of Berliner's discs: The original recording was cut into zinc plates coated with beeswax. After etching this disc in acid, electroplated molds were created that were negatives of the original recording. From these negatives, stampers were made and the final, positive records were pressed. Due to the disc shape of the medium, this process was less complicated and error-prone compared to Edison's cylinders and a large number of copies could be produced easily [6]. While the first disc records consisted of hard rubber, a new material, shellac, was used starting around 1900 [9], leading to highly increased mechanical stability of the carrier. The frequency response was comparable to the frequency response of wax cylinders [7] with a slightly lower dynamic range, reaching approximately 40 dB [10].

With the availability of technically mature recording and playback devices and mass-produced prerecorded media, the research activity in the field of audio technology increased steadily. Important milestones on the path from shellac discs to today's digital hard disc recording were numerous. For example, the *telegraphone*, invented by Valdemar Poulsen in 1889, was the first device to record an audio signal *electrically*, onto a steel wire [5]. Several years later, after important inventions like the condenser microphone and vacuum tube amplifier had been made in the 1920s, the *Magnetophon* was a sensation at the Berlin exhibition in 1935 [11]. The frequency range now was 50 Hz to 5000 Hz, with a dynamic range that was similar to the 40 dB obtained with shellac discs [12]. A huge improvement in recording quality was presented shortly after, in 1941, namely the application of AC bias. This new technology increased the dynamic range to 60 dB and the recordable frequency range to between 50 Hz and 10 000 Hz [11], which can be seen as the first "noise-free" recording that exhibits practically no perceivable noise while covering large parts of the frequency range of the human auditory system. As a consequence, tape recording started to become the standard audio recording medium, e.g., for recording studios—with the famous singer Bing Crosby's enthusiasm for the high sound quality stimulating its success [5]. In the following, magnetic tape recording was improved and developed further in many ways—with Philips' introduction of the *Compact Cassette* in 1963 (reaching a dynamic range of 50 dB and a frequency response of 30 Hz to 18 kHz [13]) and the increase of the dynamic range by up to 10 dB by Ray Dolby's invention of his noise-reduction system in 1967 [14, 15] fostering tape as the number one audio recording medium. Further important developments were, e.g., stereophonic sound systems, invented in 1931 by Alan D. Blumlein, and in 1948 the long-playing record (LP) by Peter Goldmark which achieved a dynamic range of approximately 60 dB and a frequency range of 30 Hz to 15 kHz [9, 16]. For several decades, until around 1990, LPs and tape were the main carriers for audio, when Compact Disc sales overtook [17]. The Compact Disc (CD), introduced by Philips and Sony in 1982 [18] was the first digital consumer audio carrier and was the climax of research activity going back to the 1930s, when pulse code modulation was invented. The dynamic range now was larger than 90 dB with a frequency range of 2 Hz to 20 kHz [19]. In the same year, a new technology was introduced to the market that might be seen as a factor for the decline of the CD: *hard disc recording*, invented by New England Digital [20, 21]. With the increase of available memory and increasing network transfer rates, analog storage and distribution via physical carriers has become less and less important in recent years. Important related developments of digital signal processing techniques for audio processing, e.g., the invention of the efficient coding and compression algorithm *MPEG-1 Audio Layer III (mp3)*, standardized in 1992 [22], were crucial in this process. Nowadays, a large number of high-quality audio recordings with practically unlimited dynamic range and frequency range can be stored even on mobile phones, and sent around the globe within mere seconds.

1.2 Archive Audio

> "What are we to expect from this wonderful invention? Mainly, we fear, an immense storing up of sounds that it might be better not to store up, an immense accumulation of those winged words whose wings are best employed in carrying off into nothingness what deserves only temporary life."
>
> —Excerpt from "What Will Come of the Phonograph?", The Spectator, No. 3.131, p. 881, June 30, 1888, London, England.

1.2.1 *The Motivation for Media Archives*

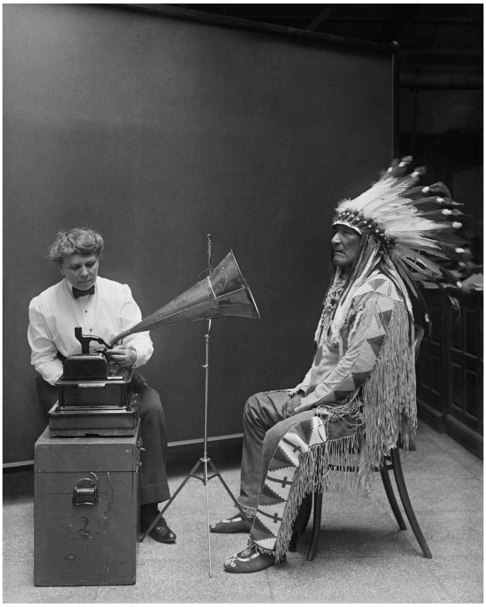


Figure 1.1: Ethnographer Frances Densmore with Blackfoot chief Mountain Chief in 1916, during a phonograph recording session.[1]

With the wide availability of sound recording and storage technology, the number of existing audio documents is increasing steadily. While the production and sale of music and spoken word recordings has become a profitable market since the availability of the first prerecorded cylinders and discs, the storage and archival of audio documents receives more and more attention. The screening of available recordings and the management and storage of valuable ones is crucially important. In addition to providing access to information, media archives serve as a safe storage of the world's cultural heritage. The number of audio documents stored in archives around the globe can only be estimated, but it is immense: e.g., the Library of Congress of the United States reports around 3.5 million audio documents in 2014 [23], the National Archives and Records Administration of the United States reports more than 400 000 sound and video recordings [24] and the German *Bundesarchiv* reports approximately 44 000 audio media of various kinds [25]. These media exhibit a large diversity and comprise recordings of extinct languages, early music but also wildlife recordings and bird songs [26]. Examples for historically relevant documents are recordings of Tasmanian Aboriginal songs and spoken word from 1899 [27], early

American Indian music (see also Figure 1.1) [28] and piano recordings of famous composers, e.g., Edvard Grieg from 1903 [29].

Analog audio documents demonstrate the challenges of information preservation much more radical than print media. Due to the low redundancy regarding the information storage and the high data density, the structural decay of analog carriers (e.g., [30]) has severe consequences for the stored information [31]. While safeguarding of the world's documentary heritage has received great attention in recent years, it has become more and more apparent that special consideration has to be taken regarding sound recordings on mechanical carriers such as cylinders and discs, magnetic carriers such as tapes, hard discs and floppy discs, and optical carriers such as Compact Discs [32]. This is caused on the one hand by the obsolescence and future unavailability of machinery that is required to access the information stored on various media, an on the other hand by the quick decay process of some carriers. Is has become evident that the only possibility for long-term storage of the information stored on analog carriers is digitization. In the digital domain, using error detection and correction techniques, it is possible to copy the information to a new carrier without a decrease of quality once a certain number of errors is reached [33]. This so-called *re-recording* process [34] makes it possible to create a practically unlimited number of copies from an original with identical information. Moreover, the availability of the recording in the digital domain allows for global access [35].

While the re-recording from the original carriers and subsequent digitization interrupts the disintegration process of the carrier material in many cases, the digital copy of the signal contains different types of disturbances caused by the aging process. The degree of perceptual quality degradation varies depending on the type of the original carrier, age and storage conditions. As a consequence, restoration may be unnecessary for a certain recording, while in extreme cases a recording may already be completely useless. For many recordings, the audio quality is far below the quality modern listeners are accustomed to. Furthermore, large amounts of noise impede the study of the vast amount of oral recordings,e.g., for research purposes, as listeners tend to lose concentration when listening to noisy signals for a longer period of time [36, 37].

The perceptual and technical characteristics of the disturbances vary. Each carrier typically exhibits different types of disturbance, depending on how the information is stored on the material. Generally, all analog carriers are prone to multiple disturbance types. For example, shellac discs usually contain broadband noise and impulsive disturbances. While tape media typically do not contain impulsive disturbances, they may exhibit other types of disturbances, such as signal drop outs caused by magnetic stray fields. Altogether, a multitude of disturbances exist, some having more dramatic influence on the signal than others and all differing in perceived sound and origin.

---

[1] Photo source: Library of Congress, Reproduction Number: LC-DIG-npcc-20061, Washington, D. C., United States of America.

1.2.2 *Disturbance Types*

As a basic classification, disturbance types can be separated into *acoustic* and *technical* disturbances [38]. Acoustic disturbances arise when undesired acoustic sources produce sound that leaks into the recording microphone, e.g., an air conditioning unit humming in the background of a speech recording, applause or cheering in a classical orchestra recording or the hissing of the ocean when recording the singing of birds. In contrast, technical disturbances either arise due to shortcomings of the carrier material, e.g., bubbles in shellac composite material, aging processes or shortcomings of the used recording or playback equipment. In the context of this thesis, only technical disturbances will be considered. Technical disturbances can be defined in a systematic way, justified by their origin, e.g., the technical properties of the carrier or the recording equipment. In contrast, the definition of acoustic disturbances is rather subjective and a matter of personal taste. To decide whether, e.g., the hissing of the ocean in the background should be removed from some live coverage requires a large amount of semantic information that is usually not available in the context of archive audio restoration. Of course, the intensity of technical disturbances becomes worse if, in addition to aging, the carriers are stored improperly and dust and other particles, e.g., fungus, is attracted by the information-carrying surface. The main factors causing a physical degradation are the same for all carriers [35]:

- humidity,
- temperature,
- mechanical deformation, and
- dust/dirt.

In addition, for magnetic tapes severe damage can occur to the signal by strong magnetic stray fields [35] that directly manipulate the information stored on the tape.

Depending on the medium a recording was originally made on, the technical disturbance types vary. Technical disturbances are often classified into two groups (cf. Table 1.1), depending on whether a recording is affected as a whole (global disturbances) or if only portions are disturbed (local disturbances) [39]. Major global disturbances comprise, e.g., broadband noise (often called *hiss*), additive tonal disturbances (*hum*), pitch variation defects (depending on the modulation frequency called *wow* or *flutter*), linear and non-linear distortions or stereo issues. Broadband noise is caused by (thermal) noise generated in the amplifiers that are used for a recording and by the restrictions of the recording medium. These restrictions are, e.g., the surface roughness of wax, shellac and vinyl for cylinder and disc media and the size of the magnetic particles for tape media. Additive tonal disturbances are typically caused by power line interference during the recording process, e.g., when audio signal lines are placed close to power cables, or by faulty electric circuits. The alternating electromagnetic field caused by mains power lines leads to an induction of an alternating current in audio signal lines. The frequency of this alternating current typically corresponds to the mains frequency, i.e., 50 Hz or 60 Hz, depending on where the recording is made. In addition, non-linearities in the signal chain and interference from phase-fired controllers, that are often used in light dimmers, in many cases cause harmonics of the mains frequency [40] which may reach up to frequencies of several kHz [41]. As a result, the overall hum disturbance may be described by a harmonic tone complex, consisting of one or more partial tones: the fundamental frequency and possibly harmonics with frequencies of integer multiples of the fundamental frequency. Pitch variation defects are caused by inconstant angular velocity of tape machine capstans or by non-centered spindle holes with disc media. Linear distortions may be caused, e.g., by the frequency response of a mechanical or analog transmission path involved in a copy process. Non-linear distortions are generally caused by saturation of electronics or magnetic recording media [39]. Stereo issues are caused, e.g., by improperly aligned playback heads of tape machines [42]. Examples of local disturbances are impulsive types of noise that may occur with all grooved recording media and optical film sound tracks, where dust particles or scratches on the recording surface affect only short parts of the signal. Figure 1.2 shows photomicrographs of a clean and a dirty microgroove vinyl record. It can be seen that scratches, dust and dirt particles have sizes in the order of the groove width and, hence, can be expected to interfere with the stored audio information. Due to its perceived sound, this type of disturbance is commonly called *click*, *crackle* or *thump*, depending on the specific character of the disturbance. In the context of audio restoration, the term click is usually used to describe localized degradations that appear rather sporadically, e.g., caused by scratches on the surface of a disc. In contrast, crackle can be described as a static floor of small clicks, e.g., caused by a layer of dust or dirt that is more or less evenly distributed on the surface. Thumps are low-frequency pulses in the signal that are caused by the response of the pick up system to severe clicks that are, e.g., caused by large scratches on grooved media or optical film sound tracks [39].

Due to the very different properties of the used recording equipment, media, age and storage conditions, the intensity of the disturbances with respect to the desired signal typically varies largely in media archives. In some cases the desired signal may be barely perceivable while in other cases a recording may be completely free of disturbances. Some recordings may show multiple disturbances, e.g., broadband noise, clicks and crackle for shellac discs, while others may only contain broadband noise, e.g., tape recordings. In general, the most prominent disturbances will vary from archive to archive. Nevertheless, in many cases impulsive disturbances, hum disturbances and broadband noise are encountered most frequently [43].

1.2.3 *Restoration of Archive Audio*

With the availability of a recording as a digital signal, digital signal processing techniques can be used for different purposes, e.g., to extract metadata from the audio signal [25, 35] or to enhance the quality by reducing the contained disturbances [38, 39]. Metadata extraction comprises for example the classification of speech and

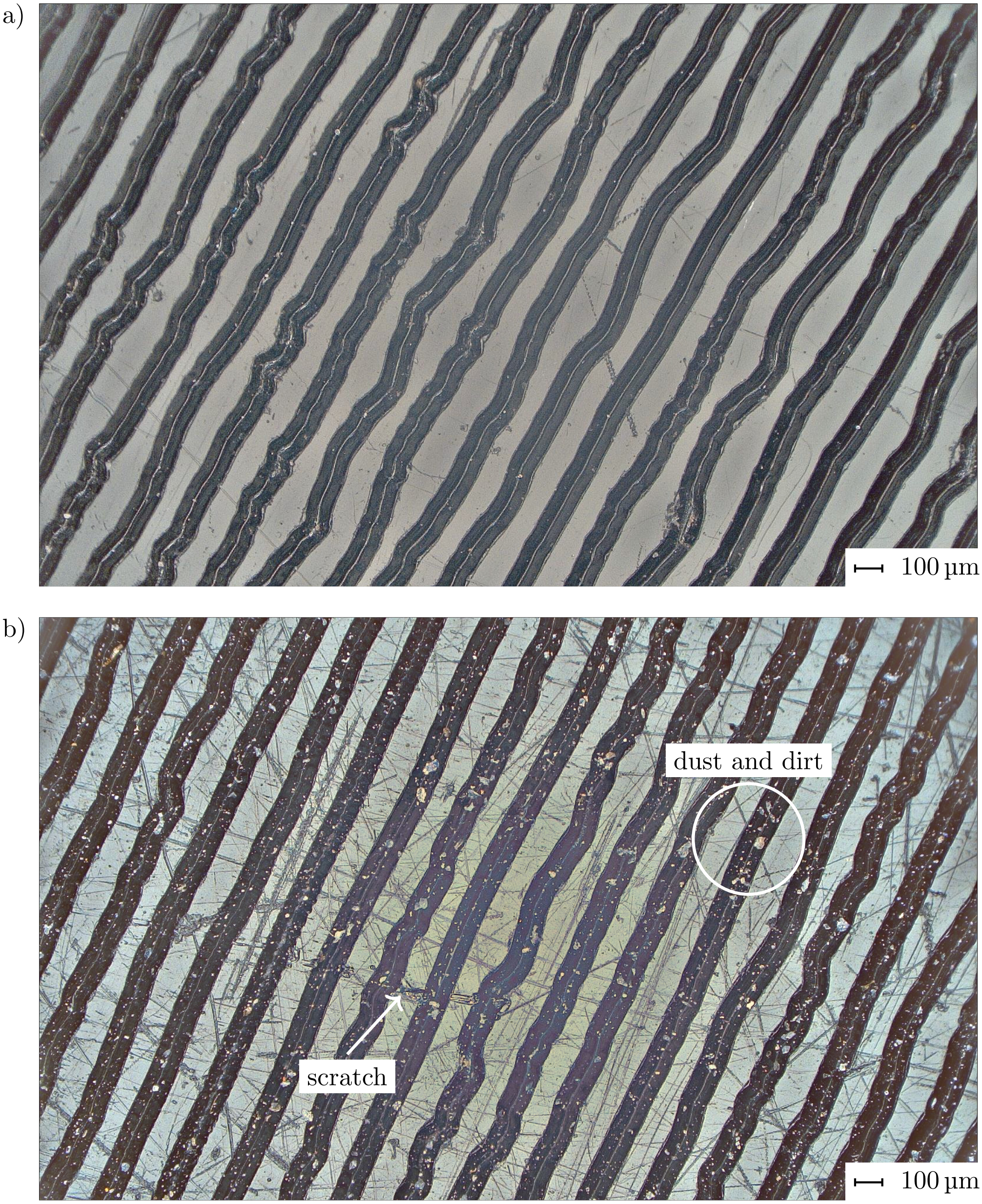


Figure 1.2: Photomicrographs of microgroove vinyl discs. a) A used but cleaned vinyl disc. b) A vinyl disc with scratches and large amounts of dust and dirt.

Table 1.1: Main disturbance types in audio signals.

Global disturbances	Local disturbances
Broadband noise	Clicks
Hum	Crackle
Pitch variation defects (wow and flutter)	Thumps
Linear distortions	
Non-linear distortions (saturation, clipping)	
Stereo issues (level imbalance, channel delay)	

music signals, the estimation of the bandwidth of the audio signal, or obtaining indications on the original carrier, e.g., through determining the presence or absence of impulsive disturbances. The additional information can then be used, e.g., to facilitate the retrieval of individual recordings and the access to media archives [44, 45].

The possibility to enhance the audio quality by using digital signal processing techniques has brought up a number of questions regarding the goals of the re-recording process. In an attempt to establish a standard re-recording procedure, it is argued in [46] that legitimate goals are the preservation of an audio document as it was heard by people of the time of the original recording, and striving to obtain the true sound of the original performer at the same time. In order to go one step further, advances in technology can be used to compensate for imperfections caused by technical limitations of the recording technology that was used to make the original recording [47]. Nevertheless, in the context of this thesis the definition from [32, p. 55] is adopted, in which the term *restoration* is defined as "the process of restoring an object to a condition as close as possible to that when it was first made." In other words, the overall goal of this thesis is to develop methods that enable to reduce the amount of signal degradation caused by carrier-related processes such as disintegration, age and wear.

As mentioned above, due to the large variety with regard to the type of desired signal, recording technology, carrier type, age and storage conditions, the recordings in archives show a large diversity of disturbances and signal-to-noise ratios SNRs. As a consequence, each recording requires a different restoration process. For example, only the subset of the archive recordings that had been stored on grooved media or optical film is prone to impulsive disturbances. To avoid detrimental effects of the application of a restoration algorithm, it is important to only process those signals that actually contain a specific type of disturbance. In this regard, the decision of whether a signal is disturbed or not is typically not straightforward. For example, a technically measurable disturbance, e.g., quantization noise, may be imperceptible by human listeners. In the context of this thesis, we assume a recording to be disturbance-free if it is perceived as such.

Unfortunately, for many recordings the original media that could give indications on potentially contained disturbances are unknown. As a consequence, information about disturbances can often only be based on an analysis of the signal itself. The individual, manual inspection and supervised restoration may surely be performed for a small number of recordings. However, a manual processing of typically very large numbers of recordings that are stored in an archive is not possible. If the restoration of complete archives is desired, the only option is automatic processing. The main challenge of archive audio restoration, therefore, is to perform unsupervised restoration for a large number of very diverse recordings (each one requiring different processing) and to achieve high restoration quality while minimizing the risk of signal degradation.

1.3 Prior Work

While a lot of research has been performed to improve the recording process and the sound storage media themselves, the last decades have shown a lot of research on reducing disturbances in audio signals *after* they have been recorded. An early publication from 1983 contains an enthusiastic description and first results regarding the restoration of music signals by means of digital signal processing [48]. Since then, a large number of algorithms for the restoration of audio signals have been proposed [38, 39]. These algorithms are usually specialized for each disturbance type, hence typically differing in the assumptions that are made regarding the signal model and, as a consequence, their functionality. The main goal of audio restoration is to improve the (perceptual) quality of a recording, i.e., to reduce the disturbance without significantly affecting the integrity of the desired signal. In doing so, the audio restoration process typically represents a compromise between the amount of disturbance reduction and the amount of signal degradation. While in many cases it is possible to achieve a very high restoration quality using existing algorithms, in order to obtain optimum results it is usually necessary to adjust one or more algorithm parameters for each recording individually. For example, a crucial aspect for broadband noise restoration is typically the manual selection of noise-only sections for the restoration algorithm to accurately estimate the noise power spectral density (PSD) (the so-called *fingerprint*). For hum restoration, typically the fundamental frequency of the hum tone complex and the number of harmonics have to be selected. For unsupervised operation, this required user interaction imposes significant restrictions on the applicability of most audio restoration algorithms and is related to the usually narrow ranges of parameter settings that lead to high-quality results. Severe degradations of a signal may occur if it is processed with inappropriate parameter settings. In general, the requirements for algorithms aiming at automatic restoration are substantially different from the requirements for algorithms aiming at manual restoration. The crucial point in this regard is the above-mentioned required robustness against a large diversity of desired signals, disturbance characteristics and disturbance intensities that occur in media archives. Furthermore, many existing audio restoration algorithms are designed for causal and real-time processing, e.g., to ease the manual parameter adjustment, or for live broadcasting applications. Algorithms for automatic restoration are not restricted to causal processing as the complete signals are usually available. Finally, the latency and computational complexity of the algorithms are of minor importance as the processing is typically not time-critical.

Existing audio restoration algorithms typically comprise two stages: the estimation of disturbance parameters and the actual reduction of the disturbance. When the disturbance parameters are known (or estimated accurately), it is usually possible to obtain high-quality restoration results for the majority of disturbed signals. To automate the restoration process, it is therefore crucial to develop a way to robustly estimate the disturbance parameters. This also includes determining whether a disturbance is absent in order to protect undisturbed recordings (or sections of a recording) from possible detrimental processing with restoration algorithms. Only a few publications on the specific problem of *automatic* audio restoration exist, and to the best of the author's knowledge do not fulfill all crucial points that are important for automatic restoration of diverse archive audio. The major restriction of existing automatic restoration algorithms is the assumption of a certain class of input signals, e.g., noisy speech recordings. In [49] a promising automatic sound restoration system for archive purposes is described that is able to automatically remove wow, flutter and broadband noise. However, no description of the broadband noise restoration algorithm or evaluation results are provided.

This thesis specifically addresses the problem of insufficient robustness of existing audio restoration algorithms against a large diversity of input signals and SNRs if no manual parameter adjustment is performed. In doing so, we focus on the *detection of individual disturbance types*, i.e., whether a disturbance type is present in (a section of) a recording or not, and the *estimation of disturbance parameters*. In order to maximize the applicability only single-channel methods are considered, since typically only a subset of the archive recordings is available in multichannel formats, .

The following sections give an overview on existing algorithms to reduce different disturbance types in audio signals and discuss potential issues regarding their unsupervised operation for automatic archive restoration applications. More specifically, we discuss the restoration of *impulsive disturbances* (Section 1.3.1), *hum disturbances* (Section 1.3.2) and *broadband noise* (Section 1.3.3). The algorithms addressing the individual disturbance types differ fundamentally as the generation process of the disturbances is different, i.e., they are described by different signal models. In addition, we briefly describe the typical audio restoration workflow (Section 1.3.4) and discuss several options to evaluate the quality of audio restoration algorithms (Section 1.3.5).

1.3.1 *Restoration of Impulsive Disturbances*

Impulsive disturbances represent one of the most prominent disturbance types, especially for grooved recording media. While early publications only deal with the detection of impulsive disturbances in audio recordings [50], approaches to remove clicks and crackle followed soon after, e.g., [51]. Typically, the signal portions affected by impulsive disturbances are determined first and corrected in a subsequent step. Early detection approaches consist of high-frequency pre-emphasis to enhance transient elements in the input signal followed by thresholding of the preprocessed signal [50]. Early correction methods simply consist of replacing the disturbed signal portions with silence or linearly interpolating the neighboring samples [51]. The linear-prediction-based algorithm for the detection and interpolation of impulsive disturbances in speech signals proposed in [52] represented a big improvement regarding the achieved restoration quality and is the basis for many state-of-the-art algorithms. These algorithms, which are based on an autoregressive (AR) model representation for the clean signal, make use of the fact that the clean signal at a certain time can be approximated adequately based on the surrounding signal. In contrast, impulsive disturbances do not follow this AR model and can therefore be detected by a large AR model prediction error. Corrupted samples of the detected signal portions are then replaced using the AR model and the uncorrupted signal surrounding the detected portion [38].

More recent versions of the AR-model-based algorithms aim at improving the impulse detection accuracy on the one hand, and the replacement of disturbed samples on the other hand. For example, in [53, 54] it is proposed to process the input signal in forward as well as backward direction to reduce the number of erroneously detected impulses (so-called *false alarms*). Other algorithms use impulse templates to increase the detection accuracy [55] or are based on machine learning techniques [56, 57, 58]. Interpolation algorithms based on true linear prediction [59] or frequency-warped prediction are reported to achieve high-quality results for gap lengths as long as approximately 45 ms [60, 61].

Many existing impulse restoration algorithms have been shown to yield high-quality results for a variety of disturbed input signals, e.g., speech signals and music signals from different genres that contain different types of impulsive disturbances. Figure 1.3 shows a section of a music signal with impulsive disturbances that has been copied from a shellac disc. This figure also shows the output signal of a standard AR-model-based impulse restoration algorithm [39, Ch. 5.2.3.2] with iterative estimation of the clean signal AR parameters [39, Ch. 5.3.1] using the implementation from [62]. It can be observed that impulses are removed effectively from the disturbed input signal. However, these algorithms typically require the adjustment of a threshold parameter, which is related to the disturbance intensity, and assume that the input signal contains impulsive disturbances. As mentioned before, it can usually not be assumed that all audio recordings in a media archive actually contain impulsive disturbances. As a consequence, the unsupervised application of an impulse restoration algorithm may lead to insufficient disturbance reduction or to a degradation of the signal, primarily due to false alarms in the detection stage. These false alarms are typically caused by impulse-like elements of the desired signal, e.g., drum transients, guitar pickings, distorted guitar or attacks of brass instruments and synthesizers. Figure 1.4 shows the undesired reduction of transients from the desired signal if inappropriate impulse restoration is performed. In this example, the input signal does not contain impulsive disturbances, but certain elements of the desired signal (the transients of a synthetic drum sound) are erroneously detected as such and removed. A second possible issue with unsupervised application of an impulse restoration algorithm is shown in Figure 1.5, where high-frequency content of a distorted guitar recording is attenuated because part of the signal is erroneously recognized as an impulsive disturbance, leading to a significant change in sound character. In Chapter 2, we present evaluation results with a large number of test signals that underpin these observations and suggest that the application of an impulse restoration algorithm with a fixed threshold parameter is beneficial for signals that contain severe disturbances, but may lead to a signal degradation if the SNR is high and especially for undisturbed signals.

1.3.2 *Restoration of Hum Disturbances*

In many cases, hum disturbances can be removed effectively by using comb filters or by placing narrow-band notch filters at the hum tone frequencies. In order to do so, at least the fundamental frequency of the hum tone complex is required. It should be noted that care has to be taken to avoid the generation of processing artifacts that are, e.g., related to the fact that comb filtering is achieved by adding a delayed version of the signal to itself, causing an echo effect, or that are related to ringing effects with narrow-band notch filters. Chapter 4 contains a detailed comparison of existing hum reduction filter algorithms. More specifically, we compare the performance of comb filters, subband comb filters and notch filters in terms of hum reduction and signal degradation by means of an evaluation with artificially disturbed test signals. We show that comb filters generally lead to a large amount of hum reduction (as all hum partials are removed), but also a comparatively large amount of signal degradation. Nevertheless, comb filters may be useful when only the fundamental frequency of the hum tone complex is known and no information about the number of hum harmonics is available. It is also shown that notch filters allow for a large amount of hum reduction with a lower signal degradation compared to comb filters.

While hum reduction with known disturbance parameters is possible with a small amount of signal degradation, it is important to refrain from applying a hum reduction algorithm if a recording does not actually contain hum. In archive audio restoration, the information whether a recording contains a hum disturbance or not is typically not available. Therefore, the presence of hum needs to be detected and the hum parameters need to be estimated from the signal itself. To the best of the author's knowledge, only two publications exist that propose algorithms to automatically estimate hum tones in audio recordings [66, 67]. In [66] an algorithm is proposed that is shown to efficiently remove hum in an automatic manner. However,

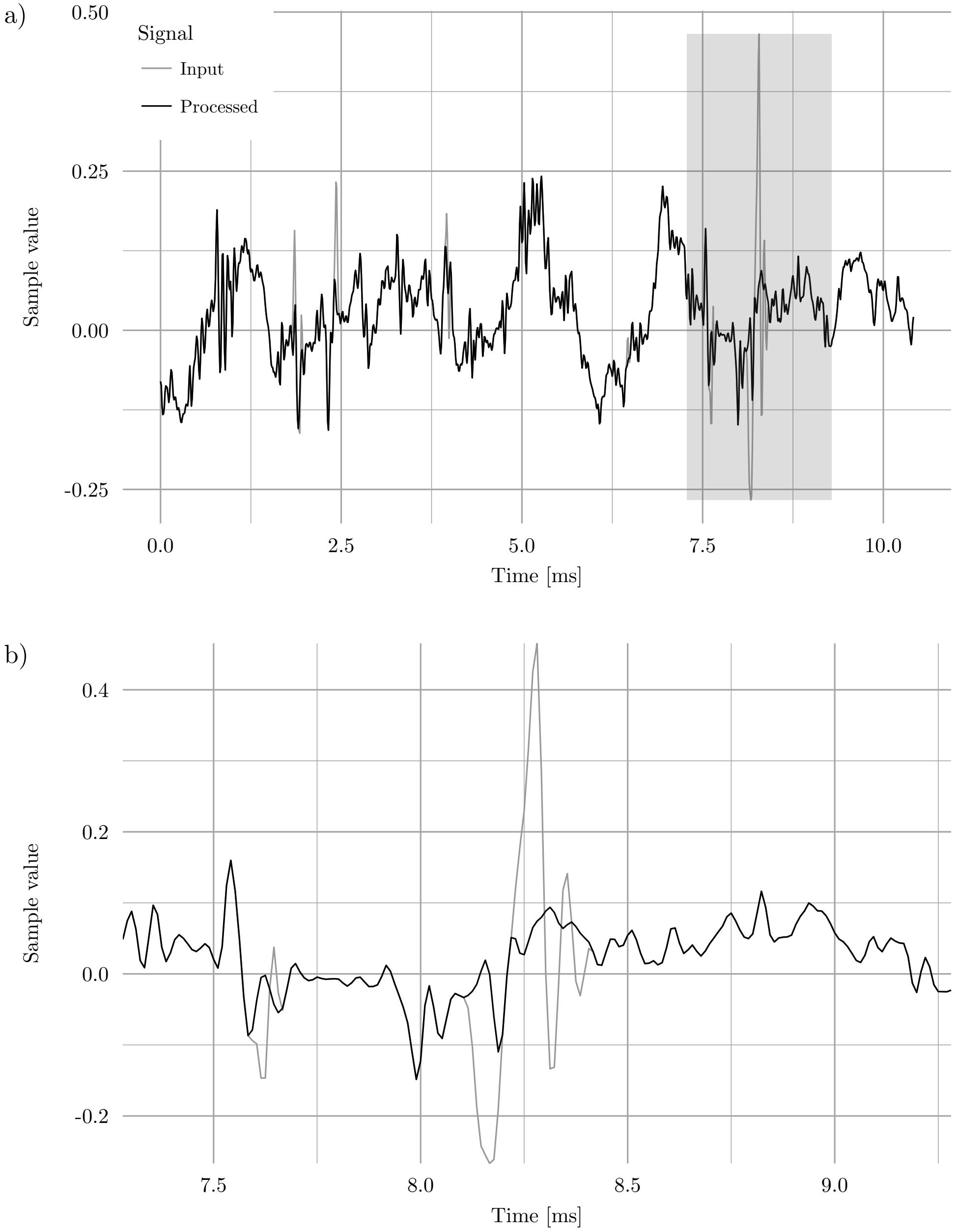


Figure 1.3: Impulse restoration with a standard AR-model-based algorithm [39, Chs. 5.2.3.2 and 5.3.1]. a) Short section of a gramophone recording on a shellac disc [63]. A number of impulsive disturbances have been detected and interpolated by the algorithm. b) Zoom of the region that is marked in plot a) with a grey rectangle.

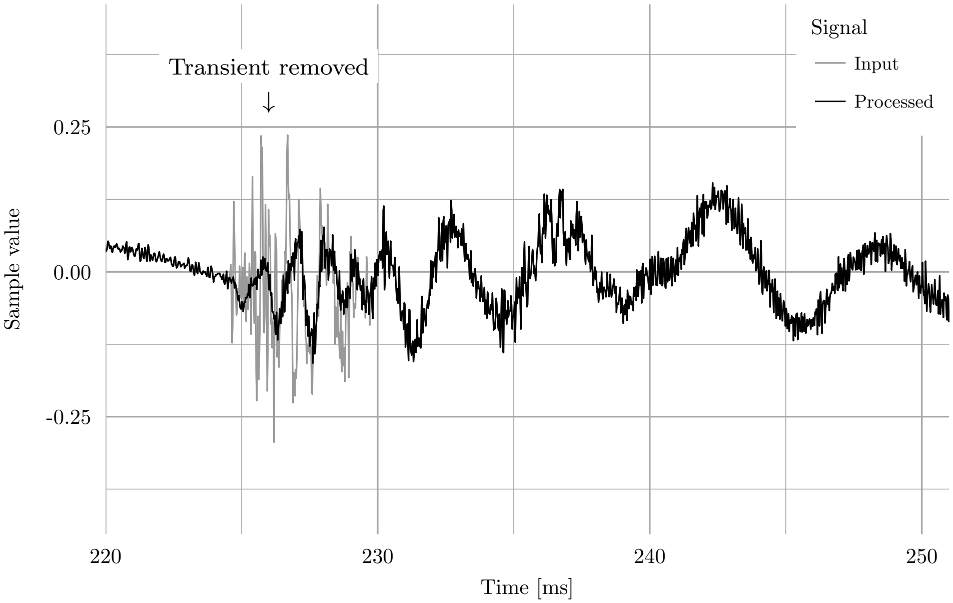


Figure 1.4: Erroneous reduction of transients by inappropriate processing of a music signal with a standard AR-model-based impulse restoration algorithm [39, Chs. 5.2.3.2 and 5.3.1]. In this case, the attack of a synthetic drum sound (in [64]) was removed.

the desired signal is assumed to be speech and speech pauses are used to estimate the hum parameters. As a consequence, this algorithm is highly prone to estimation errors if the desired signal does not contain pauses, as is the case in many music signals. Although the algorithm proposed in [67] is not restricted to speech signals, it is based on the assumption that a hum disturbance is present in the input signal. To reduce the disturbance, narrow-band notch filters are placed across the frequency spectrum. A subjective evaluation presented in [67] documents that a substantial quality improvement can be achieved with the proposed algorithm. However, the evaluation is restricted to signals containing severe hum disturbances. It can be expected that applying a large number of notch filters will lead to a signal degradation, which becomes especially noticeable if no hum disturbance is present. Hence, none of the aforementioned algorithms solves the problem of being robust against a large range of desired signals and SNRs.

1.3.3 *Restoration of Broadband Noise*

The restoration of broadband noise in audio signals has been an active field of research for more than 50 years. Manfred Schroeder was probably the first to develop a system to reduce continuous noise in speech signals in 1960. His patent [68] describes a system that splits the input signal into multiple frequency bands, atten-

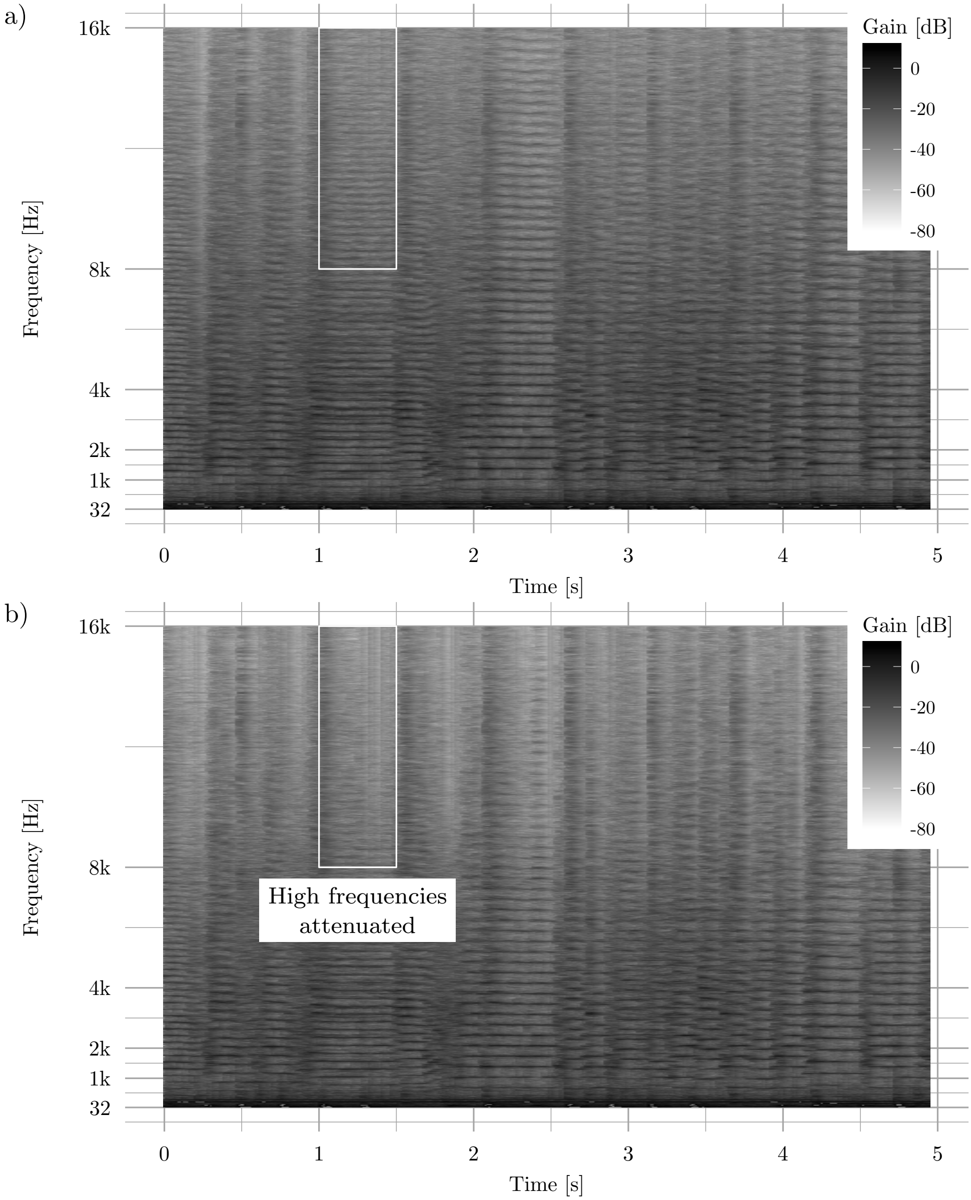


Figure 1.5: Loss of high-frequency energy caused by inappropriate processing of a music signal with a standard AR-model-based impulse restoration algorithm [39, Chs. 5.2.3.2 and 5.3.1]. The plots show spectrograms of an excerpt of [65] that features a distorted guitar. a) Spectrogram of the unprocessed input signal. b) Spectrogram of the signal that has been processed by the impulse restoration algorithm.

uates frequency bands with a low SNR, and combines the scaled frequency bands to obtain a noise-reduced output signal. This principle—applying a time-varying (real-valued) gain to each frequency band—is still the basis of many single-channel audio enhancement algorithms. Although noise reduction in the time-domain is possible, frequency-domain methods are usually preferred as the spectro-temporal characteristics of the audio signal can be exploited and the computational complexity is reduced (e.g., through the fast Fourier transform). These methods typically consist of block-based spectral analysis of the input signal, subsequent filtering and finally block-based synthesis to obtain the time-domain output signal. To alleviate spectral leakage between neighboring frequency bins and to reduce audible artifacts caused by time-variant filtering, analysis and synthesis windows are commonly used, resulting in the well-known weighted overlap-add (WOLA) processing [69, 70]. Estimating the noise characteristics is also typically performed in the frequency domain, leading to the schematic diagram shown in Figure 1.6.

In [71] the *spectral subtraction* method is proposed, which can be seen as the basis of many noise reduction algorithms. In this method the short-time Fourier transform (STFT) coefficients of the input signal with low SNR are attenuated, coining the well-known term short-time spectral attenuation (STSA). While the SNR in the individual frequency bands is not changed, the overall broadband SNR can still be improved [72]. A major challenge with STSA-based noise reduction algorithms is to avoid unnatural sounding residual noise, often denoted as *musical noise*. Research on single-channel audio signal enhancement has been very active since the publication of the spectral subtraction algorithm and a number of different methods have been devised [38, 39, 73, 74, 75]. Typically these methods differ in the gain rules (also called weighting functions) that are used to perform the frequency-domain filtering [75]. Widely used gain rules are based on, e.g., spectral subtraction [71] and Wiener filtering [76]. Other gain rules are based on the minimum mean-square error (MMSE) of the short-time spectral amplitudes, or their logarithm, [77, 78], and have been shown to result in less processing artifacts compared to earlier gain rules. Due to their widespread application, these algorithms can be considered state of the art. Over the years, various extensions of STSA-based algorithms have been proposed to further reduce processing artifacts and enhance the audio quality. For example, it has turned out that in many cases it is beneficial to limit the maximum amount of attenuation [79], which may even take psychoacoustic properties of the human auditory system into account [80, 81].

Although most of the STSA-based noise reduction algorithms have been designed for speech communication applications they have also been applied successfully to music signals [38, 39, 82, 83]. It should however be noted that the aim of broadband noise restoration in speech communication applications is generally different from the aim in archive audio applications. First, in speech communication applications the input signal is typically assumed to be a noisy recording of a single speaker, while in archive audio applications the input signal is generally more complex, e.g., a noisy music recording. Second, in speech communication applications the latency and the computational complexity should typically be kept low, while both are of minor concern in archive audio applications. Third, the general goal in speech

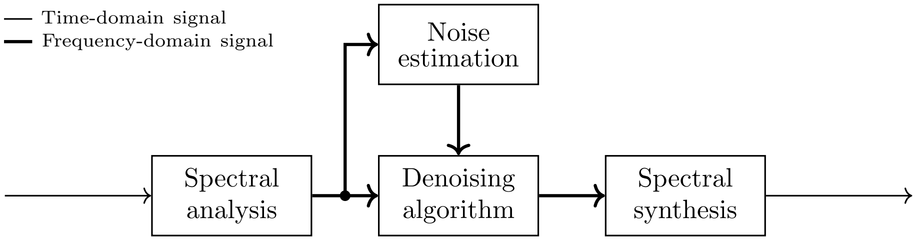


Figure 1.6: Frequency-domain denoising.

communication applications is typically the enhancement of the speech intelligibility, while in archive audio applications the main goal is to obtain a pleasant-sounding high-quality restoration result.

Besides STSA-based algorithms, some recently proposed noise reduction algorithms exploit assumptions about the time-frequency structure of the desired audio signal and specifically aim at signals different from speech. In [84] a noise reduction algorithm is proposed which takes the neighborhood for each time-frequency point of the disturbed input signal into account to compute time and frequency dependent attenuation factors. In [85] an algorithm is proposed that performs a sparse approximation of the noisy input signals, also taking the neighborhood of each time-frequency point into account. Both algorithms are reported to yield increased restoration quality compared to STSA-based algorithms. In [86] an algorithm is proposed that incorporates an AR model for the desired signal to achieve low signal degradation. Broadband noise restoration algorithms based on the Bayesian framework may have advantages in critical applications, but in general are not (yet) expected to outperform STSA-based algorithms [39].

With the exception of [85], all aforementioned broadband noise restoration algorithms require the noise characteristics to be available. Early noise estimation algorithms for speech communication applications rely on the use of a voice activity detector (VAD) [87, Ch. 33.3.2] to determine noise-only sections in the input signal. As an accurate estimate of the noise PSD is crucial for STSA-based noise reduction algorithms, research on this topic has received great attention in recent decades. On the one hand sophisticated VAD methods have been developed, e.g., [88, 89]. On the other hand, noise PSD estimation algorithms have been proposed that do not rely on a VAD and are able to update the noise PSD estimate even during speech activity, which is important to cope with non-stationary noise [74, 75]. Algorithms based on minimum statistics compute a noise PSD estimate by tracking minima of the short-time PSD estimates of the noisy input signal within a sliding time window [90, 91]. Other algorithms compute a noise PSD estimate by recursive smoothing of the short-time periodogram coefficients of the noisy input signal, where different methods have been proposed to determine the smoothing constants. In the minima controlled recursive averaging (MCRA) and improved MCRA (IMCRA) algorithms, the smoothing constants are determined based on the ratio of the short-time PSD estimate of the noisy input signal and its minimum within a certain time window

[92, 93]. In [94] the smoothing constants are determined based on a statistical model for the speech presence probability (SPP). Quantile-based algorithms compute an estimate of the noise PSD as a low quantile of the short-time PSD estimates of the noisy input signal [74, 95, 96]. In [97] an algorithm which simultaneously performs signal activity detection and noise PSD estimation based on dynamic Bayesian networks is proposed, specifically for the application with music signals. However, the evaluation is restricted to comparatively low SNRs around 15 dB.

It should be noted that all aforementioned noise PSD estimation algorithms except [97] have been designed for speech communication applications and assume that the desired signal contains a number of pauses. The requirements for archive audio restoration, however, are substantially different. While in speech communication applications the input signal is typically assumed to be a noisy recording of a single speaker, in archive audio restoration the signals are much more diverse and complex, often with hardly any pauses. Furthermore, the typical SNRs of speech communication signals are often below 20 dB [98], while the SNRs in media archives vary largely, with many signals even being noise-free. As a consequence, the direct application of the aforementioned noise PSD estimation algorithms to non-speech signals, e.g., music, in many cases leads to large estimation errors. Therefore, many audio restoration systems require the user to select noise-only sections in order to compute an accurate estimate for the noise PSD and obtain a high-quality restoration result. Figure 1.7 shows the noise PSD estimation results obtained with the minimum statistics-based algorithm, an SPP-based algorithm, the IMCRA algorithm and the estimation algorithm proposed in Chapter 5 of this thesis. It can be observed that all noise PSD estimation algorithms except the proposed algorithm significantly overestimate the noise PSD. This is due to a violation of the assumptions regarding the clean signal on which these algorithms are based. As can be seen in subplot a), the input signal contains no pauses where the noise PSD is accessible. Large noise PSD estimation errors in turn may result in a large signal degradation when the noise PSD estimate is used in a broadband noise restoration algorithm. For example, in Chapter 5 we show that when using the minimum-statistics-based noise PSD estimate in the state-of-the-art MMSE STSA noise reduction algorithm, the perceptual quality of the processed signal in many cases is even lower than the quality of the unprocessed input signal.

1.3.4 *Automatic Restoration Workflow*

A crucial aspect of a typical audio restoration workflow is the order in which the individual disturbance types are removed [39, 100]. This is because different disturbances may mask each other, or the reduction of a certain disturbance may change the characteristics of another disturbance, making it more difficult to estimate its parameters. For example, the presence of impulsive disturbances can be detrimental to hum reduction, since impulses may lead to perceivable ringing artifacts caused by notch filters for hum reduction. In addition, the presence of hum may impede the PSD estimation of broadband noise.

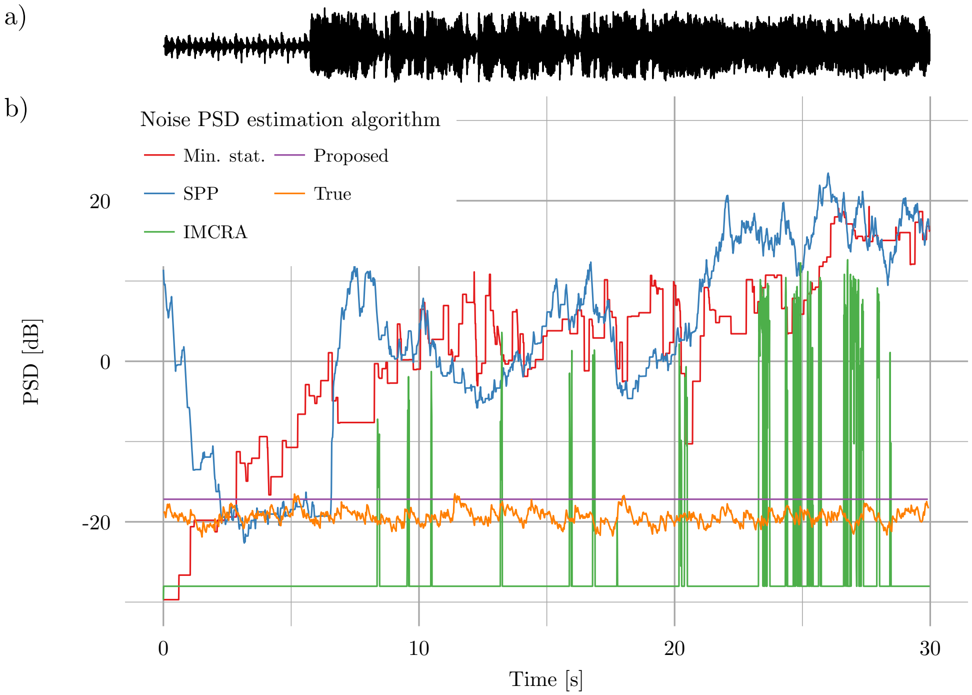


Figure 1.7: Noise PSD estimates obtained with state-of-the-art algorithms and the algorithm proposed in Chapter 5 for a music signal (excerpt from [99]). a) Noisy time-domain input signal. b) Estimated noise PSDs in the frequency bin at approximately 2 kHz. The SNR in this bin is approximately 40 dB and the true noise PSD was estimated as the recursively smoothed short-time periodogram coefficients of the noise signal.

In a typical audio restoration workflow, disturbances are handled in the following order (see Figure 1.8): 1. restoration of impulsive disturbances, 2. restoration of hum disturbances, 3. restoration of broadband noise. For each disturbance type, first the presence of the disturbance is determined, followed by a potential reduction of the disturbance. The output of the processing chain is a signal with all detected disturbances reduced.

1.3.5 *Performance Measures*

As already mentioned, the main objective of audio restoration algorithms is to achieve a high perceptual quality of the processed signal. The human auditory system has exceptional capabilities when it comes to noticing the tiniest differences to what it has learned sounds *normal*. The design of audio restoration algorithms therefore involves bridging the gap between the auditory sense and a mathematical description of what “sounds good” in terms of cost functions and objective perfor-

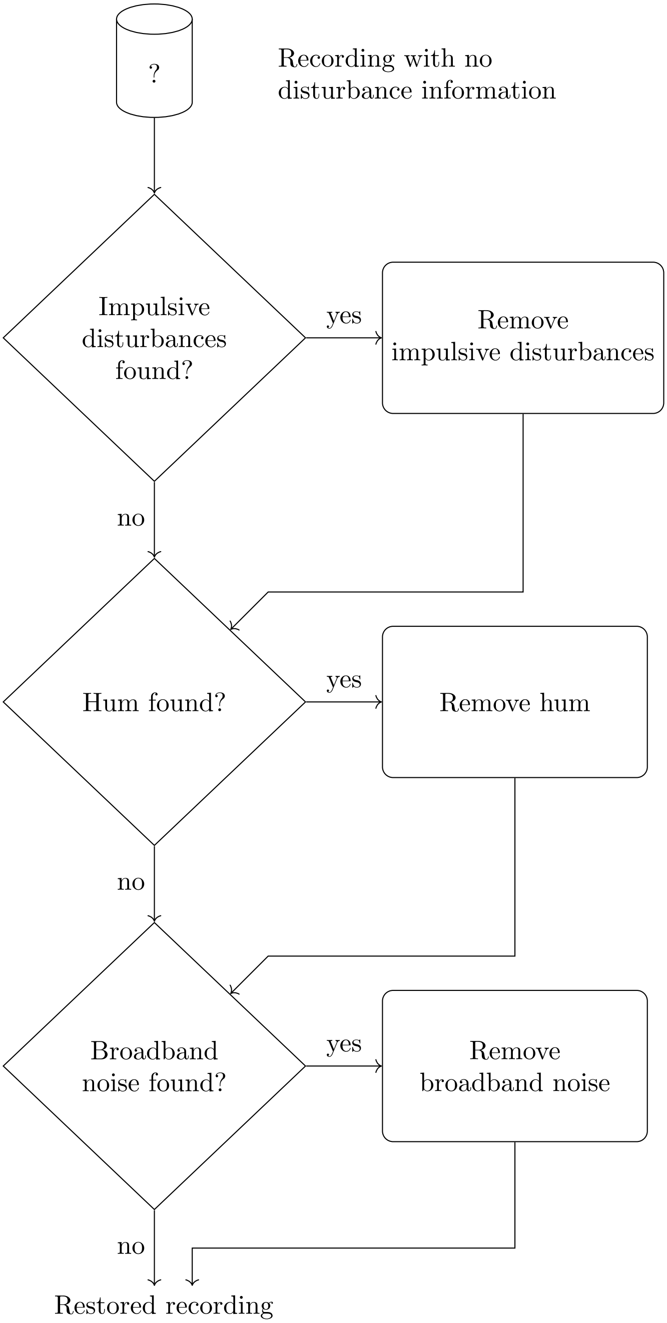

?
Recording with no disturbance information
Impulsive disturbances found?
yes
Remove impulsive disturbances
no
Hum found?
yes
Remove hum
no
Broadband noise found?
yes
Remove broadband noise
no
Restored recording

Figure 1.8: Automatic audio restoration workflow.

mance measures. To evaluate the developed algorithms, it is important to determine the quality of the obtained restoration results and, in doing so, to take the human auditory system into account. Generally, two approaches can be taken [101]: *subjective* evaluation, based on listening experiments with human listeners, and *objective* evaluation, based on measures that can be described mathematically.

Subjective evaluation based on listening experiments has the major advantage of directly determining the audio quality as perceived by human listeners. However, these experiments are usually time-consuming—not only because the tests themselves have to be performed, but also because appropriate listeners have to be acquired, e.g., listeners that have experience in rating restoration algorithms and are able to notice possible processing artifacts. Furthermore, it may be hard to determine whether individual subjective evaluation results actually reflect the auditory aspect that is sought for, e.g., whether a listener takes the quality of the desired signal into account while the experiment asks to rate the residual disturbance. In order to minimize these effects, proper instruction is required, and effects like fatigue have to be avoided. Several procedures to perform and evaluate listening tests have been proposed, e.g., the method of paired comparison [102, 103, 104] or the multiple stimulus test with hidden reference and anchor (MUSHRA) [105]. While subjective evaluation in general leads to very accurate perceptual ratings, their applicability is usually restricted to a small number of test signals—the number of required ratings grows linearly with the number of signals for the MUSHRA test and even quadratically for the method of paired comparison [102]. As a consequence, subjective evaluation is not feasible for a large number of signals.

As an alternative, objective performance measures can be used to obtain quality ratings for a large number of signals. Although objective measures are not yet able to yield ratings that correlate well with subjective ratings under all conditions, they often yield valuable information that can serve as an indication for the performance of audio restoration algorithms. While objective measures have been proposed for speech signals [87] that do not require the clean signal as a reference (so-called *non-intrusive* measures), we will only consider *intrusive* measures that are based on computing a difference to the clean reference signal. The SNR improvement between the unprocessed input signal and the output signal of the restoration algorithm is a purely mathematical measure that describes the relative change in desired signal power to disturbance signal power but usually shows little correlation to the perceptual quality [87]. While more advanced objective measures, e.g., the cepstral distance [106] and the Itakura-Saito distance [107], often show good correlation with the subjective impression, they have typically been designed for speech signals [87]. As this thesis addresses the restoration of a large diversity of desired signals, an objective measure is required that works well with speech signals as well as with non-speech signals, e.g., music. In this regard, the perceptual evaluation of audio quality (PEAQ) measure represents a viable solution that has been specifically designed to be applied with many types of audio signals [108, 109, 110]. The PEAQ measure computes a so-called objective difference grade (ODG) from the difference between the internal representations of the signal under test and the clean reference signal. The ODG ranges from -4, corresponding to "very annoying" impairments, to 0, corresponding to "imperceptible" impairments [109]. To determine the internal representation of a signal, properties of the human auditory system, e.g., the non-linear perception of pitch and masking effects, are taken into account. Although the PEAQ measure was designed primarily to rate artifacts generated by audio coding algorithms [109], we believe it is still meaningful to use this measure to rate the large number of test signals that are used for evaluating the considered audio restoration algorithms. On the one hand, this can be justified by the fact that different disturbance types, e.g., broadband noise and digital errors, were considered during the development of the PEAQ measure [108]. On the other hand, in informal listening tests we found that ODG ratings generally correspond well with the subjective impression. In general, we believe that the loss in precision by using objective measures instead of subjective listening experiments is outweighed by the gain in generality by being able to use a large number of test signals. A number of test signals are available for listening on the websites accompanying Chapters 2 and 5, along with their respective ODG ratings [111, 112].

1.4 Contributions and Thesis Outline

The main goal of this thesis is to develop algorithms that allow for an unsupervised restoration of large numbers of audio recordings, with a large diversity regarding the desired signal type, the disturbance type and the intensity of the disturbance. More specifically, we propose novel algorithms to robustly detect three important disturbance types in archive audio restoration applications, i.e., **impulsive disturbances**, **hum disturbances** and **broadband noise**. A key element in the design of these algorithms is the desire to allow for an automatic restoration, resulting in the smallest possible amount of signal degradation, and especially protecting signals that do not contain perceivable disturbances from detrimental processing with audio restoration algorithms. In addition, for hum disturbances and broadband noise we propose novel algorithms to accurately estimate the disturbance parameters that are required to reduce these disturbances. Furthermore, we present an overview and an evaluation of three state-of-the-art hum reduction algorithms. A schematic overview of the thesis is shown in Figure 1.9. The chapters on the three disturbance types are self-contained, as the algorithms for each disturbance type are fundamentally different regarding the signal model and, as a consequence, their functionality.

In order to alleviate robustness issues of existing impulsive disturbance restoration algorithms, in **Chapter 2** we propose a machine-learning-based algorithm to classify frames of the input signal as either clean or disturbed. In doing so, we specifically address the fact that many existing impulse restoration algorithms lead to a quality degradation for undisturbed signals. The proposed algorithm is based on supervised learning, using a logistic regression model and using features that are computed from the appropriately prewhitened input signal. In contrast to the detection stages of typical impulse restoration algorithms that have a time resolution in the order of the sampling interval, the proposed impulse classification algorithm uses relatively long 1 s-frames to achieve a high classification accuracy. The evaluation results with a large number of test signals show that well-known AR-model-based impulse restoration algorithms are prone to a significant number of false alarms, especially for high input SNRs and undisturbed signals. It is also shown that combining the proposed impulsive disturbance classification algorithm with a state-of-the-art AR-model-based impulse restoration algorithm leads to an increase in overall restoration quality, especially protecting signals with a high SNR and undisturbed signals from a detrimental impulse restoration processing. This work was published in the Journal of the Audio Engineering Society [113].

Hum disturbances can usually be removed effectively if the hum frequencies are known. However, as described in Section 1.3.2, existing algorithms to estimate hum disturbance parameters make certain assumptions about the input signal that prevents their use for automatic archive audio restoration. Therefore, in **Chapter 3** we propose an algorithm to detect hum in audio recordings and to estimate all required hum disturbance parameters, i.e., the frequencies of the hum partial tones, and their start and end times. The proposed algorithm uses a quantile-based statistical analysis of the short-time PSD estimates of the input signal to detect stable hum tones and uses post-processing to increase the accuracy of the detection. The accuracy of the frequency estimation increased by means of adaptive notch filters that converge towards the true frequencies of the hum partial tones. Evaluation results with real and artificial test signals show that most perceivable hum disturbances are detected with a low false alarm rate and that the hum parameters are estimated with a high accuracy. This work was published in the Journal of the Audio Engineering Society [114].

In **Chapter 4** we compare the performance of three state-of-the-art hum reduction algorithms with regard to the amount of disturbance reduction and signal degradation. More specifically, based on an evaluation with different desired signals and artificial as well as real hum disturbances, we analyze comb filters, subband comb filters, and notch filters. The evaluation results indicate that the performance of the three considered hum reduction algorithms differs significantly. While comb filters generally allow for the largest amount of hum reduction, they also result in the largest amount of signal degradation. Subband comb filters represent a compromise between the amount of hum reduction and signal degradation by splitting the input signal into a low and a high frequency band and only processing the low frequency band with a comb filter. Notch filters provide the highest flexibility as they can be placed on individual hum partial tones and their attenuation can be adjusted individually, compared to comb filters that place notches at integer multiples of the fundamental frequency. This work was published in the Proceedings of the 132nd Audio Engineering Society Convention [115].

In order to perform high-quality broadband noise restoration, an accurate estimate of the noise PSD is required. As mentioned in Section 1.3.3, existing noise PSD estimation algorithms have insufficient robustness for diverse audio material, prohibiting their application for unsupervised automatic audio restoration. In many recordings broadband noise is caused by insufficiencies of the original carrier and can therefore be assumed to be rather stationary for the complete recording. Based on this observation, in **Chapter 5** we propose a novel noise PSD estimation algorithm, assuming that the noise PSD is constant and that the short-time periodogram coefficients of the broadband noise follow an exponential distribution. The noise PSD is estimated as the mean value of the exponential distribution that corresponds to the empirical distribution of the truncated short-time periodogram coefficients of the disturbed input signal. In addition, the proposed algorithm provides a confidence measure reflecting the reliability of the noise PSD estimate, which can be used to decide whether restoration should be applied or not in a certain frequency band. Evaluation results with a large number of desired signals and different artificial and real-world broadband noise disturbances show that the proposed algorithm yields significantly lower noise PSD estimation errors compared to the state-of-the-art minimum statistics algorithm for a large range of SNRs. The evaluation results also show that using the proposed noise PSD estimates in the MMSE STSA noise reduction algorithm allows for an unsupervised restoration, leading to an increased perceptual quality for the majority of signals and only marginal signal degradation for practically undisturbed signals. This work has been submitted to the Journal of the Audio Engineering Society [116].

The proposed algorithms constitute important steps for automatic restoration of audio recordings, over a wide range of SNRs and input signals, which are typically encountered in large audio archives.

1.4.1 *Publications*

Chapters 2 to 5 contain the contents of the following articles:

> M. Brandt, J. Bitzer, "Hum Removal Filters: Overview and Analysis," *Proceedings of the 132nd Audio Engineering Society Convention*, Budapest, Hungary (2012 Apr.).
>
> M. Brandt, J. Bitzer, "Automatic Detection of Hum in Audio Signals," *Journal of the Audio Engineering Society*, vol. 62, no. 9, pp. 584–595 (2014 Oct.). `https://doi.org/10.17743/jaes.2014.0034`
>
> M. Brandt, S. Doclo, T. Gerkmann, J. Bitzer, "Impulsive Disturbances in Audio Archives: Signal Classification for Automatic Restoration," *Journal of the Audio Engineering Society*, vol. 65, no. 10, pp. 826–840 (2017 Oct.). `https://doi.org/10.17743/jaes.2017.0032`
>
> M. Brandt, S. Doclo, J. Bitzer, "Automatic Noise PSD Estimation for Archive Audio Restoration," *Submitted to the Journal of the Audio Engineering Society* (2018 Mar.).

In addition, the following articles have been published whose contents have not been incorporated in this thesis:

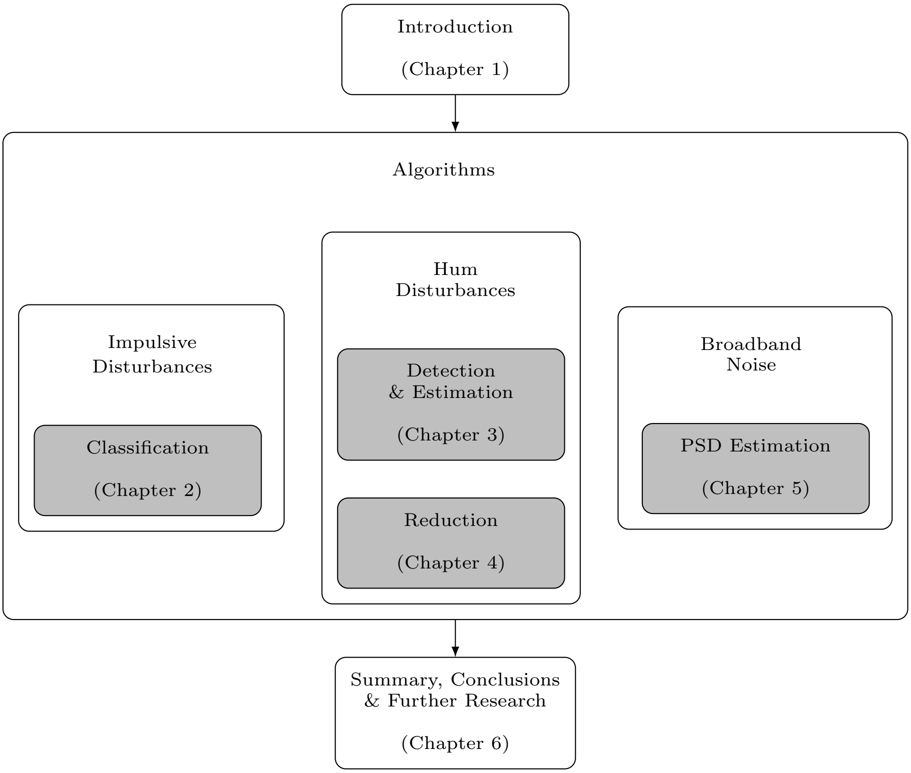


Figure 1.9: Schematic overview of the thesis.

M. Brandt, J. Bitzer, "Optimal Spectral Smoothing in Short-Time Spectral Attenuation (STSA) Algorithms: Results of Objective Measures and Listening Tests," *Proceedings of the 17th European Signal Processing Conference*, Glasgow, England, pp. 199–203 (2009 Aug.).

J. Bitzer, M. Brandt, "Speech Enhancement by Adaptive Noise Cancellation: Problems, Algorithms, and Limits," *Proceedings of the 39th International Audio Engineering Society Conference*, Hillerød, Denmark (2010 June).

M. Brandt, J. Bitzer, "Detection of Hum in Audio Signals," *Proceedings of the 12th International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Tel Aviv, Israel (2010 Aug.).

M. Brandt, T. Schmidt, J. Bitzer, "Evaluation of a New Algorithm for Automatic Hum Detection in Audio Recordings," *Proceedings of the 130th Audio Engineering Society Convention*, London, England (2011 May).

M. Ruhland, S. Goetze, M. Brandt, S. Doclo, J. Bitzer, "A New Approach for Reduction of Supergaussian Noise Using Autoregressive Interpolation and

Time-Frequency Masking," *Proceedings of the 13th International Workshop on Acoustic Signal Enhancement (IWAENC)*, Aachen, Germany (2012 Sept.).

M. Ruhland, J. Bitzer, M. Brandt, S. Goetze, "Reduction of Gaussian, Supergaussian, and Impulsive Noise by Interpolation of the Binary Mask Residual," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1680–1691 (2015 Oct.). `https://doi.org/10.1109/TASLP.2015.2444664`

2

CLASSIFICATION OF IMPULSIVE DISTURBANCES

This article presents a new algorithm to classify whether each one-second long frame of an audio recording contains impulsive disturbances or not. The developed classification algorithm is based on supervised learning and appropriate prewhitening of the input signal. It is shown that existing impulse restoration algorithms suffer from degradation of the desired signal if the input SNR is high and if no manual parameter adjustment is possible, which makes automatic restoration of large amounts of diverse archive audio material infeasible. The proposed classification algorithm can be used as a supplement to an existing impulse restoration algorithm to alleviate this drawback. An evaluation with a large number of test signals shows that a high classification accuracy can be achieved, making fully automatic impulse restoration possible.

2.1 Introduction

The number of audio documents that are stored in archives around the globe is immense. Since the development and widespread introduction of the phonograph at the end of the 19th century, all kinds of music recordings, speeches, interviews, film sound tracks, and other audio documents have accumulated and represent the world's audio heritage. Due to age, improper storage, and shortcomings of the original storage media, the degradation of audio signal quality is a common problem, especially in historic recordings. Impulsive disturbances are one of the

---

This chapter contains a copy of the article

M. Brandt, S. Doclo, T. Gerkmann, J. Bitzer, "Impulsive Disturbances in Audio Archives: Signal Classification for Automatic Restoration," *Journal of the Audio Engineering Society*, vol. 65, no. 10, pp. 826–840 (2017 Oct.). `https://doi.org/10.17743/jaes.2017.0032`

While the layout of the article has been adapted for a uniform presentation within this thesis, the contents printed here are identical to those in the published article.

most prominent types of disturbance, besides broadband hiss and hum. These so-called *click* and *crackle* phenomena are caused by deficiencies of grooved recording media, e.g., wax cylinders, shellac, and vinyl discs. After digitalization and storage in archives, these defects remain in the digital version of the signal.

To improve the listening experience, recordings that suffer from impulsive disturbances can be processed by impulse restoration algorithms that aim at removing the disturbance impulses and obtaining an estimate of the original clean signal. For these restoration algorithms to achieve optimum results, however, their parameters have to be adjusted for each recording individually, in order to make the algorithm detect and remove most of the disturbance impulses while leaving the desired signal unimpaired. In doing so, the optimum choice of parameters depends substantially on the relative level of the disturbance impulses compared to the level of the desired signal. Existing impulse restoration algorithms are typically not able to distinguish between actual disturbance impulses and certain impulse-like elements of the desired signal with a similar level, e.g., drum transients, guitar pickings or sharp synthesizer attacks.

In the specific context of audio archive restoration, individual parameter adjustment for each recording is usually not feasible. This is due to the sheer amount of audio material that is currently stored in archives around the globe: The Library of Congress, e.g., reports about more than 3.5 million audio media in 2014 [23]. Millions of further recordings are stored in a multitude of archives in the United States alone [117]. Due to the fact that grooved recording media were superseded by media that inherently are not subject to impulsive disturbances (e.g., tape, compact disc), only a subset of the recordings that are stored in an archive are prone to this type of disturbance. Unfortunately, in many cases the original type of medium of a digitally stored recording is unknown. Therefore, the decision whether a recording should be processed with an impulse restoration algorithm often can only be based on an analysis of the signal itself. As a consequence, the overall restoration quality for a full archive depends on the robustness of the restoration algorithm against a large range of input SNRs—in many cases the majority of recordings may even be undisturbed while some recordings contain severe impulsive disturbances. And while existing impulse restoration algorithms achieve high quality restoration results for the class of signals that contain typical impulsive disturbances, e.g., in a recording copied from a vinyl disc, we show in Section 2.4.4.3 that degradation of the desired signal can occur if a recording does not contain impulsive disturbances at all. Therefore, the main challenge in archive restoration comes down to the diversity of the material. Examples for especially challenging recordings, in this regard, are radio documentaries or live recordings of the program that had been broadcast by a radio station, containing a sequence of music pieces from differing original media, alternating with voice-overs from a studio speaker.

2.1.1 *Main Idea*

The main idea of this paper is to alleviate the robustness problems of existing impulse restoration algorithms by *classifying* whether a recording contains impulsive disturbance or not. Specifically, we propose a classification algorithm that determines for each frame of 1 s duration of the input signal whether impulsive disturbances are present or not. This information can then be used, for example, to control an existing impulse restoration algorithm and only restore those frames that actually contain impulsive disturbances.[1] In order to achieve accurate classification, the input signal is preprocessed in a prewhitening step. This is done in a blockwise manner using blocks of $\approx 23\,\mathrm{ms}$ length.

As the classification algorithm provides a confidence measure for the disturbance of a frame, it is possible to adjust the classification behavior either in the conservative or progressive direction.

An overview of the proposed classification algorithm, consisting of the prewhitening and classification stages, is shown in Figure 2.1, each stage with its associated signals and notation.

2.1.2 *Related Work*

For quite a number of years, attempts have been made to detect and suppress impulsive disturbances from wax cylinders, gramophone, and vinyl records. As a consequence, a number of algorithms have been developed that are able to yield high quality restored signals if their parameters are adjusted properly to a signal at hand. Most of these algorithms consist of two steps: after detecting the affected signal portions, impulses are removed by extrapolating the known signal surrounding the affected portions. Early detection schemes were typically based on first enhancing impulsive elements in the input signal and then applying cleverly devised threshold criteria to detect the individual disturbance impulses. Enhancing impulsive elements in the input signal was, for example, based on high-frequency pre-emphasis [50] or on subtracting the median filtered version from the input signal [51]. Early interpolators consisted in replacing the damaged part of the signal with silence or linear interpolation of the neighboring sample values [51]. Restoration methods based on linear prediction, introduced in [118, 119, 120, 121], constituted a big leap forward concerning the quality of restoration and are now state of the art in commercially available solutions. More recent interpolation approaches based on true linear prediction [59] or frequency-warped prediction achieve high audio quality even for gap lengths of around 45 ms [60, 61]. Different approaches have been developed that aim at improving the impulse detection accuracy on the one hand, and the replacement of affected samples on the other hand. E.g., the two-

---

[1] The presented *classification* algorithm that works with 1 s frames is not a replacement for *detection* stages working on the sample-by-sample level that are part of typical impulse restoration algorithms.

channel approach proposed in [122] gains advantage from using two signals obtained with a stereo replay cartridge, compared to only single-channel processing. Other recent methods use bidirectional processing [53] or click templates [55] to increase the detection accuracy. In [56, 57, 58] detection (and interpolation) schemes based on machine learning techniques have been proposed. Detection methods that are based on the Bayesian philosophy, developed in [39], are shown to have advantages in critical applications but with high computational requirements.

In recent years, classification algorithms based on deep learning have shown remarkable results for a variety of audio signal processing tasks, e.g., audio tagging [123], or acoustic event detection [124]. In this paper, however, we use a traditional classifier, due to the fact that deep learning based approaches are known to often suffer from limited generalization capabilities to unknown data and that their training is computationally expensive. The proposed algorithm achieves high classification performance with comparatively low computational requirements.

2.1.3 *Paper Structure*

The structure of this paper is as follows. In Section 2.2 the characteristics of the signals to be processed are described. A thorough explanation of the proposed classification algorithm is given in Section 2.3. To analyze the performance of the proposed algorithm, the evaluation method and the results for a large number of test signals are given in Section 2.4.

2.2 Signal Model

In the context of audio restoration, disturbing impulses are usually assumed to be localized degradations of the signal that are of short duration—ranging from 20 µs to 4 ms [39], corresponding to about 1–200 samples at a typical sampling rate of 44 100 Hz. For wax cylinders, shellac or vinyl records, the disturbing impulses are usually caused by scratches and dust particles in the grooves of the medium.

Depending on the severity of the damage, clicks can be assumed to be either additive to the clean signal or—in the case of severe damage—fully replacing the original signal (cf., [39, p. 100]). In this article we will assume that the impulsive disturbances are additive, i.e.,

$$x[n] = s[n] + d[n] \quad \text{for } 0 \leq n < L, \tag{2.1}$$

with the sample index n, L the length of the signals, the disturbed signal $x[n]$, the clean (unobservable) signal $s[n]$ and the sparse disturbance $d[n]$ (with $d[n] = 0$ for most n).

To evaluate the proposed algorithm and to determine optimum model parameters we use artificial disturbances. This has the major advantage of obtaining a fully controlled environment—i.e., the location of the clicks and the SNR of the disturbed

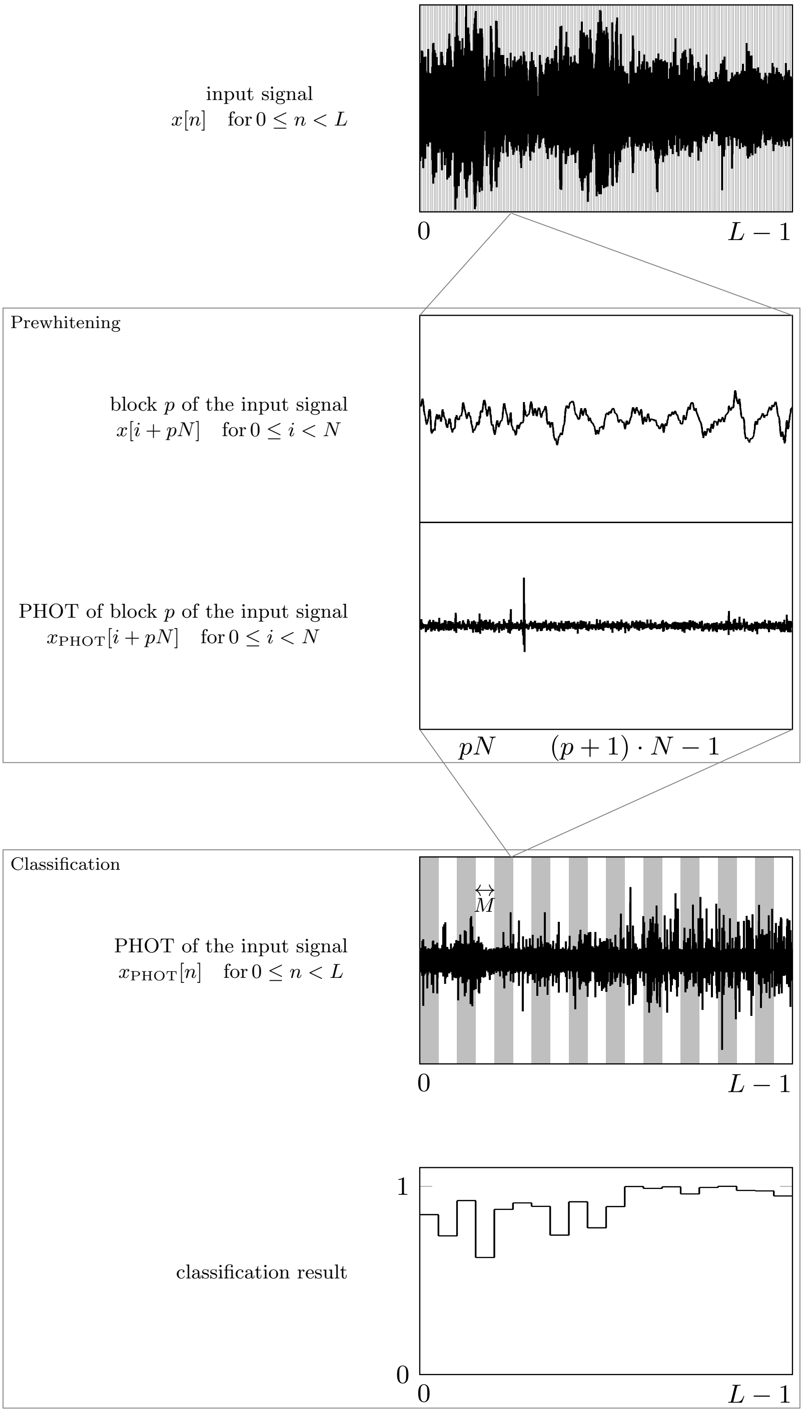

input signal
$x[n]$ for $0 \leq n < L$
0
$L-1$
Prewhitening
block p of the input signal
$x[i+pN]$ for $0 \leq i < N$
PHOT of block p of the input signal
$x_{\mathrm{PHOT}}[i+pN]$ for $0 \leq i < N$
pN
$(p+1) \cdot N - 1$
Classification
$\overset{\leftrightarrow}{M}$
PHOT of the input signal
$x_{\mathrm{PHOT}}[n]$ for $0 \leq n < L$
0
$L-1$
1
0
classification result
0
$L-1$

Figure 2.1: Schematic overview of the classification algorithm. The prewhitening and classification stages are shown with associated signal notation, block and frame lengths.

signal are known. Furthermore, our preliminary experiments have shown that the manual annotation of real-world signals is too time-consuming to be feasible for large amounts of audio recordings and the obtained accuracy is not sufficient to yield meaningful evaluation results. In addition, it is very difficult to obtain a recording of a real impulsive disturbance signal, without any desired signal, that can be used as an additive disturbance. This is due to the fact that real recordings of, e.g., the blank groove of a vinyl record, always contain additional disturbances, for example hiss or low frequency mains hum. On the one hand, processing such a real recording to remove everything except the impulsive disturbances would lead to a change in the waveform of the impulses, for example caused by the response of the hum removal filter. On the other hand, using the unprocessed recording, including hiss and hum, makes it very difficult to properly set the SNR of the artificially disturbed signals to allow for a precise evaluation. However, we have found in informal experiments that the performance of autoregressive (AR) model based impulse restoration algorithms when used with signals containing these artificial disturbances is comparable to the performance for real disturbed signals. For the reasons explained above, we did not include signals containing real disturbances in the evaluation. However, on the website that accompanies the manuscript [111] we demonstrate the performance of the proposed classification algorithm when used with real disturbances (i.e., the recording of blank grooves of shellac and vinyl discs).

In Section 2.2.1 the used model for the artificially generated disturbances will be reviewed, while in Section 2.2.2 two ways to set the SNR will be discussed.

2.2.1 *Artificial Impulsive Disturbance Generation*

Impulsive disturbances are often modeled in a probabilistic way as the output of a filter that is excited by amplitude-modulated impulses with random time of occurrence (see [38]). Different distributions for the time between impulses and for their amplitudes can be used. To generate the artificial impulsive disturbances we used a method based on [125, Sec. 3.1]. The underlying probabilistic process and its parameters were selected to fit real-world disturbed signals. More specifically, the inter-occurrence time τ (in samples) of the impulses is modeled with a gamma distribution, i.e.,

$$\begin{aligned} f(\tau; k, \Theta) &= \frac{1}{\Theta^k \cdot \Gamma(k)} \cdot \tau^{k-1} e^{-\frac{\tau}{\Theta}} \\ &\qquad \text{for } \tau > 0 \text{ and } k,\ \Theta > 0, \end{aligned} \tag{2.2}$$

with shape parameter k, scale parameter Θ and $\Gamma(\cdot)$ the gamma function (see [126]). The magnitude A of the impulses is modeled with a log-normal distribution, i.e.,

$$f(A;\mu,\sigma) = \frac{1}{A\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(A)-\mu)^2}{2\sigma^2}\right) \quad \text{for } A > 0 \text{ and } \sigma > 0, \tag{2.3}$$

with location parameter μ, scale parameter σ and $\ln(\cdot)$ the natural logarithm.

The impulsive disturbance signal is constructed by first placing unit impulses with inter-occurrence times according to the gamma distribution in Equation (2.2). The individual impulses are scaled according to the log-normal distribution in Equation (2.3) and multiplied by 1 or -1 with equal probability. To take the response of the pickup system and variations in the click-generation process into account, this intermediate signal is then filtered with a third-order Butterworth low-pass filter with time-varying cut-off frequency. Each block of 25 ms is filtered with a random cut-off frequency according to a uniform distribution between ≈ 2.2 kHz and ≈ 11 kHz. For simplicity, we did not model the duration of the impulses explicitly as in [125]. Besides, the application of the low-pass filter leads to a varying duration of the generated impulses as the length of the filter's impulse response changes in dependence on its cut-off frequency.

2.2.2 *Two SNR Concepts*

Since the disturbance signal is modeled as localized impulses with gaps between occurrences, defining an appropriate measure rating the perceptual *amount of disturbance* is not straightforward. Obviously, the average magnitude of the disturbance impulses comes into consideration as a signal sounds more disturbed as the disturbance gets louder. However, in practice the interval between impulses, i.e., the *impulse density*, is a second characteristic of the disturbance signal that is at least of equal importance. This is motivated by the fact that a large proportion of typical vinyl and shellac degradations are caused by dust and dirt particles in the grooves of the disc. The size of these particles (corresponding to the energy of the impulses) can be expected to change only little [127] compared to the number of dust particles (corresponding to the impulse density) that are distributed on the disc surface.

For this reason, throughout the article, we will consider two ways to set the SNR, either by adjusting the gain of the disturbance signal, or by adjusting the impulse density. In the first case, the disturbance signal is generated with the default parameters given in [125] (see Table 2.1), where only the gain is adjusted to obtain the desired SNR. In the second case, the scale parameter of the gamma distribution in Equation (2.2) is adjusted to obtain the desired SNR. Changing the scale parameter has the effect of changing the average time between impulses. Signals

Table 2.1: Default parameters of the impulsive disturbance generation method from [125]. The values related to the inter-occurrence time hold for a sampling rate of $f_s = 44\,100\,\text{Hz}$.

	Parameter		
	Symbol	Description	Value
Gamma distribution (inter-occurrence time)	k	Shape	0.2
	Θ	Scale	2 433.8
Log-normal distribution (impulse magnitude)	μ	Location	−3.63
	σ	Scale	0.74

demonstrating the two characteristics of the disturbances are available online on the website accompanying this article [111].

2.2.2.1 *SNR via Gain Factor*

In this case the default disturbance signal, $d_{\text{def}}[n]$, generated with the default parameters from [125], is scaled with a gain factor, i.e.,

$$d[n] = d_{\text{def}}[n] \cdot \sqrt{\frac{\sum_{i=0}^{L-1} s^2[i]}{\sum_{i=0}^{L-1} d_{\text{def}}^2[i]} \cdot 10^{-\text{SNR}/20}},$$

and added to the clean signal $s[n]$.

2.2.2.2 *SNR via Impulse Density*

Setting the desired SNR via the impulse density is based on an iterative approach. First, the scaling factor is determined for the default disturbance signal to yield an SNR of $\text{SNR}_{\text{def}} = 30\,\text{dB}$, as informal listening tests have shown that this represents a medium disturbance, corresponding well with real-world audio material, i.e.,

$$f_{\text{scale}} = \sqrt{\frac{\sum_{i=0}^{L-1} s^2[i]}{\sum_{i=0}^{L-1} d_{\text{def}}^2[i]} \cdot 10^{-\text{SNR}_{\text{def}}/20}}.$$

Second, the scale parameter Θ of the gamma distribution in Equation (2.2) that corresponds to the desired SNR is determined in an iterative manner.

If the SNR is too small, the scale parameter is increased, leading to a higher mean inter-impulse time. If the SNR is too large, the scale parameter is reduced, lowering the mean inter-impulse time. This iteration is repeated until the deviation from the desired SNR is smaller than $\Delta_{\mathrm{SNR}} = 0.1\,\mathrm{dB}$. The appendix at the end of this paper contains a table of the mean shape parameters required to obtain different SNRs.

2.3 Classification Algorithm

The complete impulsive disturbance classification algorithm is shown in Figure 2.2. In the training stage a model is trained based on artificially disturbed data to distinguish between clean and disturbed one-second long input frames using a supervised learning approach. To enhance impulses in the input signal the signal is *prewhitened* in a first step (cf., Section 2.3.1). To do so, much shorter block lengths are used in the order of 23 ms. From the prewhitened signal, a number of *features* are computed that have been selected to efficiently separate between the two classes *clean* and *disturbed* (cf., Section 2.3.2). Using these features as input data, a classifier is trained to determine the class of each frame of the input signal (cf., Section 2.3.3). In an application scenario, the resulting classification model is then used to classify whether the frames of an unknown input signal contain impulsive disturbances or not. The output of this model is not a hard binary decision but rather a probability for each frame to belong to the *clean* and *disturbed* class, respectively. This can be viewed as a confidence measure and is important information that in principle allows for deciding about the overall desired behavior of the classification algorithm: One option is to decide for a conservative strategy, which would be to classify frames to be disturbed only if the disturbance probability is very high. Another option is to reduce the number of missed impulses and accept a certain number of false alarms by classifying frames to be disturbed even if the disturbance probability is comparatively low. In conjunction with an impulse restoration algorithm, it is then possible to choose a compromise between removing all impulsive disturbances and accepting a certain amount of desired signal degradation or rather avoiding desired signal degradation with the risk of leaving some impulsive disturbances unremoved. In our experiments the threshold for assuming a frame to be disturbed is set to 0.5, making no assumptions about preferred weighting of the classes, to allow for an evaluation as general as possible.

2.3.1 *Prewhitening*

In many cases impulsive disturbances are audible even if their amplitude is very low. As a consequence, it may be a difficult task to automatically find impulses in the input signal. Therefore, existing approaches for impulse detection employ different types of prewhitening to make the disturbing impulses stand out from the desired signal (see, e.g., [38, Ch. 13]). The most common type of prewhitening is to use the prediction error signal of a *linear predictor*, which is briefly reviewed

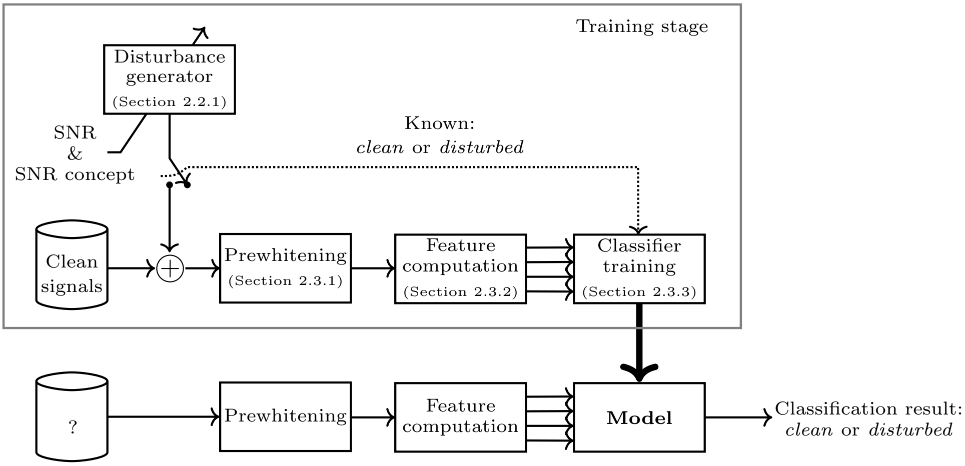


Figure 2.2: Flow diagram of the impulsive disturbance classification system.

in Section 2.3.1.1. However, since for impulsive disturbance classification we found that prewhitening based on linear prediction performs only suboptimally (see the evaluation results in Section 2.4.4.1), we also investigated *phase-only transform (PHOT)* prewhitening, which is described in Section 2.3.1.2.

2.3.1.1 *Prediction Error of a Linear Predictor*

The use of the prediction error of a linear predictor has proven to be an effective prewhitening step to reduce the energy of the desired signal $s[n]$ and make the disturbance stand out more clearly [52, 119, 128].

In forward linear prediction (see, e.g., [70]) the current sample is modeled as a linear combination of previous samples, i.e.,

$$\hat{x}[n] = -\sum_{i=1}^{P_{\mathrm{LP}}} a[i]x[n-i] + e[n], \tag{2.4}$$

where $\hat{x}[n]$ is an approximation of $x[n]$, $e[n]$ is the *prediction error*, $a[i]$ are the predictor coefficients and P_{LP} is the prediction order. The predictor coefficients for the pth input signal block of length N are determined by minimizing the least squares prediction error:

$$\begin{aligned} \mathrm{E}^{(p)} &= \frac{1}{N}\sum_{i=0}^{N-1}\left(e[i+pN]\right)^2 \\ &= \frac{1}{N}\sum_{i=0}^{N-1}\left(x[i+pN]-\hat{x}[i+pN]\right)^2, \end{aligned}$$

where the superscript $\bullet^{(p)}$ denotes values of the pth *block* (of length N) of the input signal. The block length N is typically chosen to correspond to a block length in the order of 23 ms because of the assumed short-time stationarity of the desired signal.

Depending on the prediction order and the block length, slowly-varying deterministic elements can be predicted with high accuracy, compared to stochastic elements and quickly changing parts of the signal. This has the desired effect of reducing the energy of the desired signal and thus enhancing the impulsive disturbances in the prediction error signal.

2.3.1.2 *Phase Only Transform*

The phase only transform (PHOT), also known as the phase transform (PHAT), has been successfully employed to increase the robustness of sound source localization systems in noisy and reverberant environments [129, 130] and for surface defect detection in images [131]. It is computed for the pth block of the input signal x defined in Equation (2.1) as follows:

$$X^{(p)}[k] = \sum_{i=0}^{N-1} x[i+pN] \cdot e^{-j2\pi ki/N} \tag{2.5a}$$

$$X^{(p)}_{\mathrm{PHOT}}[k] = \frac{X^{(p)}[k]}{\left|X^{(p)}[k]\right|} \tag{2.5b}$$

$$x_{\mathrm{PHOT}}[n+pN] = \frac{1}{N} \sum_{i=0}^{N-1} X^{(p)}_{\mathrm{PHOT}}[i] \cdot e^{j2\pi ni/N} \tag{2.5c}$$

with N both the DFT length and block length. The PHOT of the full-length input signal x is computed by using a weighted overlap-add method as described in [69].

The reason why the PHOT enhances transients can be illustrated intuitively. The spectral magnitude of music signals usually decays with higher frequencies [132]—Figure 2.3 shows the mean power spectral density of music signals from several decades of the 20th century. The PHOT in Equation (2.5) can be interpreted as filtering the input signal with a filter that emphasizes high-frequency content of the signal:

$$H[k] = \frac{1}{|X[k]|}.$$

As impulsive disturbances usually contain much more high-frequency energy than the target audio signal, the effect of this filter is a relative enhancement of the

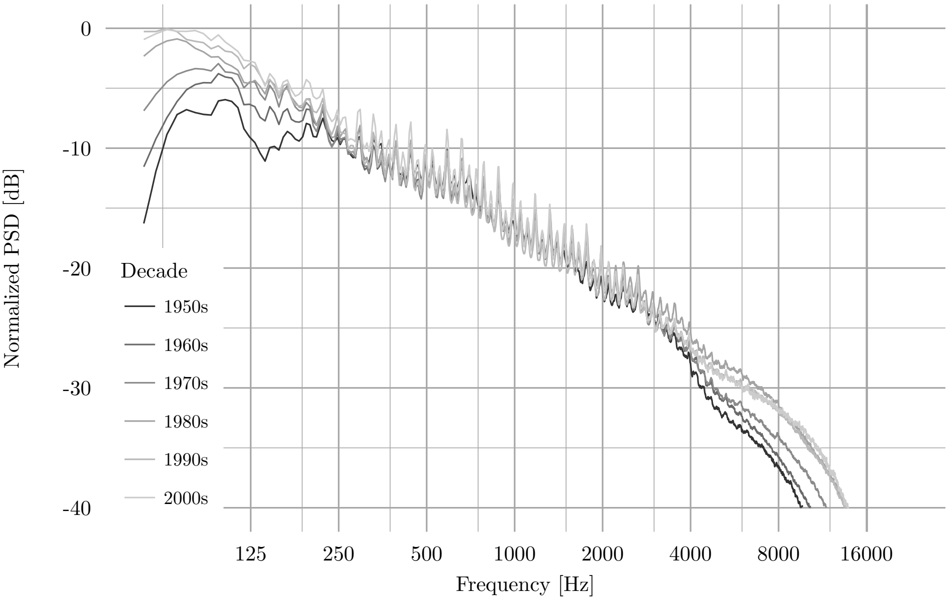


Figure 2.3: Mean power spectral density of music signals from several decades of the 20th century. This figure has been generated from the database described in Section 2.4.1. The PSDs were estimated using the Welch method and were normalized such that the overall maximum value is 0 dB. The PSD axis is clipped at −40 dB for reasons of clarity.

impulses compared to the audio signal. A more thorough examination of why the phase only transform makes irregularities stand out is given in [131].

2.3.2 *Feature Computation*

After prewhitening the signals (using the prediction error of a linear predictor or the PHOT), the features are computed for each frame of the prewhitened signal:

$$x_{\text{pre}}^{(q)}[n] = x_{\text{pre}}[n + qM] \quad \text{for } 0 \leq n < M,$$

with M the frame length of the feature computation. $\bullet^{(q)}$ denotes values of the qth frame (of length M) of the prewhitened signal. As mentioned before, we use frames of 1 s duration, corresponding to a frame length of $M = 44\,100$ samples at a sampling rate of $f_s = 44\,100\,\text{Hz}$. Informal analyses have shown that this choice represents a good compromise between classification accuracy and time resolution.

To make the feature values independent of the energy the input frames are normalized. To reduce the influence of potentially existing impulses on the level scaling, this is done in a robust way using the 5 %-*truncated standard deviation* [133]:

$$x'^{(q)}_{\text{pre}}[n] = x^{(q)}_{\text{pre}}[n] / \sigma_{x^{(q)}_{\text{pre}},\, 5\%},$$

where $\sigma_{x^{(q)}_{\text{pre}},\, 5\%}$ is the standard deviation of $x^{(q)}_{\text{pre}}$ whose 5 % smallest and greatest elements have been removed. By using the truncated standard deviation instead of the regular standard deviation, the salience of impulses possibly contained in a frame is not reduced by the normalization which is desirable to allow for good separability between clean and disturbed input frames.

For classification we have considered a variety of features (see Section 2.A.2). Using recursive feature elimination [134], we found that good performance can be achieved using the *crest factor*, i.e.,

$$C^{(q)} = \frac{\max\limits_{0 \le i < M} \left| x'^{(q)}_{\text{pre}}[n] \right|}{\sqrt{\frac{1}{M} \sum_{i=0}^{M-1} \left(x'^{(q)}_{\text{pre}}[i] \right)^2}}, \tag{2.6}$$

and the *sample kurtosis*,

$$\text{Kurt}^{(q)} = \frac{\frac{1}{M} \sum_{i=0}^{M-1} \left(x'^{(q)}_{\text{pre}}[i] - \overline{x'^{(q)}_{\text{pre}}} \right)^4}{\left(\frac{1}{M} \sum_{i=0}^{M-1} \left(x'^{(q)}_{\text{pre}}[i] - \overline{x'^{(q)}_{\text{pre}}} \right)^2 \right)^2}, \tag{2.7}$$

which are both relatively easy to compute.

2.3.3 *Classifier Training*

After computing the features as described in the previous section, they are used to train a binary classifier that labels each input frame either as *clean* or *disturbed*. The training happens in form of a supervised learning approach, using artificially disturbed signals (cf., Section 2.2.1) and the corresponding information whether a frame contains impulsive disturbances or not as training labels. As classifiers we considered L2-regularized logistic regression and a support vector machine (SVM) with radial basis function kernels, both in the implementation from [135]. The optimal hyperparameters (amount of regularization for logistic regression and SVM and kernel coefficient for SVM) were determined via 5-fold cross-validation [136]. Depending on the specific evaluation goal (cf., Section 2.4.4) either the complete data set was used for training *and* testing or the available data was split into training and test subsets. Details will be given in the respective sections.

2.4 Evaluation Method & Results

To determine the classification performance of the developed algorithm and to find optimum values for its parameters we use an evaluation based on a database of test signals and different error measures. In a first experiment, cf., Section 2.4.4.1, we optimize the parameters of the prewhitening stage, i.e., block length N, and prediction order P_{LP} for the linear predictor, and investigate the classification performance for different classifiers. In a second experiment, cf., Section 2.4.4.2, we analyze the classification performance, based on the optimized parameters, for a large database of signals unknown to the classification algorithm. A third experiment, cf., Section 2.4.4.3, investigates the audio quality improvement of three existing impulse restoration algorithms. The aim of that section is to assess the ability of these restoration algorithms to deal with a wide variety of input signals when no individual parameter adjustment is performed. A final fourth experiment investigates the audio quality improvement that is obtained when using the classification algorithm in conjunction with a standard impulse restoration algorithm based on an AR model of the clean signal [39, Ch. 5]. In all cases, the tests are performed for different SNRs and both SNR concepts. The frame length for feature computation is set to $M = 44\,100$ samples, corresponding to a frame duration of 1 s at the used sampling rate of $f_s = 44\,100\,\mathrm{Hz}$.

In Section 2.4.1 we describe the database of test signals. After that, Section 2.4.2 presents different error measures that are used to rate the classification performance in the first two experiments on the one hand and the perceptual audio quality improvement obtained in the third and fourth experiments on the other hand. In Section 2.4.3 we briefly describe the three reference impulse restoration algorithms that are used for the evaluation. Section 2.4.4, finally, presents and discusses the results of the four experiments.

2.4.1 *Test Signals*

For the development and evaluation of the classification algorithm, we used a database of clean music recordings [137] that contains 20 recordings from each of the years $1955 - -1985$, resulting in 620 clean signals. This time span was chosen since this is the main targeted period of application for the impulsive disturbance classification algorithm. Before around 1955 most commercial music recordings were distributed on wax cylinders or shellac discs, and thus can be assumed to generally contain impulsive disturbances. In contrast, recordings that have been produced after around 1985 are available in digital format and can be assumed impulsive disturbance-free. Starting at the end of the 1940s, magnetic tape recordings gained widespread popularity and coexisted with the hill-and-dale recording technologies for several decades, until the introduction of digital recording and the compact disc (cf., e.g., [4]). As a consequence, no assumptions concerning impulsive disturbances can be made for recordings from this time span and we show that it is beneficial to use an impulsive disturbance classifier.

Each test signal was a randomly selected 20 s long monaural segment of the corresponding recording from the clean music database. As the database consists of two-channel CD recordings the monaural test signals were obtained by extracting the left channels of the original recordings.

As already mentioned, we used artificial additive disturbances that were generated using the method described in Section 2.2. As we used four different SNRs plus undisturbed signals (SNR $= \infty$), and used the two SNR concepts explained above, the overall amount of test data consisted of $620 \cdot 5 \cdot 2 = 6200$ signals, corresponding to an overall duration of $6200 \cdot 20\,\mathrm{s} \approxeq 34\,\mathrm{h}$. However, due to the random nature of the disturbance signal generation, not all frames of the disturbance signal actually contain disturbance impulses. This is caused by high inter-occurrence times between the individual impulses which may exceed the frame length of 1 s. Therefore, all frames from the disturbed class that did not contain any impulses were removed from the training set to prevent two identical disturbance-free signal frames being used for classifier training.

2.4.2 *Error Measures*

This section describes both the measures that are used to evaluate the classification performance of the proposed algorithm and an instrumental measure to evaluate the audio quality of three existing impulse restoration algorithms and also a full restoration chain where only those frames that have been classified to contain impulsive disturbances are processed by an impulse restoration algorithm.

2.4.2.1 *Classification Performance*

The performance of a classification system is typically rated based on three measures: *accuracy*, *precision* and *recall* [138, 139]. The accuracy is simply the proportion of correctly identified instances:

$$\text{Accuracy} = \frac{\text{TPos} + \text{TNeg}}{\text{Pos} + \text{Neg}},$$

with TPos and TNeg the number of true positive and true negative instances, respectively—in our context this translates to *disturbance present & correctly classified as disturbed* and *no disturbance present & correctly classified as disturbance-free*, respectively. Pos and Neg are the overall number of positive (disturbed) and negative (clean) instances, respectively. In our context, an *instance* corresponds to a *frame* of the input signal, and all frames of all test signals considered in each experiment are combined to determine the values of TPos, TNeg, Pos and Neg.

If the classes (*clean* and *disturbed*) are skewed, i.e., the number of instances in each class differ, the accuracy measure may not be very useful. The most extreme example would be when all instances are disturbed. In that case, a classifier always assuming an instance to be disturbed will yield an accuracy of 100%. Obviously, such a classifier would perform very poorly in real-world scenarios as no clean instance would be classified as such.

Additional performance measures can be used that take the number of positive (disturbed) and negative (clean) instances into account. The *precision* specifies the number of disturbed instances compared to the number of instances assumed to be disturbed:

$$\text{Precision} = \frac{\text{TPos}}{\text{TPos} + \text{FPos}},$$

with FPos the number of undisturbed instances erroneously assumed to be disturbed (false alarm). The *recall* value is the proportion of disturbed instances that have been classified as disturbed:

$$\text{Recall} = \frac{\text{TPos}}{\text{Pos}}.$$

2.4.2.2 *Instrumental Measures for Audio Quality*

In order to rate the quality improvement of existing impulse restoration algorithms and also to determine the benefit of the proposed impulsive disturbance classification algorithm when integrating it with an impulse restoration algorithm, we will rate the perceived audio quality of the processed signal using an intrusive instrumental audio quality measure. In this context, "intrusive" means that the quality is determined by computing a similarity measure between the processed signal and a (clean) reference signal. More specifically, the instrumental measure used in this article is the "Perceptual Evaluation of Audio Quality" (PEAQ) measure [108, 109, 110]. It yields a so-called *Objective Difference Grade* (ODG) describing the perceptual difference to a reference signal that ranges from -4 ("very annoying") to 0 ("imperceptible"), cf., Table 2.2.[2]

Although PEAQ was originally developed to assess artifacts of audio coders, we still decided to use this measure to evaluate the performance of impulse restoration algorithms, since this measure has also been used to evaluate other audio enhancement algorithms [85] and informal listening experiments showed that the obtained ODG scores generally correspond well with subjective auditory impression (cf., demonstration signals on the website accompanying this article [111]).

2.4.3 *Reference Impulse Restoration Algorithms*

One reasonable application of the proposed impulsive disturbance classification algorithm is in combination with an impulse restoration algorithm. A straightforward way to make automatic restoration possible without compromising the quality of

---

[2] As the PEAQ algorithm requires its input signals to have a sampling rate of 48 kHz, the processed and reference signals were resampled accordingly before running the PEAQ algorithm.

Table 2.2: The ODG scale.

ODG	Impairment Description
0	Imperceptible
-1	Perceptible but not annoying
-2	Slightly annoying
-3	Annoying
-4	Very annoying

undisturbed signal portions is to only process those 1 s frames of the input signal with an impulse restoration algorithm that have been classified to contain impulsive disturbances. These processed frames can be concatenated with undisturbed, unprocessed frames. To do so, of course, possible processing delay of the restoration algorithm has to be taken into account.

We use three impulse restoration algorithms for reference. All of them are based on an AR model of the clean signal for impulse detection and interpolation [39, Ch. 5]:

- LSAR – A standard least squares AR algorithm that combines the AR model with a sinusoidal model for the input signal to increase the detection and interpolation performance [39, Ch. 5.2.3.2]. In addition, the AR model parameters and clean signal are estimated iteratively [39, Ch. 5.3.1] as informal listening tests have shown that the achieved restoration quality benefits greatly from doing so. We use this algorithm in the implementation and with parameter values from [62].
- DT-LSAR – An impulse restoration algorithm that uses an improved detection stage by using a double-threshold based approach [140]. Specifically, the algorithm is able to merge closely spaced impulses and processes each block of the input signal multiple times to reduce the number of missed disturbance impulses.
- Auto-LSAR – A recently published algorithm that incorporates ideas from [140] and is reported to achieve good restoration performance for a wide range of input material without manual parameter adjustment [141].

2.4.4 *Results*

In this section we present results of four experiments to determine the optimum prewhitening, the classification performance of the proposed algorithm with unknown signals, the restoration performance of the three reference impulse restoration algorithms with no parameter adjustment, and the perceptual audio quality improvement of a fully automatic impulsive disturbance restoration chain.

2.4.4.1 *Optimum Prewhitening*

It is expected that the prewhitening method and the prewhitening parameters (e.g., block length N, prediction order P_{LP}) have a major influence on the performance of the classification algorithm. Based on a subset of 31 clean signals (one randomly selected from each year, cf., Section 2.4.1) from the signal database, the disturbed signals were generated with SNRs ranging from 20 dB to 50 dB, using both SNR concepts. As mentioned above, those frames from the disturbed class that, due to the random nature of the disturbance signal generation, did not contain any impulses were removed from the corrupted class of the data set. The classification algorithm was trained per condition, i.e., per combination of block length N, choice of prewhitening (none, PHOT or linear prediction), classifier (logistic regression or SVM), SNR concept and, for prewhitening based on linear prediction, also prediction order P_{LP}. For each condition, $31 \cdot 20 = 620$ clean frames were used with an equal number of disturbed frames that were randomly selected from the available $31 \cdot 4 \cdot 20 = 2480$ frames. This was done in order to find an optimal prewhitening working well both at high and low SNR conditions. We did not use separate training and test data sets as the aim was to determine the specific prewhitening that allows for the best classification performance for all data; in this experiment we were not interested in the generalization performance of the classification algorithm, i.e., how accurately it classifies unknown data. In this section we will rate the classification quality solely based on the accuracy. Despite what was said about the disadvantages of the accuracy measure in Section 2.4.2.1, these results are still meaningful as we selected an equal number of clean and disturbed instances for our experiments. The fraction of disturbed frames in an actual archive restoration application scenario may differ from our assumptions, but as we were not able to find more detailed information on this topic, we think that this approach allows for an evaluation as general as possible.

Figure 2.4 shows the classification accuracy for several prewhitening algorithms, for different classifiers and for both SNR concepts. The accuracy values are averaged over all SNRs per condition. The two columns of Figure 2.4 contain the results for the two SNR concepts. The results for prewhitening based on linear prediction are those obtained with the optimum prediction order P_{LP}. The optimum prediction order was determined beforehand as that P_{LP} that allows for the highest accuracy, individually for each block length N. As can be observed the choice of classifier seems to be of minor importance, as the curves for logistic regression and SVM lie almost on top of each other. However, the choice of prewhitening has a large influence on the classification performance. Although employing no prewhitening at all allows for a classification accuracy that is above chance level, the use of linear prediction and PHOT yields a much better classification accuracy, with PHOT clearly outperforming linear prediction. Figure 2.4 shows that

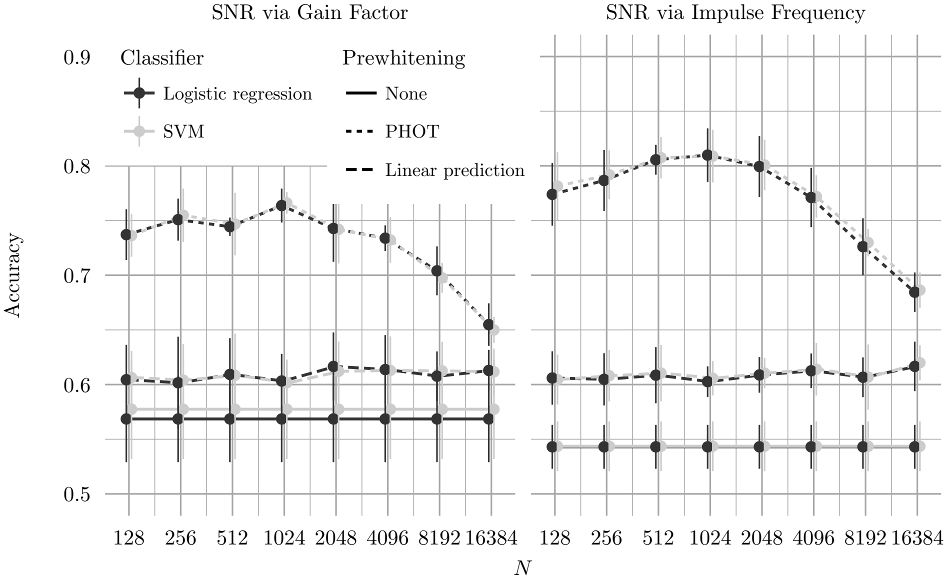


Figure 2.4: Classification accuracy for all analyzed types of prewhitening with varying block lengths. The two figures show classification results averaged over all SNRs (20, 30, 40, and 50 dB). For the linear predictor, for each block length the optimum prediction order was selected. To enhance the clarity, plots have been separated in terms of the SNR concept. The length of the error bars is twice the standard deviation of the five cross-validation runs.

the achieved overall accuracy is higher if the SNR is set by modifying the impulse density (cf., Section 2.2.2.2). This is plausible as the amplitude—which corresponds to the detectability—remains the same independent of the SNR. Although there is no clear optimum choice for all conditions, we chose the combination of PHOT prewhitening with a block length of $N = 1024$ samples and the logistic regression classifier for all further experiments.

2.4.4.2 *Classification Performance in Dependence on the SNR*

Using the optimal prewhitening parameters determined in the previous section we evaluate the classification performance of the proposed algorithm using the complete test signal database based on 620 clean recordings. As in the last experiment, the disturbed signals were obtained using artificial disturbances, SNRs of 20 dB to 50 dB using both SNR concepts. Only those frames of the disturbed class that actually contain any impulses are used for training, supplemented by an equal number of clean signal frames. To determine the generalization capabilities of the classifier, the available data was split into training and test sets, comprising 60% and 40% of the data, respectively. Classifier training and hyperparameter optimization was performed with only the training data as described in Section 2.3.3. The classifica-

Table 2.3: Classification performance in dependence on the SNR. All accuracy, precision and recall values are in percent.

	SNR Concept					
	Gain			Impulse Density		
SNR	Accuracy	Precision	Recall	Accuracy	Precision	Recall
20 dB	93.2	88.7	99.1	92.2	88.4	97.1
30 dB	85.9	86.9	84.4	86	87	84.6
40 dB	66	77.9	44.7	68.2	75.5	46.3
50 dB	53.1	59.9	19	76.1	33.8	27.8

tion performance was then evaluated based only on the test data. Table 2.3 lists the classification error measures in dependence on the SNR and the SNR concept.

As expected, the classification performance improves as the SNR decreases; at an SNR of 20 dB approximately 92% of all frames are classified correctly ("Accuracy" columns in Table 2.3). At an SNR of 40 dB the accuracy decreases to ≈ 67%, however note the precision and recall values: The recall value drops to ≈ 45% for both SNR concepts whereas the precision value indicates that ≈ 76% − 78% of the frames that have been classified to be disturbed actually contain impulses. The relatively low recall values in high SNR scenarios will in many cases not pose a severe problem as the disturbance is inaudible anyway (compare the demonstration signals on the article website). Furthermore, the classification performance is similar for both SNR concepts except at 50 dB.

2.4.4.3 *Impulse Reduction Performance of Existing Restoration Algorithms*

The goal of this section is to determine a baseline for the performance of existing impulse restoration algorithms when used with very diverse audio material in an automatic manner, i.e., with no parameter adjustment. Therefore, we processed our test signal database (cf., Section 2.4.1) with the LSAR, DT-LSAR and Auto-LSAR algorithms (cf., Section 2.4.3) and rated the restoration capabilities in terms of the perceptual quality of the restored signal. As described in Section 2.4.2.2 the perceived quality is determined using an instrumental measure, namely the PEAQ algorithm. This algorithm compares two signals, the *reference* and the *test* signal, and computes a single number, indicating the perceptual similarity of both. The results were obtained using the clean signal for reference.

The box plots [142] in Figure 2.5 display the distribution of the ODG scores obtained using PEAQ: The lower and upper edges of each box correspond to the first and third

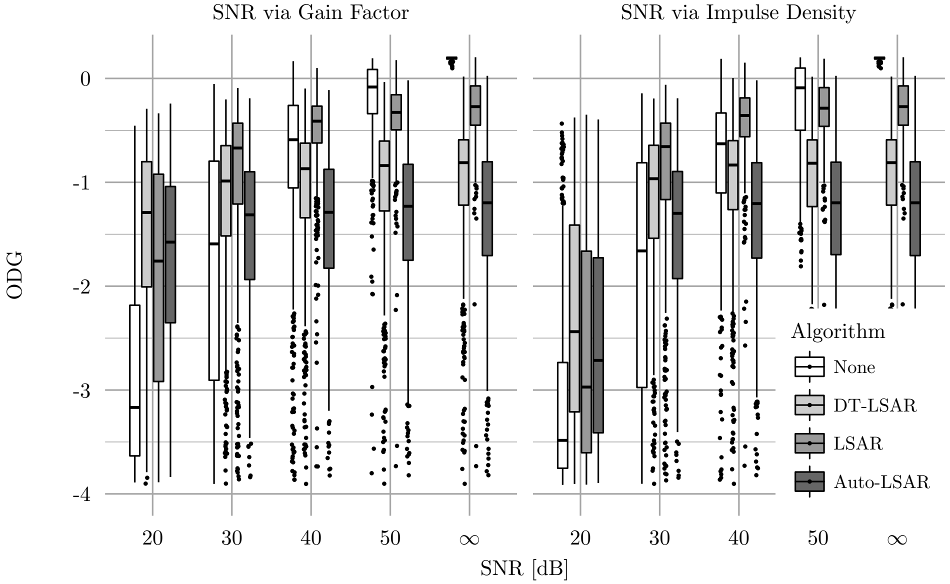


Figure 2.5: Results of the instrumental audio quality evaluation of the three impulse restoration algorithms described in Section 2.4.3. The ODG scores were obtained with the PEAQ algorithm, and for each SNR concept ("Gain" or "Impulse Density") and SNR all 620 signals from the test signal database were used. The reference signal for the PEAQ algorithm is the clean signal in all cases. The leftmost boxes ("None") represent the results for the disturbed signal that has not been processed by a restoration algorithm, the other boxes ("LSAR," "DT-LSAR," and "Auto-LSAR") represent the results obtained when the complete signal is processed by the respective restoration algorithm. Refer to Section 2.4.4.3 for the interpretation of this box plot.

quartiles of the data, respectively. Consequently, the height of each box represents the inter-quartile range (IQR). The horizontal line inside each box represents the median value and the lines extending vertically from each box indicate the smallest and largest data point, respectively, that is still within $1.5 \cdot \text{IQR}$ distance from the lower, or upper, edge of the box, respectively. All data outside of this interval are considered outliers and represented by dots.

As can be seen in the figure, considering the leftmost group ("None") which represents the results for the unrestored, disturbed input signals, the perceptual audio quality is severely impaired by impulsive disturbances at low SNRs (compare Table 2.2). For SNRs above 40 dB the ODG attains median scores around zero, indicating mostly unnoticeable signal quality degradation. For most SNRs, a number of outliers extend to low ODG scores around -4. Informal listening tests have shown that these results correspond to test signals whose desired signal exhibits certain peculiarities. For example, low ODG scores for unprocessed signals at SNRs of 40 dB and 50 dB are caused by signals that have very low high-frequency content or that contain very quiet sections. In both cases, even soft impulses are perceptually striking, resulting in low ODG scores.

Higher ODG scores for the signals processed by the impulse restoration algorithms ("LSAR," "DT-LSAR," and "Auto-LSAR") compared to the disturbed signals: ("None") for SNR values of 20 dB and 30 dB indicate that impulse restoration processing leads to an improvement of audio quality for heavily disturbed signals. For severe degradations at an SNR of 20 dB and especially for the SNR concept "Gain" the "DT-LSAR" algorithm yields a severe increase in audio quality, outperforming the other two algorithms. This is likely to be caused by its improved detection stage featuring less missed detections (compare [140]). The "LSAR" algorithm yields less quality improvement for very low SNRs, but is able to increase the audio quality up to an SNR of 40 dB. However, note that for SNRs above 40 dB the uninformed processing with any of the evaluated impulse restoration algorithms leads to a median *decrease* in quality, compared to the unprocessed input signal. This is especially evident in the results for undisturbed signals (represented here with an SNR of ∞ dB). This observation suggests that in these cases all three impulse restoration algorithms produce a high number of erroneous detections, with the consequence of removing parts of the desired signal.

2.4.4.4 *Restoration Performance with the Classification Algorithm*

The last experiment evaluates the gain in audio quality that can be obtained when combining the presented impulsive disturbance classification algorithm with the LSAR impulse restoration algorithm. We decided to use the LSAR algorithm for this experiment as the results in Figure 2.5 indicate that the LSAR algorithm, of all three analyzed impulse restoration algorithms, performs best when used with a wide variety of input signals. The improvement in perceived audio quality, as in the last section, is determined via the PEAQ algorithm, using the clean signal for reference. Figure 2.6 shows three groups of data, subdivided by the type of impulse restoration processing: "None," "Classified," and "All". The first group, "None", represents the ODG scores for the disturbed, unprocessed input signal. The "All" group displays the ODG scores for the signals with all frames processed with the LSAR impulse restoration algorithm and corresponds to the "LSAR" boxes in Figure 2.5.[3] The "Classified" group represents the results obtained for signals where only frames indicated by the classifier to actually contain impulsive disturbances were processed by the restoration algorithm. The "Classified" group in the figure reveals that for SNR values of $\geq$40 dB the ODG benefits from the application of the impulsive disturbance classification algorithm, saving (mostly) clean signal frames from being distorted by the impulse restoration algorithm. The ODG scores in these cases are significantly higher than the scores of the fully processed signals, becoming more evident with increasing SNR values and yielding the largest gains with clean signals. For low SNR values, the application of the impulsive disturbance classification algorithm has practically no drawbacks as the classification accuracy

---

[3] However, note that in this section, only the 248 signals from the test set were used for the evaluation, while all 620 test signals were used in the last section.

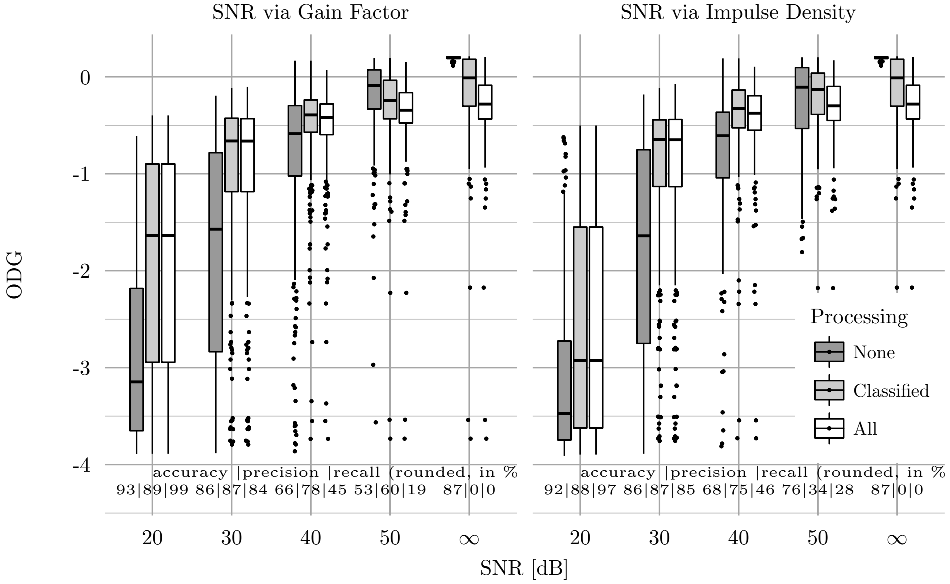


Figure 2.6: Results of the instrumental audio quality evaluation of the full restoration processing chain. The ODG scores were obtained with the PEAQ algorithm, and for each SNR concept ("Gain" or "Impulse Density"), SNR and type of processing only the 248 signals from the test set, previously unknown to the classification algorithm, were used. The reference signal for the PEAQ algorithm is the clean signal in all cases. The leftmost boxes (processing "None") represent the results for the disturbed signal which has not been processed by the restoration algorithm, the rightmost box ("All") represent the results obtained when all frames of the signal are processed by the restoration algorithm. The middle boxes ("Classified") show the results with the classification algorithm applied: only the frames classified to contain impulsive disturbances are processed by the restoration algorithm. The tables on the bottom of the figure copy the classification performance measures from Table 2.3 for convenience.

is very high in these cases—compare the classification performance measures in the bottom of the plot. Hence, for SNRs of 20 dB and 30 dB almost all frames are correctly classified to contain impulsive disturbances, yielding almost identical results to the fully processed signals.

In summary, we find that for signals that only contain marginal disturbances or that are completely clean, the presented impulsive disturbance classification algorithm shows its main improvement: Prevent clean signals from being processed unnecessarily and avoid a reduction of audio quality.

2.5 Conclusions

In this article we presented a novel classification algorithm to automatically determine whether an audio recording contains impulsive disturbances or not. The proposed algorithm is based on a supervised learning approach. Using a large clean music database and artificially generated but plausible disturbances we could show that the algorithm is capable of classifying most audible disturbances correctly while featuring a small false alarm rate. Compared to existing impulse detection schemes, which exhibit a time resolution in the order of the sampling interval, our approach yields classification results for input signal frames of 1 s duration. Hence, it is able to take advantage from the additional information, however at the cost of a decreased time resolution. Furthermore, our results show that prewhitening the input signal by means of the phase only transform is an important step to increase the detectability of disturbance impulses which can also be used as a detection enhancement method for impulse restoration algorithms.

Based on an instrumental audio quality measure, we have presented evaluation results that suggest that well-known, AR model based impulse restoration algorithms suffer from a significant number of false alarms, especially for high input SNRs. Thus, it is important to determine whether a restoration is actually required. The developed classification algorithm can be used in conjunction with legacy impulse restoration algorithms to reduce the number of erroneous detection results and, as a consequence, to increase the audio quality of the restored signal.

We conclude that the presented method constitutes a crucial step towards fully automatic restoration of media archives.

The website accompanying the article [111] makes a number of disturbed and restored signals available for listening, including their ODG scores.

2.A Appendix

2.A.1 *Shape Parameter Values of the Gamma Distribution*

The mean shape parameters of the gamma distribution, $\bar{\Theta}$, that are required to obtain specific SNR values are given in Table 2.4, including the standard deviation σ_Θ over all of the test signals. The standard deviation is zero for an SNR of 30 dB because this is the default case and the standard parameters for the impulsive disturbance generator are used for all signals. For all other SNRs, the shape parameter depends on the individual clean input signals.

Table 2.4: Mean and standard deviation of the shape parameter of the gamma distribution for different SNRs.

SNR	$\bar{\Theta}$	σ_Θ	
20 dB	230	23	
30 dB	2 434	0	(default)
40 dB	24 083	6 374	
50 dB	227 976	99 481	

2.A.2 *List of Features*

Table 2.5 lists all statistical measures that have been investigated as features of the prewhitened and normalized signal $x'^{(q)}_{\text{pre}}[n]$. As described in Section 2.3.2, recursive feature elimination was used to determine a set of two features that provide good classification results while reducing the computational requirements.

For the computation of the crest factor with trimmed mean, $\text{Trimmean}_{i\%}\{\cdot\}$ is the $i\%$ trimmed mean as described in [143, Ch. 3.3].

Table 2.5: Features of the prewhitened signal that have been investigated.

Feature	Computation
Crest factor	See (2.6)
Crest factor – $l\%$ trimmed mean, $l \in \{1,\ 2,\ 5,\ 10\}$ (see [143, Ch. 3.3])	$C_{l\%}^{(q)} = \frac{\max\limits_{0 \leq i < M} \left\lvert x_{\mathrm{pre}}^{\prime(q)}[i] \right\rvert}{\sqrt{\mathrm{Trimmean}_{l\%}\left\{ \left(x_{\mathrm{pre}}^{\prime(q)}[n] \right)^2 \right\}}}$
Peak-to-Root-Median-Squared ratio	$\mathrm{PRMedS}^{(q)} = \frac{\max\limits_{0 \leq i < M} \left\lvert x_{\mathrm{pre}}^{\prime(q)}[i] \right\rvert}{\sqrt{\mathrm{Med}\left\{ \left(x_{\mathrm{pre}}^{\prime(q)}[n] \right)^2 \right\}}}$
Kurtosis	See (2.7)
Kurtosis of absolute value	$\mathrm{Kurt}_{\mathrm{abs}}^{(q)} = \frac{\frac{1}{M} \sum_{i=0}^{M-1} \left(\left\lvert x_{\mathrm{pre}}^{\prime(q)}[i] \right\rvert - \overline{\left\lvert x_{\mathrm{pre}}^{\prime(q)} \right\rvert} \right)^4}{\left(\frac{1}{M} \sum_{i=0}^{M-1} \left(\left\lvert x_{\mathrm{pre}}^{\prime(q)}[i] \right\rvert - \overline{\left\lvert x_{\mathrm{pre}}^{\prime(q)} \right\rvert} \right)^2 \right)^2}$
Skewness	$\mathrm{Skew}^{(q)} = \frac{\frac{1}{M} \sum_{i=0}^{M-1} \left(x_{\mathrm{pre}}^{\prime(q)}[i] - \overline{x_{\mathrm{pre}}^{\prime(q)}} \right)^3}{\left(\frac{1}{M} \sum_{i=0}^{M-1} \left(x_{\mathrm{pre}}^{\prime(q)}[i] - \overline{x_{\mathrm{pre}}^{\prime(q)}} \right)^2 \right)^{3/2}}$
Sparseness (see [144])	$\mathrm{Sparseness}^{(q)} = \frac{\sqrt{M} - \sum_{i=0}^{M-1} \left\lvert x_{\mathrm{pre}}^{\prime(q)}[i] \right\rvert / \sqrt{\sum_{i=0}^{M-1} \left(x_{\mathrm{pre}}^{\prime(q)}[i] \right)^2}}{\sqrt{M} - 1}$

3

DETECTION OF HUM DISTURBANCES

This article examines the automatic detection of low frequency additive sinusoidal disturbances in audio signals, usually termed *hum*. We present a method to automatically determine whether an audio signal contains hum or not, and, if necessary, to determine its parameters – e.g., the fundamental frequency and the number of harmonics. The developed algorithm does not require a priori information, and we show its good detection capabilities by an evaluation with artificial signals and real recordings.

3.1 Introduction

Since the end of the 19th century, a large number of audio recordings have been produced. Recording media vary from wax cylinders, shellac and vinyl discs, to various tapes and photographic films to modern digital data storage. The types of disturbances that degrade recordings are manifold, and there are digital signal processing algorithms to reduce the audible artifacts, e.g., [39]. One very common disturbance is hum. In contrast to clicks and crackles, hum disturbances can even be found in modern (e.g., live) recordings and they cause several problems. Obviously, the additive tonal signal can distract the listener. Furthermore, the strong tonal components at low frequencies can cause amplifiers to unnecessarily drain huge amounts of power or even overload loudspeakers. These low-frequency tones can also affect dynamic processing devices such as compressors or noise gates since most of these units rely on the broadband power of the input signal. The undesirable low-frequency power might obscure the “true” dynamic of the content.

This chapter contains a copy of the article

M. Brandt, J. Bitzer, “Automatic Detection of Hum in Audio Signals,” *Journal of the Audio Engineering Society*, vol. 62, no. 9, pp. 584–595 (2014 Oct.). `https://doi.org/10.17743/jaes.2014.0034`

While the layout of the article has been adapted for a uniform presentation within this thesis, the contents printed here are identical to those in the published article.

Some specific algorithms to remove hum are commercially available. However, these methods need manual adjustment. In this article we present a method to determine the required parameters to enhance a signal without depending on user interaction. Yet we will not be dealing with the removal of hum here but focus on that topic in a subsequent article. A first overview of removal algorithms was already published in [115], though. In order to control removal algorithms we need to know the fundamental frequency of the hum and how many harmonics are present. These parameters have to be estimated with the desired music signal as a non-stationary disturbance that masks the hum signal.

In recent years some articles on related subjects have been published, although no article could be found that deals with the specific complex problems of detecting hum in music signals. For example, Grigoras [145] shows a method to estimate the fundamental frequency in the context of forensic audio and how this electric network frequency and its unique variations over time can be used to determine when a recording was made and whether pieces have been cut out. Czyzewski et al. [146] give a thorough description of a method to track the fundamental frequency of a hum disturbance with high accuracy to use this information to remove disturbances known as *wow* and *flutter* in older recordings. These methods do not focus on the harmonics or the removal of hum. Liu and Chen [147] present a method to estimate the amplitudes and phases of harmonics of power systems when the fundamental frequency is known a priori. Unfortunately, this and other well-known approaches to the related topic of fundamental frequency estimation in speech or music signals (e.g., [87, 148]) do not represent a solution to our problem as they address tracking the non-stationary foreground desired signal at positive signal-to-noise-ratios (SNRs). In contrast, we aim at estimating the fundamental frequency and harmonics of a very stable background disturbance with a strongly fluctuating music signal in the foreground.

In this paper we present an extended and refined version of our hum detection algorithm [149]. Furthermore, this article shows the algorithm in much more detail and also contains a deeper analysis of parameters and evaluation results. The paper is organized as follows: In Section 3.2 the disturbance, its cause, and its parameters are explained in more detail and the applied signal model is introduced. Based on this signal model we present an overview of the detection algorithm in Section 3.3. The algorithm is tested with artificially disturbed and real recordings, and the evaluation method and results are described in Section 3.4. Finally, we end the paper with some conclusions.

3.2 Problem Statement

During recording and/or copying processes of audio material, power line interference (PLI) can cause the addition of one or more stationary sinusoidal tones to the desired signal. This is often the case if audio signal lines are placed close to power cables and with poorly designed or faulty electric circuits. Usually, the disturbing signal is a harmonic tone complex with a fundamental frequency of approximately 50 Hz or

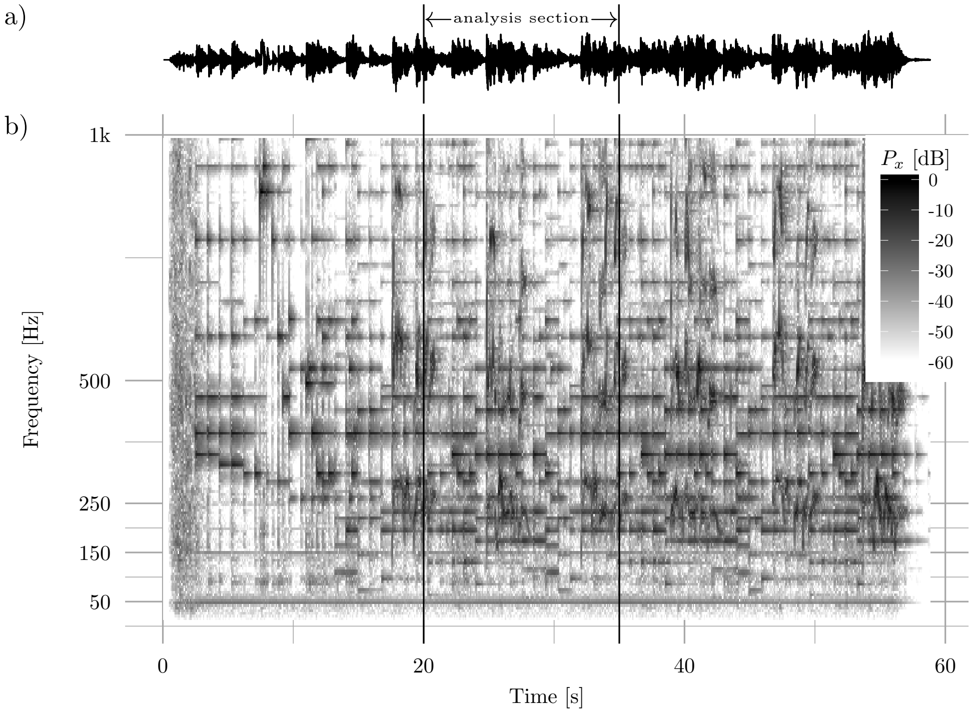


Figure 3.1: Time signal (a) and spectrogram (b) of a music recording containing a hum disturbance. In this example, showing an excerpt of a BBC recording of Elton John's "Sorry Seems to Be the Hardest Word" from 1976, the hum frequencies are $\mathbf{f}_{\mathrm{hum}} \approx [50\,\mathrm{Hz},\ 150\,\mathrm{Hz}]$. An exemplary section of 15 s, starting at 20 s and ending at 35 s, is marked to illustrate the length of the moving window that is used for analysis.

60 Hz – depending on the power line frequency used. Non-linearities in the signal chain in many cases result in a number of additional harmonics. We model the disturbed signal as

$$x(t) = s(t) + \sum_{a=0}^{N_{\mathrm{tc}}-1} n_a(t), \tag{3.1}$$

where $x(t)$ is the disturbed recorded signal, $s(t)$ is the unobservable clean signal, N_{tc} is the number of tone complexes, and

$$n_a(t) = i(t) \sum_{b=1}^{N_{\mathrm{p,a}}} A_b \sin([b\,\omega_{0,\mathrm{a}}]t) \tag{3.2}$$

with

$i(t)$ – a switch variable, indicating whether the disturbance is present at time t or not ($i(t) \in \{0, 1\}$)
N_{p} – the number of partial tones
A_b – the amplitude of each partial tone
ω_0 – the fundamental frequency

are harmonic tone complexes forming the disturbance signal. Fig. 3.1 shows the spectrogram of an exemplary live music recording that contains a hum disturbance. In this case, the hum frequencies are approximately 50 Hz and 150 Hz.

In some cases, the overall disturbance may consist of several tone complexes ($N_{\mathrm{tc}} > 1$) with different fundamentals. This can be the case, for example, when a recording was made in the USA (on 60 Hz-powered machines) and copied in Europe (using a power line frequency of 50 Hz). Fig. 3.2 shows power spectral densities of two typical hum disturbances: Fig. 3.2 a) depicts the power spectral density of a hum tone complex with a fundamental frequency of 60 Hz and a number of harmonics at 120 Hz, 180 Hz, 240 Hz, etc. In this example the power of the fundamental frequency is lower than the power of the second harmonic, and the power of the harmonics decreases with the frequency. Furthermore, the power of the even-numbered harmonics is lower than the power of the odd-numbered harmonics. Fig. 3.2 b) shows a phenomenon that may occur in practice: There are two hum tone complexes with different fundamental frequencies ω_0 (50 Hz and ≈ 58 Hz) and numbers of partial tones N_{p}.

Furthermore, irregular mechanical motion of media reproduction equipment can lead to variations of the fundamental frequency during recording or playback (called wow and flutter [150, 151, 152]). They are caused, for example, by an imprecisely centered spindle hole on a record disc or by an inconstant angular velocity of a capstan of a tape machine. In other cases, low running batteries of mobile recording equipment may result in a monotonic change (increasing or decreasing, respectively, depending on whether the mobile machine is used for recording or for playback) of the fundamental frequency of the hum's tone complex. However, these cases are rare and are therefore not considered here. The algorithm though should be capable of dealing with small fluctuations of the hum frequencies.

Since the number of audible partial tones is low, the range of the detection is restricted to low frequencies. This fact distinguishes hum from buzz, as the latter typically contains a high number of harmonics, resulting in a large frequency range that is disturbed. The algorithm should be capable of detecting the individual tones of a hum disturbance, not only the fundamental frequency, i.e., determining the parameters ω, N_{tc}, A_b and $N_{\mathrm{p,a}}$ in Eq. (3.2). This is especially important as in many cases the fundamental frequency is missing, e.g., due to highpass filtering during copying processes. Moreover, many of the existing hum removal algorithms

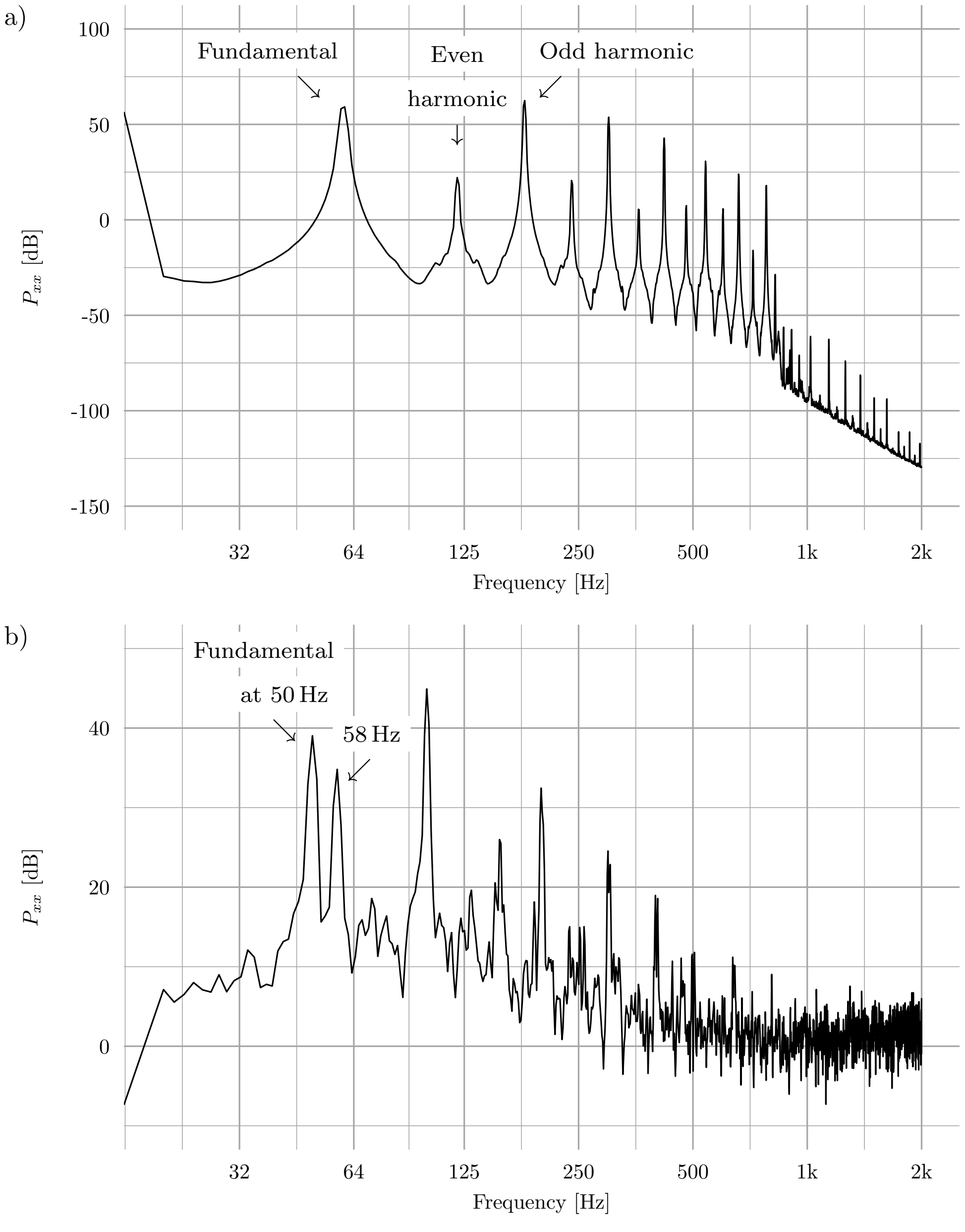


Figure 3.2: Power spectral density of hum disturbances. a) Hum tone complex consists of the fundamental frequency at ≈ 60 Hz and a number of (odd and even) harmonics. b) Two harmonic tone complexes with a fundamental frequency of 50 Hz and a tone complex with a fundamental frequency of approximately 58 Hz.

require a determination of whether the fundamental frequency is present or not. If no fundamental frequency is found it should be estimated, based on the detection of one or more harmonics. Furthermore, the algorithm must be able to determine start and end times of hum tones – this corresponds to estimating $i(t)$ in Eq. (3.2) – to avoid degradation of the desired clean signal by applying hum removal algorithms to undisturbed sections of the audio material. The most negligible parameter is the individual amplitude, since most removal algorithms are based on filtering out the affected frequency completely.

3.3 The Detection Algorithm

The detection algorithm utilizes the presumed long-term stability of the hum. This stability, in terms of power and frequency, is reflected in the signal model in Eq. (3.2) since A_b and ω_0 do not change over time. To detect stable tones, the basis of the detector is to analyze rather long sections of the input signal ($\approx 10\,\mathrm{s} \ldots 30\,\mathrm{s}$). The basic principle of the detection algorithm is a statistic analysis of the shortterm discrete Fourier transform (STDFT) of the input signal that is computed blockwise. In order to reduce the computational complexity of the processing, the input signal is sampled down to a sampling rate of $f_\mathrm{s} = 2\,\mathrm{kHz}$ in a pre-processing step that reduces the bandwidth to 1 kHz. To avoid aliasing effects we applied a 100^th order linear-phase FIR low-pass filter beforehand with a cut-off frequency of 900 Hz. This bandwidth reduction does not pose a problem since most hum tone complexes have decayed up to 1 kHz. For the block-based processing, the short-term power spectral densities of the input signal are computed by first obtaining overlapping blocks of length L_b from the input signal. Block p is defined as[1]

$$x^{(p)}[k] = x[p \cdot L_\mathrm{f} + k], \quad k = 0, \ldots, L_\mathrm{b} - 1, \tag{3.3}$$

where p and k are block and sample indices, respectively, and L_f is the length of the block-feed. With the discrete Fourier transform (DFT) of each signal block

$$X^{(p)}[n] = \sum_{k=0}^{L_\mathrm{b}-1} x^{(p)}[k] \cdot \mathrm{e}^{-2\pi j \frac{k \cdot n}{L_\mathrm{b}}} \tag{3.4}$$

a block-wise power spectral density (PSD) estimate is obtained by first-order recursive smoothing of the blocks' periodograms (see, e.g., [153, p. 147]):

$$\hat{P}_{xx}^{(p)}[n] = \alpha \cdot \hat{P}_{xx}^{(p-1)}[n] + (1 - \alpha) \cdot \left| X^{(p)}[n] \right|^2 . \tag{3.5}$$

In Eq. (3.4) and Eq. (3.5), n is the frequency index of the DFT. In Eq. (3.5), $0 \le \alpha < 1$ is a factor that adjusts the amount of averaging of the recursive smoothing.

---

[1] We denote a vector, be it in time or frequency domain, associated with a certain block p by $x^{(p)}[k]$ or $X^{(p)}[n]$, respectively.

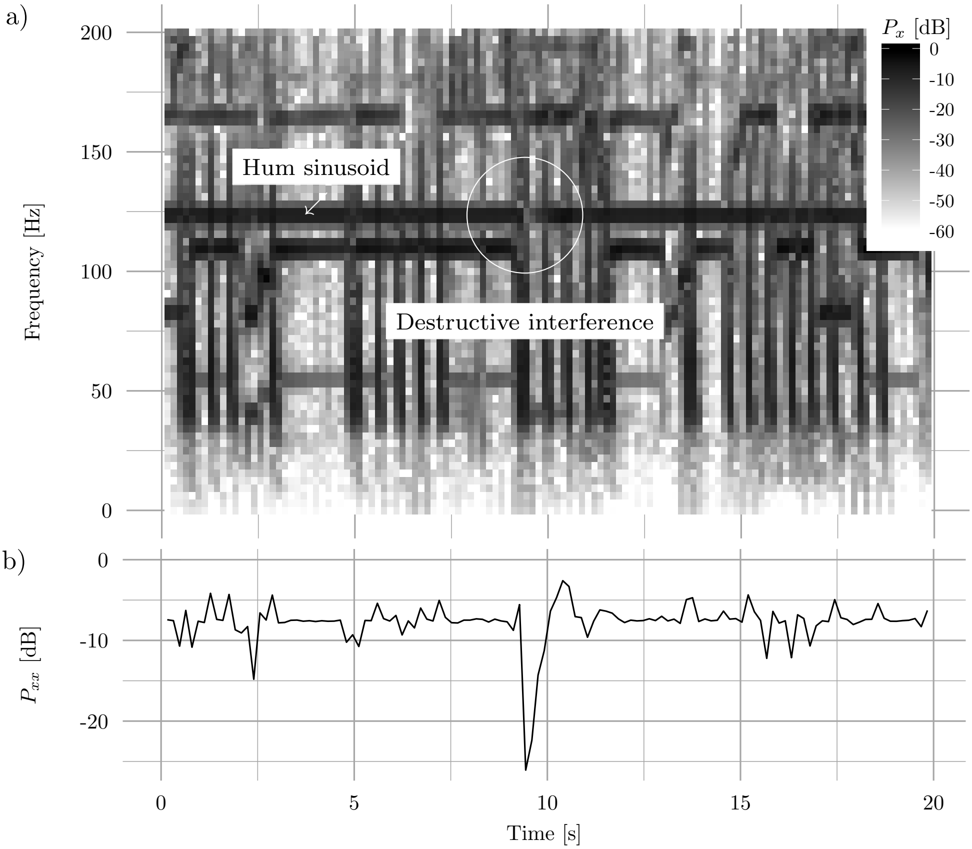


Figure 3.3: Temporary decrease of the hum sinusoid power by destructive interference with elements of the clean signal. a) Part of a spectrogram with a hum sinusoid at $f \approx 123.5\,\mathrm{Hz}$. b) Power at this frequency over time. At T$\approx 9.2\,\mathrm{s}$, the hum power drops by $\approx 20\,\mathrm{dB}$, as a bass note plays at almost exactly the same frequency, which corresponds to the musical note B2.

In order to detect hum sinusoids, the statistics of each frequency band of the spectral density are analyzed independently.

3.3.1 *Stage 1: Tone Detector*

Although in Sec. 3.2 we stated that the power of hum disturbances can be expected to be stable over a longer period of time whereas the audio signal is assumed to contain pauses in individual frequency bands, the intuitive examination of spectral minima (compare for example [91]) yields only suboptimal results. The hum power may occasionally drop well below its mean value, mainly caused by destructive interference with elements of the audio signal (e.g., bass tones). An example is

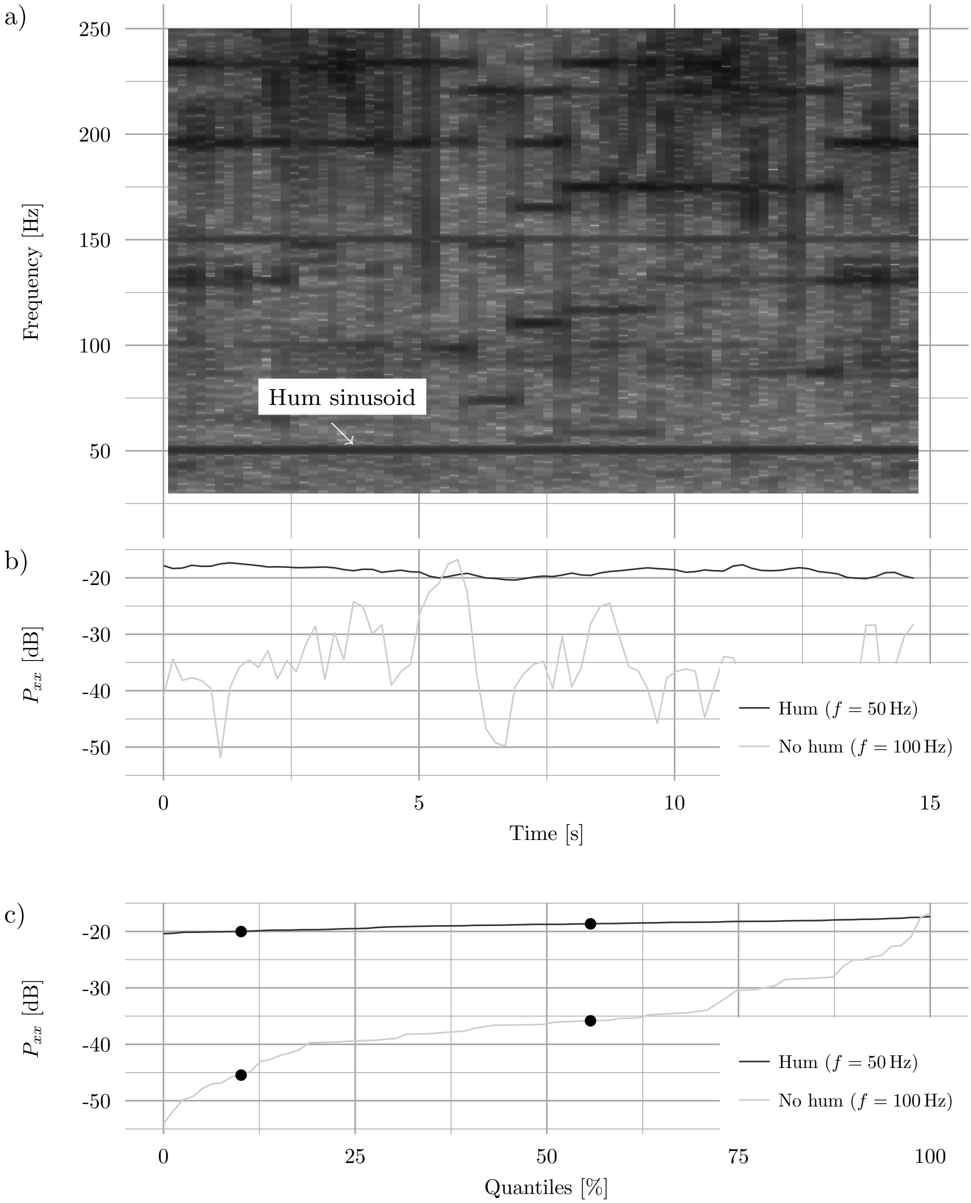


Figure 3.4: Spectrogram and quantiles that are used for hum detection. Plot a) contains a region of the spectrogram of a music signal shown in Fig. 3.1. Diagrams b) and c) show the power at frequency $f = 50\,\text{Hz}$ (solid line) and $f = 100\,\text{Hz}$ (dashed line), respectively. The power at the frequency that contains a hum sinusoid (50 Hz) is stable over the analysis interval of 15 s whereas at 100 Hz – where no hum sinusoid is present – the power is subject to strong fluctuations. Sub-figure c) is obtained from b) by sorting the power values of each frequency individually in an ascending order. The x-axis then inherently indicates the quantiles of the distribution of the PSD values.

given in Fig. 3.3, where a bass note interferes with the hum signal only at one point in time. Hence, to gain robustness, quantiles[2] are used in combination with different kinds of data smoothing.

In order to draw conclusions about the stability of the frequency components of the input signal, the short-term block PSDs representing the last T seconds[3], divided into N_b blocks, are considered:

$$\mathbf{b}_\mathrm{in}^{(p)}[n] = \Big[\hat{P}_{xx}^{(p)}[n] \quad \hat{P}_{xx}^{(p-1)}[n] \quad \cdots \quad \hat{P}_{xx}^{(p-N_\mathrm{b}+1)}[n]\Big]. \tag{3.6}$$

Fig. 3.4 a) shows a spectrogram of the input signal, consisting of the short-term block PSDs $\hat{P}_{xx}[n]$. Plot b) in the same figure shows the power of two frequencies over time, each corresponding to one row in Fig. 3.4 a). A hum frequency indication measure is created by relating the 10 % quantiles (denoted by $Q_{0.1}\{\bullet\}$) to the 55 % quantiles (denoted by $Q_{0.55}\{\bullet\}$) of $\mathbf{b}_\mathrm{in}^{(p)}[n]$. Fig. 3.4 c) shows how the quantiles $Q_{0.0} \ldots Q_{1.0}$ are obtained by sorting the values in Fig. 3.4 b) in an as ascending order. This *quantile ratio*

$$r_\mathrm{q}^{(p)}[n] = \frac{Q_{0.1}\Big\{\mathbf{b}_\mathrm{in}^{(p)}[n]\Big\}}{Q_{0.55}\Big\{\mathbf{b}_\mathrm{in}^{(p)}[n]\Big\}} \tag{3.7}$$

is a measure for the amount of fluctuation of power over time in each DFT frequency bin within the last T seconds. Fig. 3.5 a) shows r_q for the section indicated in the exemplary input signal in Fig. 3.1. Both hum tones, at $\approx 50\,\mathrm{Hz}$ and $\approx 150\,\mathrm{Hz}$, become clearly visible.

In order to reduce the influence of broad frequency ranges showing low fluctuation of power and strengthening the influence of narrowband frequency peaks, the quantile ratio measure is enhanced by subtracting its median filtered [154] version:

$$\tilde{r}_\mathrm{q}^{(p)}[n] = r_\mathrm{q}^{(p)}[n] - \mathrm{medfilt}_{\lfloor 30/f_\Delta \rfloor}\Big\{r_\mathrm{q}^{(p)}[n]\Big\}, \tag{3.8}$$

where $\mathrm{medfilt}_N\{\bullet\}$ denotes the operation of median filtering with a length of N samples and $f_\Delta = f_\mathrm{s}/L_\mathrm{b}$ is the DFT frequency resolution. $\lfloor \bullet \rfloor$ is the integer (or *floor*) operator. For the given example Fig. 3.5 a) shows $r_\mathrm{q}^{(p)}[n]$ for all frequency indices and the corresponding filtered version and Fig. 3.5 b) the result of the subtraction.

---

[2] Quantiles are characteristic values of the cumulative density function of a random variable. Their computation is straightforward: The a % quantile is obtained by finding the value that has a % of the data smaller or equal to it (compare [126, p. 68f.]).

[3] We chose $T = 15\,\mathrm{s}$ to achieve satisfactory results in our experiments.

Furthermore, we increase the detection performance by utilizing the long-term stability of hum disturbances and smoothing the time progression of $r_{\mathrm{q}}^{'}[n]$ by a median filter. Therefore, the median of the quantile ratios representing the last N_s blocks,

$$\tilde{\mathbf{r}}_{\mathrm{q}}^{(p)}[n] = \left[\tilde{r}_{\mathrm{q}}^{(p)}[n] \quad \tilde{r}_{\mathrm{q}}^{(p-1)}[n] \quad \dots \quad \tilde{r}_{\mathrm{q}}^{(p-N_{\mathrm{s}}+1)}[n]\right], \tag{3.9}$$

is computed:

$$\tilde{r}_{\mathrm{q}}^{'(p)}[n] = \mathrm{Med}\left\{\tilde{\mathbf{r}}_{\mathrm{q}}^{(p)}[n]\right\}. \tag{3.10}$$

In the last equation, Med{•} denotes the median operator. Fig. 3.5 c) depicts the time progression for all frequencies and Fig. 3.5 d) shows the resulting output of the median operator.

Stable sinusoids can be detected from $\tilde{r}_{\mathrm{q}}^{'(p)}[n]$ by applying a broadband, fixed threshold Θ. Values of $\tilde{r}_{\mathrm{q}}^{'(p)}[n]$ exceeding the threshold Θ represent frequencies that are very stable in power and can be said to contain hum tones. Finally, we obtain the set[4] of the detected hum frequencies

$$\mathcal{F}_{\mathrm{hum}} = \left\{n \cdot f_{\Delta} \middle| \tilde{r}_{\mathrm{q}}^{'(p)}[n] > \Theta\right\}. \tag{3.11}$$

3.3.2 *Stage 2: The Selection Step*

The output of the steady tone detector is then optimized by a selection stage to reduce the set of hum candidates and to raise the detection accuracy. In this step, further requirements concerning the detected sinusoids can be defined to reduce the false detection rate. Furthermore, stable sinusoids that drop out for short periods of time should still be identified by allowing certain pause durations. The most important parameter is the minimum tone duration T_{min}. Of course, choosing a rather high T_{min}, e.g., $T_{\mathrm{min}} = 30\,\mathrm{s}$, drastically reduces the false alarm rate but, on the other hand, raises the probability of missing hum sinusoids of short duration.

3.3.3 *Refinement of the Detected Frequencies*

Since the steady tone detector is based on the STDFT of the input signal, the frequency resolution is limited to f_{Δ}. Therefore, for each detected frequency a time-domain refinement method is applied. The input signal $x[k]$ is filtered with

---

[4] We denote sets by calligraphic letters $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, etc. .

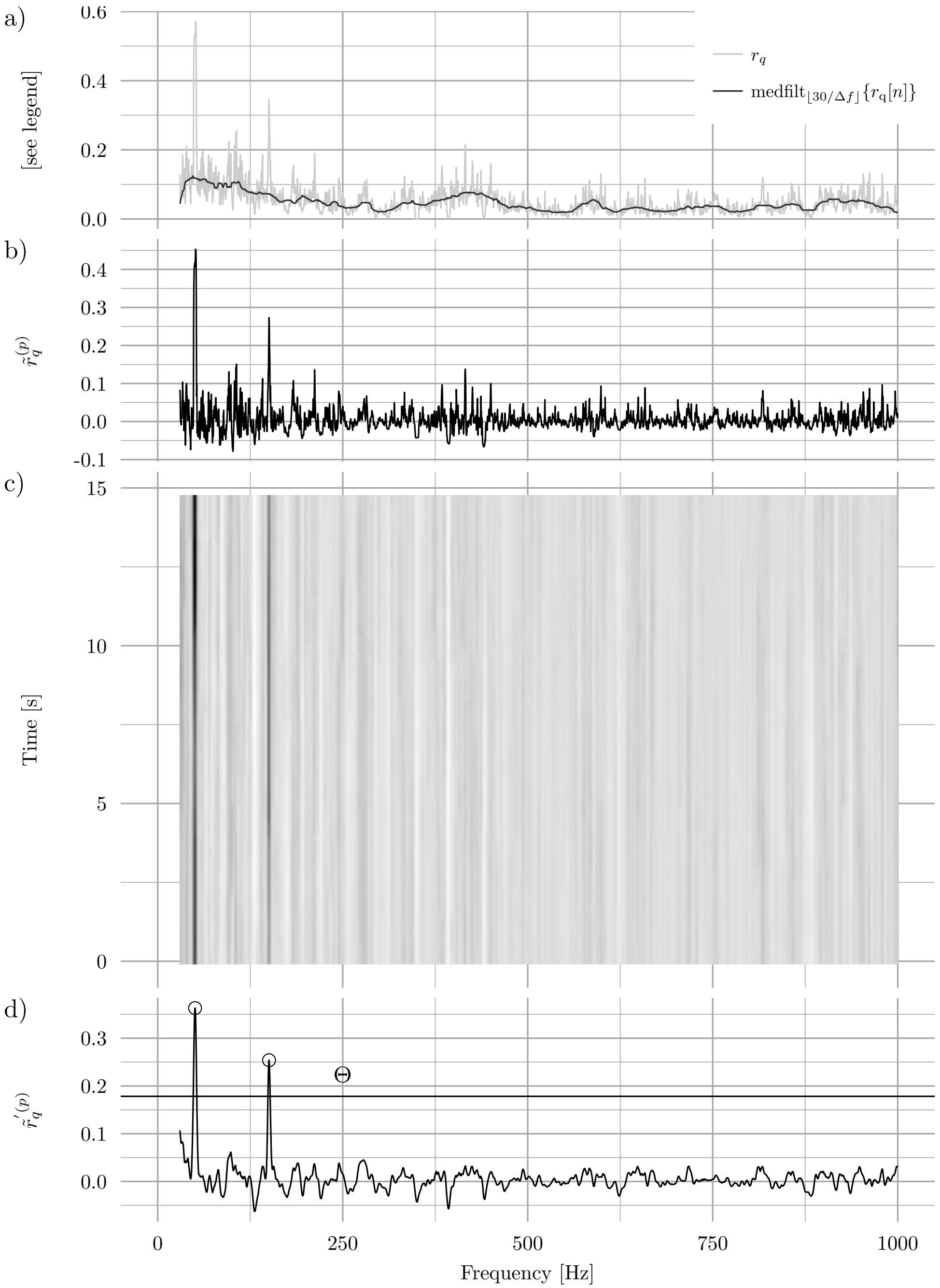


Figure 3.5: The tone detector stage. a) Example vector of the quantile ratio $r_\mathrm{q}^{(p)}[n]$ and the median filtered version of it. b) Result of the subtraction. c) Example of the matrix $\tilde{\mathbf{r}}_\mathrm{q}^{(p)}$. d) Resulting $\tilde{r}_\mathrm{q}^{'(p)}[n]$, threshold Θ and the set of detected hum frequencies, $\mathcal{F}_\mathrm{hum}$ (circles).

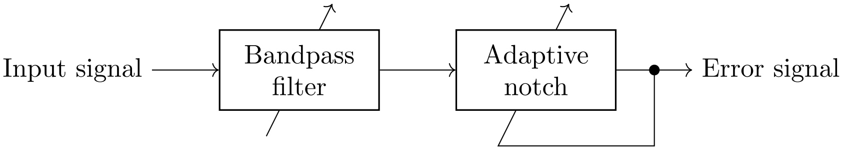


Figure 3.6: Adaptive notch filter with bandpass constrained input signal.

a bandpass filter with a bandwidth of $B_{\mathrm{BP}} = 2\,\mathrm{Hz}$, centered around the frequency that has been estimated. Finally, an adaptive notch filtering algorithm [155] with a variable center frequency but fixed bandwidth determines the desired value by converging to the frequency that leads to the smallest output power (compare Fig. 3.6).

3.3.4 *Determining the Fundamental Frequency of a Tone Complex*

The last important parameters to be estimated are the fundamental frequencies of the hum tone complexes, even if this fundamental frequency is not present in the signal and only the harmonics are observable. For this reason we used a method similar to the *frequency histogram* method proposed in [156] that is capable of determining the – potentially missing – fundamental frequency of a harmonic tone complex by analyzing its harmonics. We extended that approach to detect the number of harmonic tone complexes, determine their individual fundamental frequencies, and distinguish the harmonics belonging to either one. In order to do so, instead of solely picking the frequency with the highest histogram value, we picked all potential fundamental frequencies – starting with the frequency featuring the highest histogram value – explaining all measured harmonics. To take small estimation errors into account we allowed for a certain frequency deviation when comparing multiples of the potential fundamental frequencies with the measured harmonics.

3.4 Evaluation

The evaluation of the hum detection algorithm consists of two parts:

- The determination of the quantile combination (compare Eq. (3.7)) that leads to the best performance, i.e., high hit rates and low false alarm rates,
- The performance achieved with artificial and realworld test signals when using the optimum quantile combination.

3.4.1 *Test Signals*

In order to evaluate different aspects of the hum detector both artificial signals and real recordings were used.

3.4.1.1 *Artificial Signals*

For the purpose of testing the hum detection algorithm under controllable conditions, a variety of signals were artificially disturbed: Sinusoids of 30 random frequencies between 30 Hz and 900 Hz were added to hum-free signals at different SNRs. The audibility of the hum signal strongly depends on the non-stationary clean signals and the overall SNR could be meaningless if the hum is completely masked by the clean signal. Therefore, in addition we used the PEAQ algorithm [108, 109] in the implementation given in [110] to get a perceptually motivated measure. In order to do so, we added the disturbances to the clean signals at SNRs corresponding to varying Objective Difference Grades (ODG, the PEAQ perceptual quality measure) defined as shown in Table 3.1. The rightmost column of the table shows the median of the SNR values corresponding to the respective ODG ratings over all of our test signals. Due to the diverse nature of the clean signals the interquartile range[5] of the SNR, the IQR, is rather high.

Table 3.1: ODG values, perceived impairment, and median SNR values (compare [109]).

ODG	impairment (MOS)	SNR values (IQR)
0.0	imperceptible (5)	36 (≈ 26)
-0.5		26 (≈ 26)
-1.0	perceptible but not annoying (4)	21 (≈ 26)
-1.5		18 (≈ 26)
-2.0	slightly annoying (3)	14 (≈ 20)
-2.5		7 (≈ 23)
-3.0	annoying (2)	2 (≈ 24)
-3.5		-7 (≈ 22)
-4.0	very annoying (1)	¡ -12

For the hum-free calean signals we chose:

- Stationary white Gaussian noise,
- Random excerpts from a speech signal [157],

---

[5] The interquartile range is the difference between the 75 % and 25 % quantiles.

- Random excerpts from a classical music piece [158],
- Random excerpts from a popular music piece [159],
- Random excerpts from an electronic music piece,
- Random excerpts from a field recording from inside an airplane.

Each of these test signals had a duration of two minutes – the hum disturbance was switched on after 40 s and switched off after 80 s. The SNR in this context was defined to be the energy of the clean signal within the disturbed section compared to the energy of the hum sinusoids:

$$\mathrm{SNR}|_{\mathrm{dB}} = 10\log_{10}\left(\frac{\sum_{k=k_{\mathrm{start}}}^{k_{\mathrm{end}}} s^2[k]}{\sum_{k=k_{\mathrm{start}}}^{k_{\mathrm{end}}} n_a^2[k]}\right), \tag{3.12}$$

where k_{start} and k_{end} denotes the sample indices that correspond to a time of 40 s and 80 s, respectively. The SNR required to obtain the desired ODG values were determined with an error of $\Delta\mathrm{ODG} \leq 0.001$.

To allow for an effective determination of the false alarm probability of the hum detector an equal number of excerpts of hum-free signals was also fed through the algorithm.

3.4.1.2 *Real Recordings*

Although the use of artificial test signals allows for a systematic evaluation, practice has shown that hum disturbances in real recording signals follow the characteristics mentioned in Sec. 3.2 only to a certain extent. Usually the frequencies of hum sinusoids drift very little, but in many cases the power of the hum changes over time. Therefore, to evaluate the behavior of the detection algorithm under realistic conditions real recordings were used. These signals are 24 hours of radio program containing studio speakers, live coverage, music recordings, and jingles.

3.4.1.3 *Ground Truth*

In order to assess the performance of the detection algorithm, information about the true parameters (ground truth) of the hum disturbances is required. For the real recordings, ground truth information was obtained manually: We listened to the signals and used spectral analyzing software to first determine whether hum sinusoids were actually present. If hum was found, we manually identified its exact start and stop times. In addition, the frequency of the hum tones was determined with an accuracy of approximately 1 Hz. We separated the manually detected hum disturbances into the categories *very quiet* (could be detected by an operator using spectral analyzing software during Ground Truth creation but not audible under normal circumstances), *quiet* (compared to the useful signal, the disturbance is very low in power and just audible under normal circumstances), and *disturbing*.

3.4.1.4 *Determination of the Optimum Quantile Ratio*

The overall performance of the detection algorithm depends on the selection of the upper and lower quantiles. Therefore, it is crucial to determine the quantile combination that leads to the best performance in terms of high hit rates and low false alarm rates. Apart from the quantile ratio, the behavior of the algorithm depends on a number of parameters. To reduce the number of free parameters and allow for a clear evaluation, some parameters were set to reasonable values. The length of the median filter in Eq. (3.8), for example, was set to twice the 3 dB bandwidth of the Hann-window that is used for short-term discrete Fourier analysis [160, 161].

The test signals used in this section were the same as described in Sec. 3.4.1.1, and we added the disturbance in order to obtain 11 SNRs, where the SNR is defined in Eq. (3.12).

3.4.1.4.1 Optimization Criterion

As a performance measure (or error function) that should be minimized we use the total error probability as a function of the decision threshold Θ,

$$P_{\text{error}}(\Theta) = P_{\text{miss}}(\Theta) + P_{\text{false alarm}}(\Theta), \tag{3.13}$$

where $P_{\text{miss}}(\Theta)$ and $P_{\text{false alarm}}(\Theta)$ denote the probability of miss or false alarm, respectively. The procedure for estimating the miss and false alarm probabilities is based on histograms of the quantile ratio, compare Eq. (3.7), – for frequencies that contain a hum disturbance and for frequencies that do not contain hum disturbances independently. The histograms are obtained by taking into account the quantile ratio value for all test signals, all SNRs and all hum frequencies for every quantile combination. Thus, for each of the quantile combinations two histograms are obtained. The following lower and upper quantiles are examined:

$$\begin{aligned}\mathbf{Q}_{\text{lower}} &= \begin{bmatrix}1 & (5 & 10 & \cdots & 45)\end{bmatrix}\\ \mathbf{Q}_{\text{upper}} &= \begin{bmatrix}(5 & 10 & \cdots & 95) & 99\end{bmatrix}.\end{aligned}$$

With the restriction that the upper quantile must be greater than the lower quantile, this results in 160 combinations altogether. The miss and false alarm probabilities are estimated from the normalized histograms:

$$\begin{aligned} P_{\text{false alarm}}(\Theta) &= \sum_{l=0}^{l_\Theta} h_{\text{with hum}}[l] && (3.14) \\ P_{\text{miss}}(\Theta) &= \sum_{l=l_\Theta+1}^{N_{\text{bins}}-1} h_{\text{no hum}}[l]. && (3.15) \end{aligned}$$

where $\sum_{l=0}^{N_{\text{bins}}-1} h_{\text{with hum}}[l] = 1$ and $\sum_{l=0}^{N_{\text{bins}}-1} h_{\text{no hum}}[l] = 1$ are normalized histograms and l_Θ denotes the histogram bin index corresponding to a threshold value of Θ. The minimum achievable error and the corresponding optimum threshold follow from Eq. (3.13):

$$\begin{aligned} P_{\text{error, min}} &= \min\{P_{\text{error}}(\Theta)\} && (3.16) \\ \Theta_{\text{opt}} &= \arg\min_{\Theta}\{P_{\text{error}}(\Theta)\}. && (3.17) \end{aligned}$$

Fig. 3.7 shows the total error probability depending on the threshold value. For very small threshold values ($\Theta < 0.03$), the total error consists only of false alarms because almost all quantile ratios are interpreted as if there was a hum disturbance. For $\Theta > 0.15$ almost no false alarms are triggered but the total error increases as more and more hum disturbances are dismissed because the resulting quantile ratio is below the threshold.

The minimum error for each combination of lower and upper quantiles is shown in Fig. 3.8. The smallest overall error, computed over all SNR values, was achieved with a lower quantile of 10 % and an upper quantile of 55 %, with a corresponding optimal threshold of $\Theta_{\text{opt}} = 0.18$.

3.4.1.5 *The Error Measures*

Other aspects of the detection algorithm can be evaluated by the following error measures:

3.4.1.5.1 Hit and false alarm rate

As a first measure, the probability of the algorithm erroneously detecting hum where none is present (false alarm) or to miss hum disturbances is a very important information. To compute the hit rate, we calculated the overall time of correctly detected sinusoids, T_{hit}[6], divided by the total hum duration, T_{hum}: $h_{\text{hit}} = T_{\text{hit}}/T_{\text{hum}}$.

In a similar fashion, the false alarm rate is determined by relating the duration of indicated hum disturbances where no hum was present, T_{fa}, to the total duration

---

[6] In this article all times are specified in seconds unless stated otherwise.

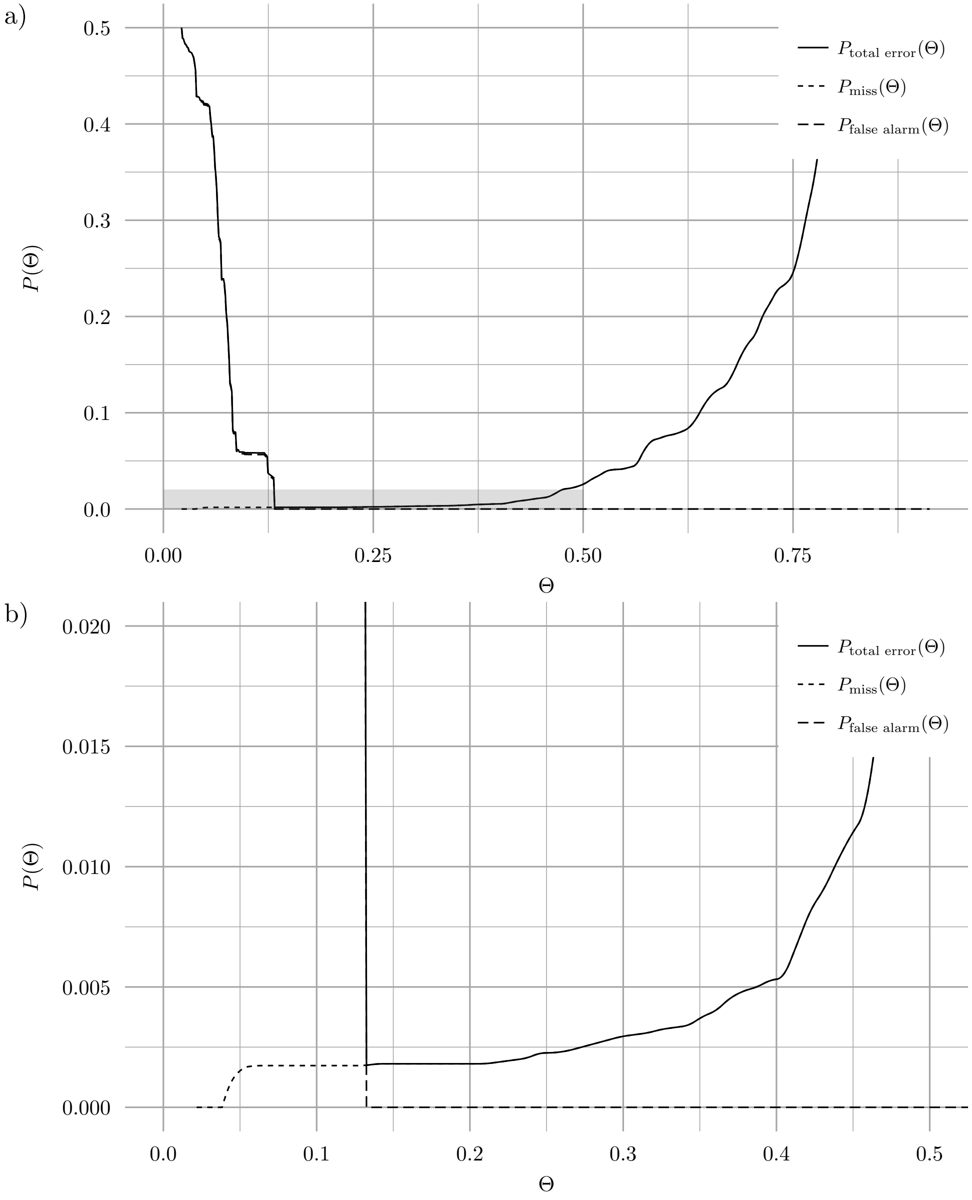


Figure 3.7: Total error probability vs. threshold Θ. (The lower quantile is 5 % and the upper quantile is 60 %, the SNR is set to 20 dB). a) Whole possible range of Θ. b) Zoom of the region that is indicated by the rectangle in plot a).

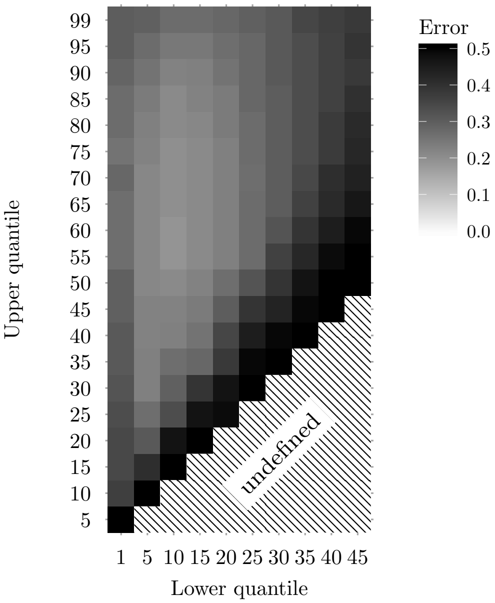


Figure 3.8: The minimum error depending on the quantile combination. For each combination the optimum threshold was used.

of hum-free signal sections that was available from ground truth, $T_{\text{no hum}}$ (see Sec. 3.4.1.3): $h_{\text{fa}} = T_{\text{fa}}/T_{\text{no hum}}$.

3.4.1.5.2 Accuracy of frequency estimation

To calculate the deviation of the detected hum frequencies from the true values, the absolute value of the difference of both frequencies was determined for each test case: $\Delta f = f - \hat{f}$. The frequency estimation accuracy was not determined for the real recordings due to the problem of determining the ground truth with sufficient precision.

3.4.1.5.3 Start and stop time deviation

To evaluate the estimated start and stop times, we computed their difference to the true values: $\Delta T_{\{\text{start,stop}\}} = T_{\{\text{start,stop}\}} - \hat{T}_{\{\text{start,stop}\}}$.

3.4.1.6 *Baseline Algorithm*

In order to compare our algorithm to an intuitive method to find hum frequencies, we implemented a simple algorithm based on minima tracking within a certain time window as a baseline algorithm (abbreviated by “MT”). The minimum tracking algorithm is based on the same framework as our algorithm – e.g., down-sampling to 2 kHz, short-time discrete Fourier transformation, and buffering (we used the same buffer length as for the presented algorithm). The *tone detector* stage (compare Sec. 3.3.1) is different though. Instead of computing the quantile ratio and smoothing intermediate values, the minima in the buffer $\mathbf{b}_{\text{in}}^{(p)}$ are picked in each frequency bin. Hum frequencies are detected by thresholding this vector of minimum powers. The threshold is computed in the following way:

$$\Theta_{\text{MT}} = \alpha_{\text{MT}} \cdot \overline{\mathbf{P}_{\text{min}}}. \tag{3.18}$$

The factor α_{MT} is set to a value that minimizes the total error. This optimum value is determined by the method described in Sec. 3.4.1.4. The threshold defined by Eq. (3.18) is motivated by the fact that hum sinusoids are characterized by their relative power compared to the rest of the spectrum, not by their absolute power.

3.4.1.7 *Results*

In this section the results of the hum detection algorithm are compared with ground truth. The threshold value in Eq. (3.11) was set to the optimal one that was determined by the method described in Sec. 3.4.1.4.

3.4.1.7.1 Artificial Test Signals

For the given artificial test scenario we got the following results for the measures depending on the adjusted ODG:

Hit and false alarm rate The hit and false alarm rates achieved with the artificial test signals in dependence on the ODG value are shown in Fig. 3.9. For ODGs of −2 and below the hit rate stays above 90 % and falls to approximately 50 % for an ODG of −0.5. However, although this seems like a sub-optimal detection performance, informal listening tests have shown that at an ODG of −0.5 the hum disturbances are barely audible. The rather low hit rate at an ODG of 0 is not likely to pose a problem in everyday practice since the disturbance is inaudible in almost all cases for an ODG of 0. Furthermore, all false alarms are caused by errors in the start and stop time estimation (compare Figs. 3.11 and 3.12). In no case was a hum free signal erroneously believed to contain hum. In comparison to the baseline system, the proposed algorithm always has a better hit rate with a relative gain up to 20 %, even though this scenario with a pure static artificial sinusoid is optimal for the MT algorithm.

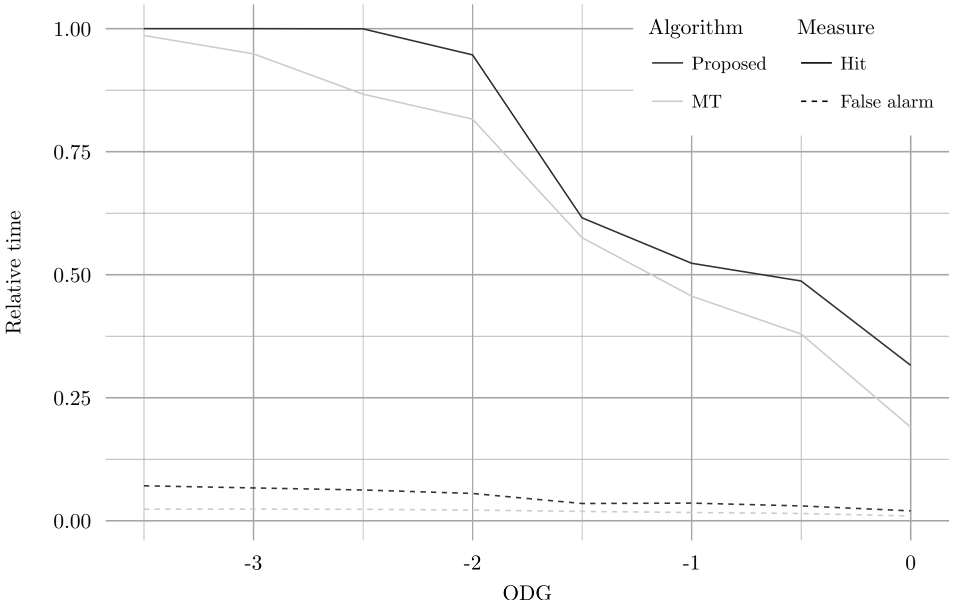


Figure 3.9: Hit and false alarm rate for artificial test signals. The results of the presented algorithm are printed in black, the results of the baseline algorithm (MT) are in grey. The false alarm rate is solely caused by imprecise start and stop time estimates – in no case was a hum free signal believed to contain a hum disturbance.

Frequency deviation The frequency estimation error achieved over ODG rating is shown in Fig. 3.10 a). The box plot shows that for all ODG ratings the absolute frequency estimation error is below 0.5 Hz. Compared to the coarse estimation given by $f_\triangle$ it is a vast improvement and allows for automatic removal algorithms. For comparison purposes the estimation error over the input SNR is shown in Fig. 3.10 b). For SNR values above 30 dB the estimation error increases compared to corresponding ODG ratings (compare the coarse ODG → SNR mapping given in Table 3.1). The frequency estimation depends partially on the input SNR and on the diverse nature of the input signals. Therefore, a direct mapping of ODG and SNR results is not possible.

Accuracy of start and end time estimation The start and end time estimation errors achieved with the artificial test signals are shown in Figs. 3.11 and 3.12, respectively. The start time error is positive for most test runs, indicating too early detections. On average, hum is detected approximately 1.5 s early for almost all ODGs and SNRs. The error in the end time estimation is approximately -1 s in most cases, where the minus indicates that the end of most hum disturbances is detected too late. Both errors could be reduced by adjusting the detection times according to the mean values. Overall, the figures show that only a few outliers

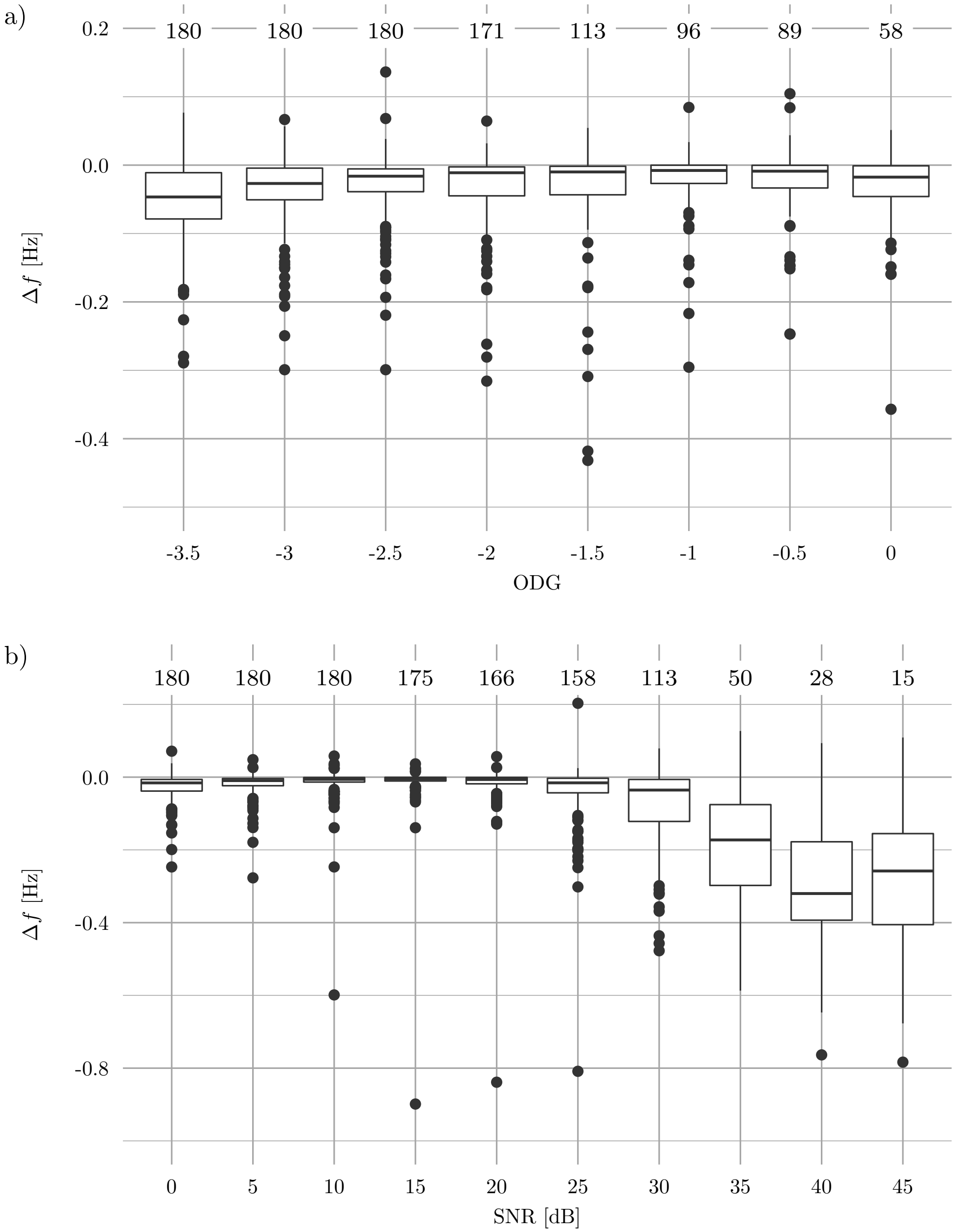


Figure 3.10: Frequency estimation error for artificial test signals. Plot a) shows the results in dependence on the ODG rating of the input signal while plot b) shows the estimation error against the input SNR. The number of values that were considered for each ODG, or SNR, respectively, are written above the boxes.

Table 3.2: Detection performance with real-world input data. Shown is the ground truth hum duration and correctly detected hum duration for the proposed algorithm (a) and the baseline method (b). Hit and false alarm rates given are determined by dividing ground truth hum duration by the duration of correctly detected hum. False alarm times and percentages excluding early detection beginnings and late endings are in parentheses.

a) Presented algorithm

Hum intensity	GT times	HD times	Rel. freq. [%]	Measure
No hum	23h05m52s	1h03m48s (52m13s)	4.60 (3.77)	False alarm
Very quiet	4m16s	1m27s	33.93	Hit
Quiet	1h03m52s	16m53s	26.44	
Disturbing	3h44m52s	2h27m08s	65.43	

b) Baseline method

Hum intensity	GT times	HD times	Rel. freq. [%]	Measure
No hum	23h05m52s	4h32m41s (4h29m35s)	19.68 (19.45)	False alarm
Very quiet	4m16s	1m24s	32.76	Hit
Quiet	1h03m52s	13m59s	21.90	
Disturbing	3h44m52s	1h40m12s	44.56	

exist, and for almost all files the start and end times would be correct if adjusted to zero by subtracting the mean value.

3.4.1.7.2 Real Recordings

Table 3.2 lists the duration of hum disturbances that were correctly detected by the hum detector, separated into the audibility categories described in Sec. 3.4.1.3. In this context, a correct detection was counted when

- The detected hum disturbance was at least partly overlapping with the ground truth hum,
- The frequency deviation between the detection result and the true value was less than 1 Hz.

Furthermore, the duration of hum actually present in the signals is shown. To determine the hit rate, the duration of correctly detected hum was divided by the true hum duration, yielding the results given in Table 3.2.

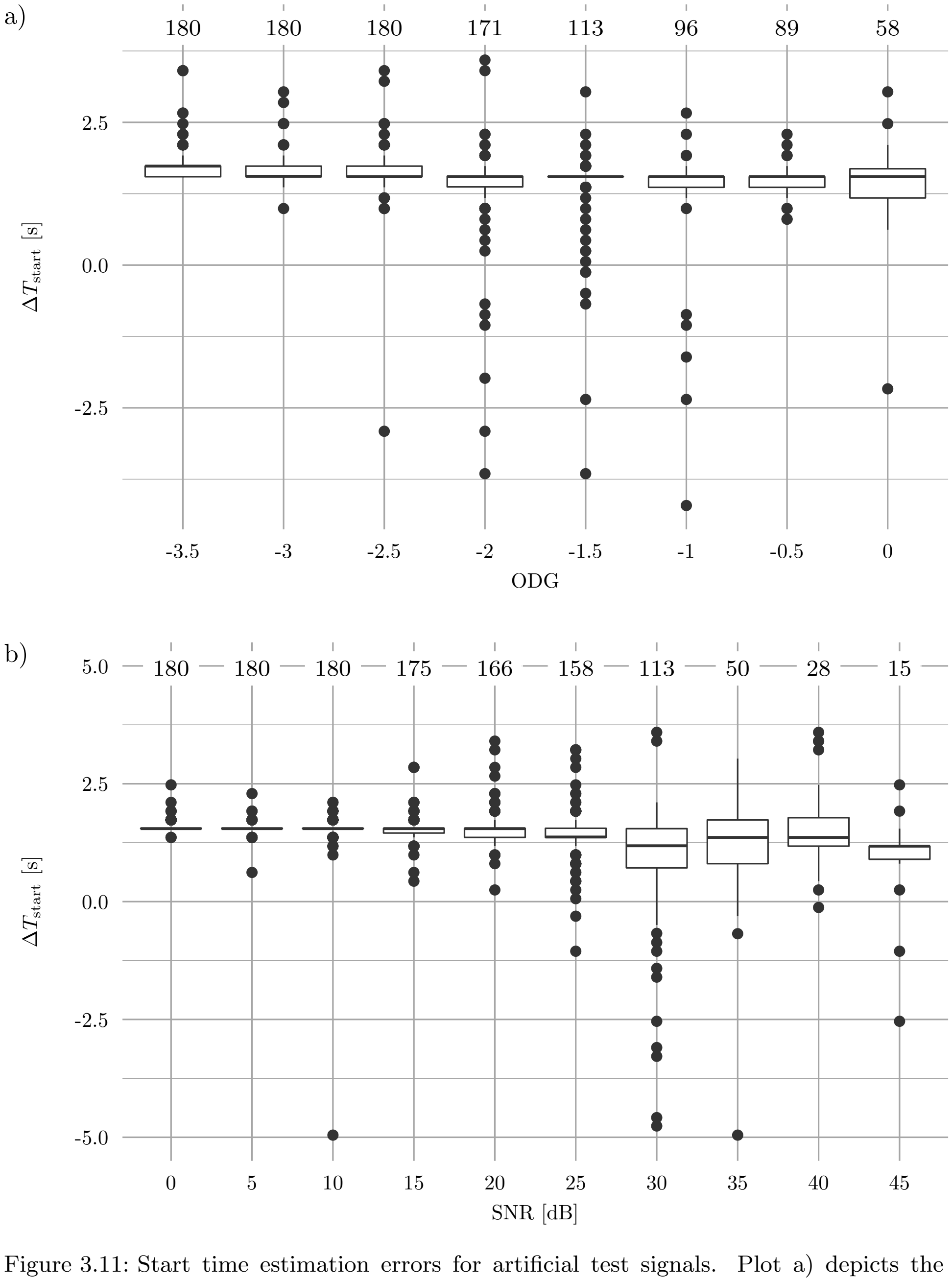


Figure 3.11: Start time estimation errors for artificial test signals. Plot a) depicts the results over the ODG rating of the input signal, plot b) shows the estimation errors in dependence on the input SNR. The number of values that were considered for each ODG, or SNR, respectively, are written above the boxes.

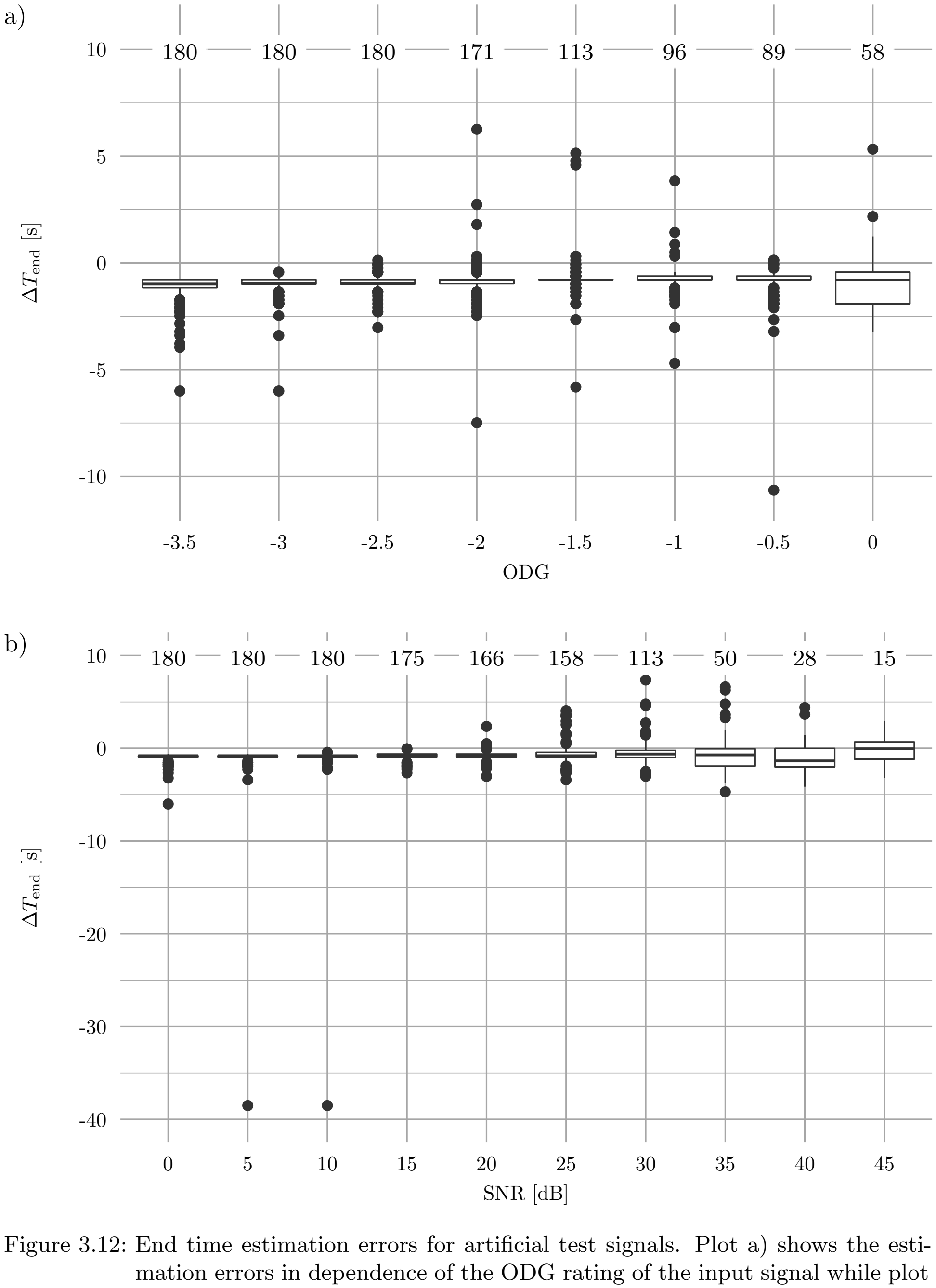


Figure 3.12: End time estimation errors for artificial test signals. Plot a) shows the estimation errors in dependence of the ODG rating of the input signal while plot b) illustrates the results in dependence on the input SNR. The number of values that were considered for each ODG, or SNR, respectively, are written above the boxes.

If we compare the baseline algorithm with the proposed method, we can see that the false alarm rate is significantly reduced. This is important to limit harm to clean material if an automatic hum removal algorithm is applied. At the same time the hit rate was increased, which shows the far better detection capabilities of the proposed algorithm. However, the overall detection rate seems small compared to the test with the artificial test signals, especially for the loud and disturbing hum, since these signals should be treated in an automatic system, so we analyzed the signals that were not detected correctly. In almost all cases the desired signal was speech. Since our test signals were not original archive material but a broadcast version of it, we found a severe dynamic compression of the material. This broadband compression algorithm changed the dynamic of the hum sinusoids up to 20 dB at the rate of the speech signal. Thus, these fluctuating hum signals violate our signal model which assumes static hum tone with minor variations. For music signals containing hum, the overall compression is much smoother and therefore our detection algorithm is able to detect the hum even in broadcast material.

3.5 Conclusions

We have shown in this article that the automatic detection of hum disturbances in audio signals is feasible. An evaluation of a detection method yields that most of the hum disturbances that are actually perceivable can be detected reliably with a low false alarm probability. The information that is obtained by the method presented can be used to control existing algorithms to effectively remove the disturbances. We believe that – next to detectors and removal algorithms for other types of degradations, like broadband noise and impulsive disturbances – the automatic detection and removal of hum is one step in the direction of a fully automatic restoration approach.

4

REDUCTION OF HUM DISTURBANCES

In this contribution we analyze different filtering algorithms for removing hum disturbances from audio recordings. In order to protect the desired signal, high frequency selectivity of the used filters is necessary. However, due to the time-bandwidth uncertainty principle, high frequency selectivity brings about long impulse responses. This can result in audibly resonating filters, causing artefacts in the output signal. Thus, the choice of the optimal algorithm is a compromise between frequency selectivity and acceptable time domain behavior. In this context, different filter structures and algorithms have different characteristics. To investigate their influence on the hum disturbance and the desired signal, we have evaluated three methods using objective measures to illustrate advantages and drawbacks of the individual approaches.

4.1 Introduction

One of the fundamental tasks in the context of audio restoration is the removal of additive sinusoidal disturbances that are commonly known as *hum*. These disturbances are usually caused by power line interference problems during recording and/or copying processes and can severely reduce the audio quality of a recording. Due to the origin of this type of disturbance, namely faulty shielding of audio equipment or signal lines in most cases, the disturbance signals can be modelled as harmonic tone complexes with fundamental frequencies of approximately 50 Hz or 60 Hz and a number of harmonics that are – in most cases – caused by nonlinearities in the signal chain.

This chapter contains a copy of the article

M. Brandt, J. Bitzer, “Hum Removal Filters: Overview and Analysis,” *Proceedings of the 132nd Audio Engineering Society Convention*, Budapest, Hungary (2012 Apr.).

While the layout of the article has been adapted for a uniform presentation within this thesis, the contents printed here are identical to those in the published article.

Ideally, the removal of hum disturbances would consist in removing the hum sinusoids only and hence leaving the desired part of the disturbed signal completely unaffected by the restoration. However, this approach would require complete knowledge about the frequency, amplitude and phase of the additive sinusoids for perfect cancellation. Although sinusoid parameter estimation is actually feasible (see e.g. [162]), this turns out to be a highly challenging task for sinusoids embedded in audio signals, since the audio is much louder. Additionally, gross estimation errors eventually lead to an addition of sinusoids, which cannot be tolerated. Therefore, state-of-the-art approaches to reduce the power of the disturbed signal at the frequencies of the hum sinusoids are based on filtering in the time domain. This requires filters that feature high frequency selectivity, i.e. very narrow dampening regions while leaving the desired part of the signal mostly unaffected. However, due to the uncertainty principle, high frequency selectivity causes long impulse responses which can result in filter resonance artefacts in the output signal. The choice of the hum removal filter properties is thus a compromise between the degree of hum reduction on the one hand and the amount of desired signal cancellation and artefacts that are introduced by the filter itself on the other hand.

To compare different approaches, we have implemented three hum removal filter algorithms and analyze their behavior with different input signal types.

Basis of our investigations is a signal model that is explained briefly in Section 4.2. The analyzed filtering algorithms are introduced in Section 4.3, followed by a description of the evaluation method in Section 4.4. After giving the results of our analysis we end with some conclusions.

4.2 The Signal Model

We model the disturbed signal $x(t)$ as follows

$$x(t) = s(t) + n_{\mathrm{hum}}(t)$$

where $s(t)$ is the unobservable clean signal and $n_{\mathrm{hum}}(t)$ is the hum signal, consisting of a number of sinusoids with different frequencies, phases and amplitudes. The process of hum removal filtering is denoted by the hum removal system operator $H_{\mathrm{HR}}\{\bullet\}$. As all of the analyzed filters are linear, the processed signal can be written

$$H_{\mathrm{HR}}\{x(t)\} = H_{\mathrm{HR}}\{s(t)\} + H_{\mathrm{HR}}\{n_{\mathrm{hum}}(t)\}.$$

and we are thus able to analyze the effect of the processing to the desired signal and the hum signal independently.

4.3 Hum Removal Filter Algorithms

We implemented the following algorithms, each having different parameters:

- FIR comb filter
  The fundamental frequency of this filter can be adjusted. Rejected frequencies are at integer multiples of the fundamental frequency, up to the Nyquist frequency (compare [163]). An exemplary comb filter transfer function and the corresponding impulse response are shown in Figure 4.1.
- Subband FIR comb filter
  The subband FIR comb filter is based on a third-order bandsplitting algorithm (see [164, 165]) to seperate the input signal into complementary high and low frequency bands. The crossover frequency can be chosen according to the properties – i.e. the harmonic frequency power decay – of the hum disturbance at hand. In Figure 4.2 the transfer functions for the low and high frequency bands are shown. After seperating the bands, only the lowpass band is processed with an FIR comb filter to reduce the hum disturbance and, afterwards, the bands are summed back together to yield the processed output signal. To take the slope of the lowpass band splitting filter into account, we set the band splitting frequency to twice the frequency of the highest hum harmonic.
- Allpass based notch filter
  The center frequency, notch depth and bandwidth can be adjusted (compare [164]). For each hum harmonic, an individual notch filter is required. An exemplary transfer function and impulse response of this filter are shown in Figure 4.4.

4.4 Evaluation

For being able to analyze the effect of the filtering to the hum signal $n_{\mathrm{hum}}(t)$ and the desired signal $s(t)$, both signals were processed independently.

4.4.1 *The Measures*

In order to determine the degree of hum power reduction on the one hand and assess the influence on the desired signal on the other hand, the following error measures are introduced which compare the input signal of the filters to their respective output signal.

- Hum reduction
  This is a measure for the amount of broadband hum power reduction that is achieved by the filter. It is the ratio of the overall energy of the hum signal $n_{\mathrm{hum}}(t)$ to the overall energy of the filtered hum signal $H_{\mathrm{HR}}\{n_{\mathrm{hum}}(t)\}$.

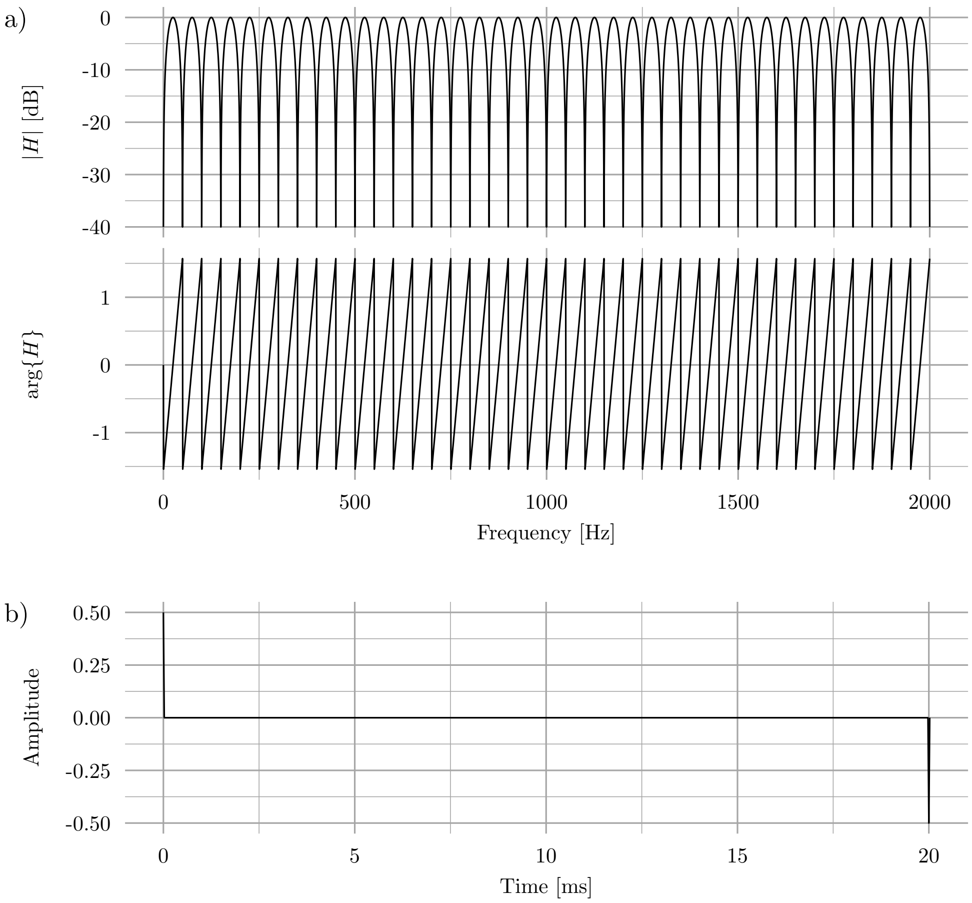


Figure 4.1: Transfer function (plot a) and impulse response (plot b) of the comb filter. The fundamental frequency has been set to 50 Hz, resulting in a magnitude transfer function with zeros at frequencies of (50, 100, ...)Hz.

- Desired signal distortion
  This is a measure for the amount of broadband distortion of the desired signal that is introduced by the algorithm. It is the ratio of the overall energy of the desired signal $s(t)$ to the overall energy of the filtered desired signal $H_{\mathrm{HR}}\{s(t)\}$.

4.4.2 *The Test Signals*

The test signals are three different types of desired signals which have been disturbed artificially by adding three different hum disturbance tone complexes, resulting in a total of nine combinations. In detail, the clean signals were

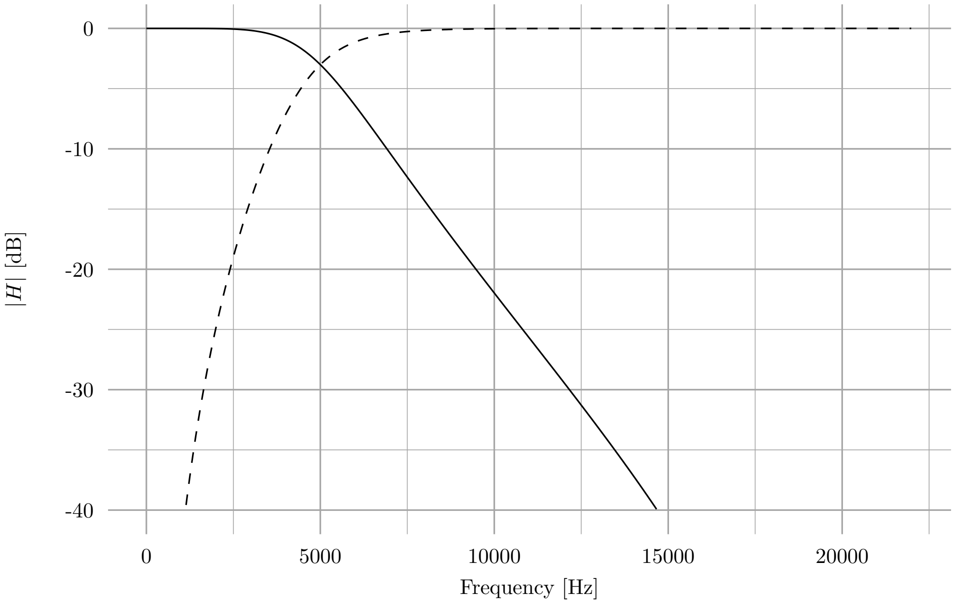


Figure 4.2: Magnitude transfer functions of the complementary band splitting filters. The crossover frequency has been set to $f_c = 5\,\text{kHz}$. The sum of the lowpass (solid line) and highpass (dashed line) transfer functions equals one.

- a speech segment from a radio recording,
- an excerpt of a classical piano music recording,
- an excerpt of a pop music recording.

The hum disturbances were

- an artificially generated sinusoid with a frequency of 50 Hz,
- an artificially generated harmonic tone complex consisting of three sinusoids with frequencies of 50 Hz, 150 Hz and 250 Hz, which is a typical hum disturbance,
- a recording of a severe real-world hum tone complex [166] with audible disturbances up to 1.5 kHz.

All test signals had a length of 40 s.

4.5 Results

To obtain the figures in this section, all test signals have been processed by the hum removal filtering algorithms and the error measures have been computed. In doing

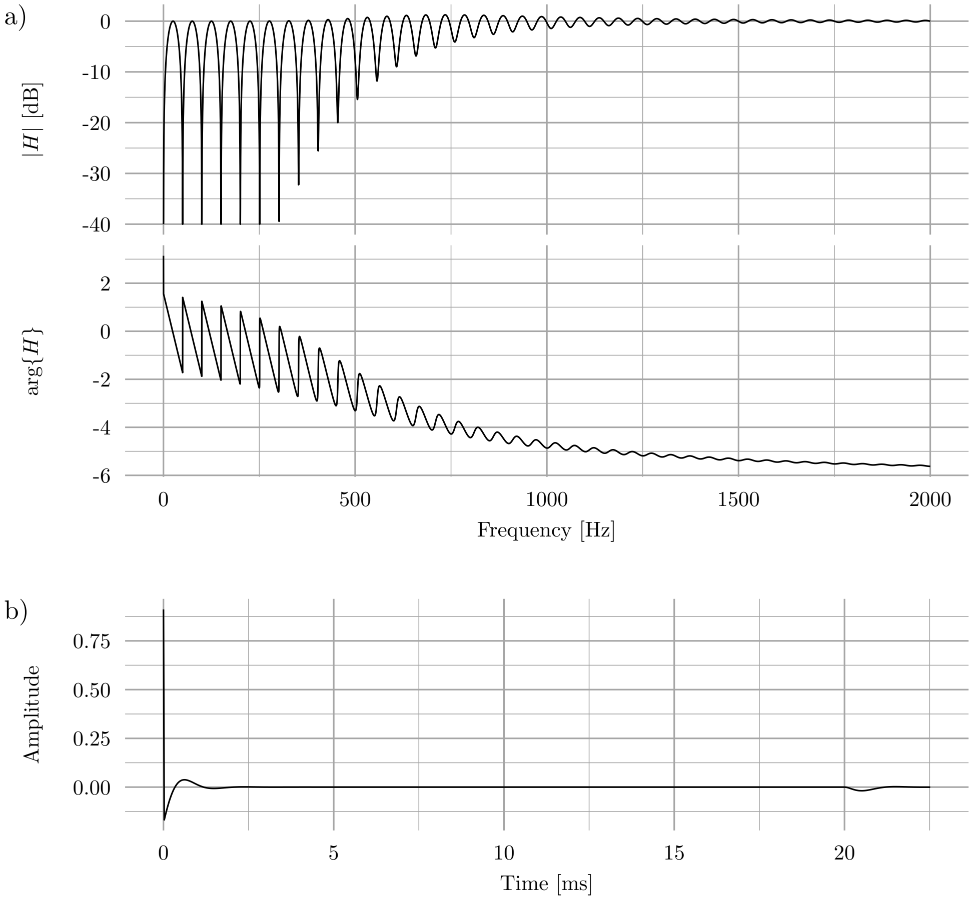


Figure 4.3: Transfer function (plot a) and impulse response (plot b) of the subband comb filter. The fundamental frequency has been set to 50 Hz and the crossover frequency of the bandsplitting filters has been set to 1.5 kHz.

so, the error measure for a specific parameter combination was averaged over all test signals.

4.5.1 *FIR Comb Filter*

The FIR comb filter does not feature parameters that can be changed to optimize its frequency and time domain characteristics for a specific application. The only parameter is the fundamental frequency whose value is dictated by the properties of the hum disturbance at hand. The hum reduction and desired signal distortion measures are given in Table 4.1. The hum reduction capabilities are good and, as the standard deviation over all test signals is very small, the comb filter reduces hum power equally well with different kinds of hum characteristics. The measure for

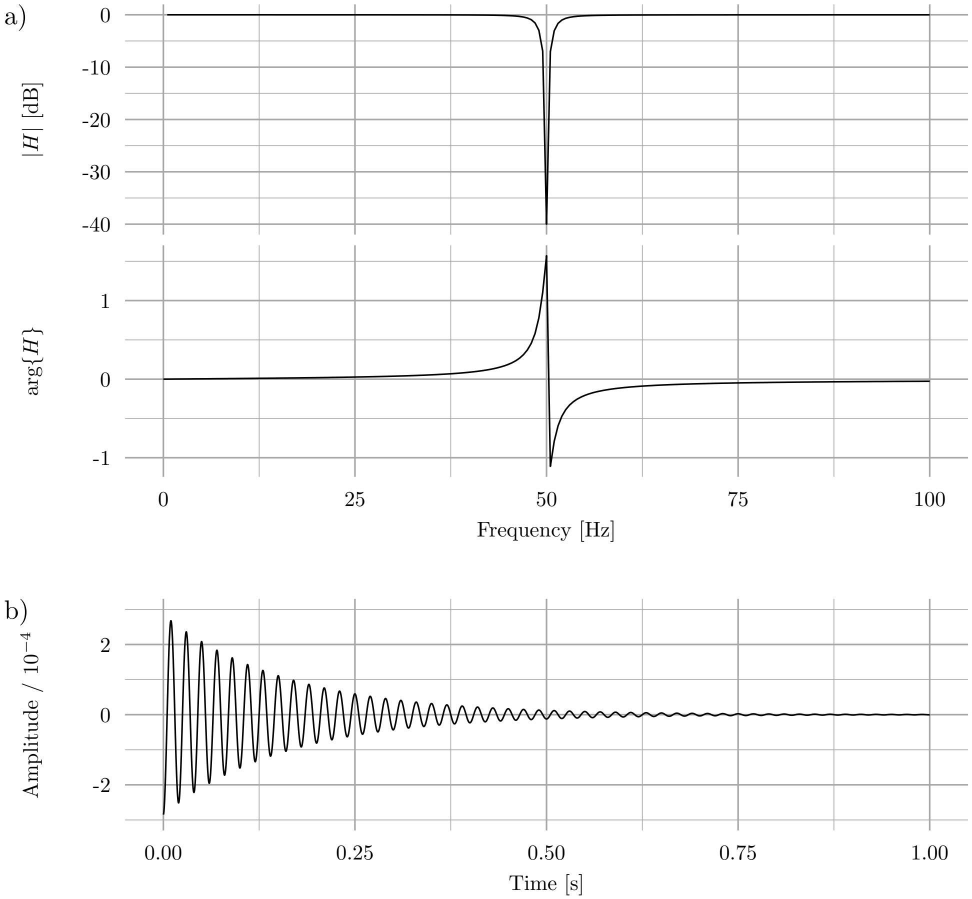


Figure 4.4: Transfer function (plot a) and impulse response (plot b) of the allpass notch filter. In this example, the notch center frequency has been set to 50 Hz, the bandwidth is 2 Hz and the notch depth is ∞ dB. The first sample of the impulse response is not plotted in order to better visualize the decay behaviour.

distortion of the desired signal contains both the power reduction due to the filtering and also artifacts caused by the characteristic FIR comb filter impulse response, as a delay effect is perceived in the output signal.

4.5.2 *Subband FIR Comb Filter*

The subband FIR comb filter does not feature parameters that can be changed to optimize its frequency and time domain characteristics for a specific application. However, the splitting frequency can be chosen to match the characteristics of a specific hum disturbance. In our evaluation, we set the band splitting frequency so that all hum harmonics were contained in the lowpass band. The resulting hum

Table 4.1: Performance measures for the FIR comb filter. Given is the mean value over all test signal combinations and the standard deviation (σ).

Error measure	Mean value	σ
Hum reduction	$-33.01\,\mathrm{dB}$	$0.01\,\mathrm{dB}$
Desired signal distortion	$3.18\,\mathrm{dB}$	$0.51\,\mathrm{dB}$

reduction and desired signal distortion measures are given in Table 4.2. Because the depth of the teeth of the comb filter in the lowpass band decreases with higher frequencies, the amount of hum power reduction is smaller than for the broadband comb filter. The large standard deviation indicates that the efficiency is largely dependent on the characteristics of the hum disturbance and the choice of band-splitting frequency. However, the desired signal distortion is decreased compared to the broadband FIR comb filter.

Table 4.2: Performance measures for the subband FIR comb filter. Given is the mean value over all test signal combinations and the standard deviation (σ).

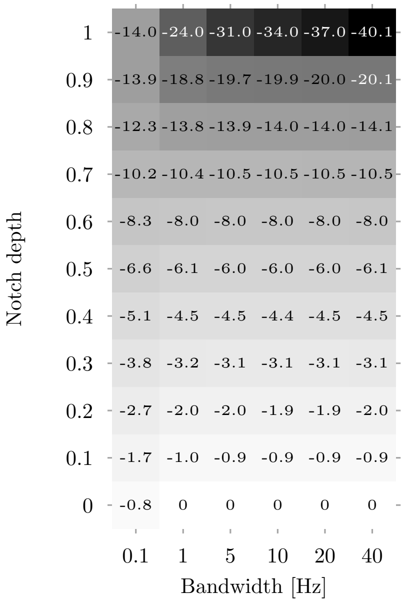
Error measure	Mean value	σ
Hum reduction	$-18.17\,\mathrm{dB}$	$8.23\,\mathrm{dB}$
Desired signal distortion	$1.55\,\mathrm{dB}$	$1.24\,\mathrm{dB}$

4.5.3 *Allpass Based Notch Filter*

Figures 4.5 and 4.6 show the hum reduction and desired signal distortion, respectively, in dependence on the parameter combination. Notch depths of 0 indicate no processing at all, and 1 indicates the maximum attenuation of $\infty\,\mathrm{dB}$. It can be seen in Figure 4.5 that the difference in hum power reduction is comparatively small. On the other hand, the cancellation of the desired signal is increased for higher notch bandwidths – compare Figure 4.6.

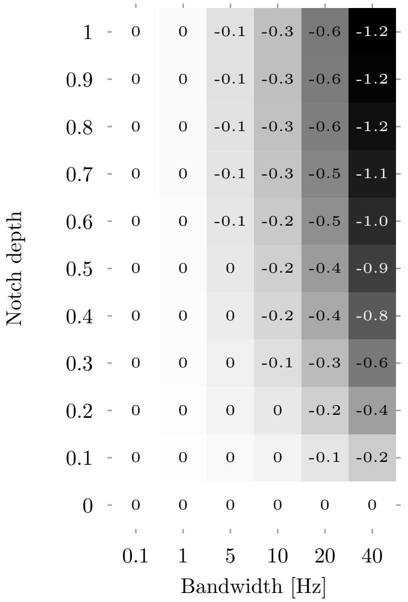
4.6 Conclusions

We showed that the three most common approaches for reducing hum disturbances in audio signals have different properties in terms of hum reduction and desired



Notch depth	0.1	1	5	10	20	40
1	-14.0	-24.0	-31.0	-34.0	-37.0	-40.1
0.9	-13.9	-18.8	-19.7	-19.9	-20.0	-20.1
0.8	-12.3	-13.8	-13.9	-14.0	-14.0	-14.1
0.7	-10.2	-10.4	-10.5	-10.5	-10.5	-10.5
0.6	-8.3	-8.0	-8.0	-8.0	-8.0	-8.0
0.5	-6.6	-6.1	-6.0	-6.0	-6.0	-6.1
0.4	-5.1	-4.5	-4.5	-4.4	-4.5	-4.5
0.3	-3.8	-3.2	-3.1	-3.1	-3.1	-3.1
0.2	-2.7	-2.0	-2.0	-1.9	-1.9	-2.0
0.1	-1.7	-1.0	-0.9	-0.9	-0.9	-0.9
0	-0.8	0	0	0	0	0

Figure 4.5: Hum reduction properties of the allpass based notch filter in dependence on the notch bandwidth and the notch depth. Zeros in the plot indicate absolute values < 0.1.



Notch depth	0.1	1	5	10	20	40
1	0	0	-0.1	-0.3	-0.6	-1.2
0.9	0	0	-0.1	-0.3	-0.6	-1.2
0.8	0	0	-0.1	-0.3	-0.6	-1.2
0.7	0	0	-0.1	-0.3	-0.5	-1.1
0.6	0	0	-0.1	-0.2	-0.5	-1.0
0.5	0	0	0	-0.2	-0.4	-0.9
0.4	0	0	0	-0.2	-0.4	-0.8
0.3	0	0	0	-0.1	-0.3	-0.6
0.2	0	0	0	0	-0.2	-0.4
0.1	0	0	0	0	-0.1	-0.2
0	0	0	0	0	0	0

Figure 4.6: Amount of desired signal distortion of the allpass based notch filter in dependence on the notch bandwidth and the notch depth. Zeros in the plot indicate absolute values < 0.1.

signal distortion. By carrying out an evaluation based on objective performance measures, we provide a means of selecting a hum removal algorithm for a specific application.

5

BROADBAND NOISE PSD ESTIMATION

This paper presents a novel algorithm to estimate the power spectral density (PSD) of stationary broadband noise disturbances in audio recordings. The proposed algorithm estimates the noise PSD as the mean value of an exponential distribution which corresponds to the truncated periodogram coefficients of the disturbed audio signal. An evaluation with a large number of speech and music test signals shows that a high PSD estimation accuracy can be obtained for a wide range of signal-to-noise ratios, allowing for unsupervised operation and thus constituting an important part of a fully automatic broadband noise restoration system for audio archives.

5.1 Introduction

The quality of audio recordings is often degraded by various types of disturbances, such as broadband noise, hum, clicks and crackles [38, 39, 113, 114]. Broadband noise is one of the most frequently occurring types of disturbance, especially in old recordings, and can be classified according to their *technical* or *acoustic* origin [38, Ch. 2.1.1]. Technical broadband noise disturbances, also known as *hiss*, typically arise because of shortcomings of recording equipment or storage media. Acoustic broadband noise disturbances, on the other hand, have their origin in acoustic phenomena, such as cars passing by, wind, or the hissing of an ocean. While it can be a difficult task to determine whether a certain acoustic element of a recording corresponds to an acoustic noise disturbance which typically requires semantic information, in many cases the identification of technical noise disturbances is much clearer. For example, it is a well-known fact that every audio recording brings about

---

This chapter contains a copy of the article

M. Brandt, S. Doclo, J. Bitzer, "Automatic Noise PSD Estimation for Archive Audio Restoration," *Submitted to the Journal of the Audio Engineering Society* (2018 Mar.).

While the layout of the article has been adapted for a uniform presentation within this thesis, the contents printed here are identical to those in the submitted article.

a degradation of the original acoustic signal. Early recording media, such as wax cylinders and shellac discs, typically had SNRs of below 40 dB [10]. The vinyl disc represented a large improvement in the dynamic range that could be stored, with SNRs of 55 dB to 60 dB [9]. The next improvement was obtained with magnetic tape storage media, leading to SNRs between 60 dB and 70 dB [167] and allowing for sound qualities that still satisfy the expectations of today's listeners. The compact disc, commercially introduced in 1982, made dynamic ranges above 90 dB possible [168]. Nowadays, digital audio formats obviously allow for dynamic ranges that are as high as desired, merely by increasing the word length for each sample of the audio signal.

Hence, due to the progress in recording and storage technology, recordings made today usually do not suffer from audible technical broadband noise disturbances. Nevertheless, in the last decades considerable effort has been spent to digitize historical recordings from a variety of original media and to reduce broadband noise disturbances in these audio recordings [33, 47, 169]. In doing so, a central problem is the estimation of the characteristics of the broadband noise disturbance. State-of-the-art audio restoration algorithms [38, 62, 84] often require the manual selection of (one or more) noise-only sections of a recording to determine a so-called *noise fingerprint* [39]. Assuming stationarity of the noise disturbance, this fingerprint is then used as an estimate for the PSD of the underlying noise disturbance. The restoration quality of a noise reduction algorithm crucially depends on the accuracy of the noise PSD estimate [38], resulting in insufficient noise reduction if the noise PSD estimate is too low and resulting in degradation of the desired signal if the noise PSD estimate is too high.

While the manual selection of noise-only sections is not a problem for a selected number of very valuable recordings, e.g., early piano recordings of Edvard Grieg [29], recorded in 1903, the restoration of huge amounts of recordings stored in audio archives is usually not feasible due to the required manual intervention for each individual recording. It should be realized that the amount of audio material stored in archives around the globe is immense: the Library of Congress alone reports more than 3.5 million audio media in 2014 [23], comprising, e.g., music recordings, interviews and field recordings. Due to the large variety with regard to the type of the desired signal, recording technology and age of the media, these recordings show a large diversity of broadband noise types at a wide range of SNRs—from below 30 dB for old wax cylinder recordings to practically noise-free recordings of today. If restoration of large audio archives is desired, the only feasible option is automatic processing. For noise reduction it is therefore crucial to automatically obtain an accurate estimate of the noise PSD, for low as well as for high input SNRs. As mentioned before, many recordings stored in an archive do not even contain audible broadband noise, implying that the optimum choice may be not to perform any restoration at all for these recordings.

To the best of our knowledge, no noise PSD estimation algorithms exist that are robust against a large range of input SNRs and against a large variety of desired signals. Although many efficient noise PSD estimation algorithms, e.g., [91, 92, 94, 170], have been proposed to enhance noisy speech signals in communication applications such as conferencing systems or hearing aids (cf. Section 5.1.2), the requirements for audio restoration of archives are substantially different. First, in speech communication applications the input signal is typically assumed to be a noisy recording of a single speaker, whereas in audio archives the input signal is much more diverse and complex (e.g., music, singing voice, multiple speakers). Furthermore, in speech communication applications the noise is typically assumed to be of acoustic nature and to be time-varying, whereas in audio archives the noise can usually be assumed to be of technical nature and to be (rather) constant for each individual recording. Finally, the main goal of noise reduction in speech communication applications is to improve speech intelligibility, whereas in archive audio restoration the main goal is to achieve well sounding, high-resolution restoration results.

5.1.1 *Main Idea of the Proposed Algorithm*

The main idea of this paper is to develop an automated procedure for audio restoration, avoiding the need for manual selection of noise-only sections and allowing for fully unsupervised broadband noise restoration of archive audio material. We propose an algorithm to estimate the PSD of stationary broadband noise disturbances, which is designed to work with diverse input signals, i.e., both speech and music signals. Assuming an exponential distribution for the noise periodogram coefficients, the noise PSD in each frequency band is estimated as the mean value of an exponential distribution which corresponds to the truncated periodogram coefficients of the disturbed input signal. The optimum truncation level is determined as the level that minimizes a distance measure between the empirical distribution of the truncated periodogram coefficients and the corresponding truncated exponential distribution. In addition, from this distance measure a frequency-dependent confidence value is computed that represents a measure for the reliability of the noise PSD estimate. This confidence value indicates whether the individual frequency bands contain broadband noise or not, and, hence should be processed by a broadband noise reduction algorithm or not.

5.1.2 *Related Work*

During the last decades the reduction of broadband noise has received steady research attention, mainly however for speech communication applications [72, 75, 77, 79]. Probably the earliest broadband noise reduction approach is described in a patent from 1965 [68]. Interestingly, many state-of-the-art broadband noise reduction algorithms are still based on a similar principle, namely splitting the noisy input signal into a number of frequency bands and attenuating the frequency bands with a low SNR. Since determining the frequency-dependent SNRs requires knowledge about the spectro-temporal characteristics of the broadband noise disturbance, a variety of noise PSD estimation algorithms have been proposed. Early algorithms used voice activity detection (VAD) [87, 88, 89] to determine noise-only sections, based on which the noise PSD was estimated by averaging short-time periodograms, e.g., using the Welch method [171]. Obviously, VAD-based noise PSD estimation algorithms perform poorly when no pauses of the desired signal are detected over a longer period. This holds especially for music signals where pauses are typically comparatively scarce. Furthermore, the VAD performance usually degrades at low SNRs, leading to an overestimation of the noise PSD as desired signal components are considered part of the noise [92]. Therefore, algorithms have been proposed that are able to estimate the noise PSD even when the desired signal is active. A well-known algorithm is the minimum statistics algorithm [91], which estimates the noise PSD by tracking minima of the noisy input PSD within a certain time window. A bias compensation factor compensates for the fact that the minimum is lower than the mean, which is the value of interest. Other algorithms estimate the noise PSD by recursively averaging the noisy input PSD using a time-varying recursive smoothing factor that depends on the probability of presence of the desired signal [92, 94].

It should be noted that all aforementioned noise PSD estimation algorithms require that the desired signal contains a number of pauses—either in the time domain (VAD-based algorithms) or in the time-frequency domain (algorithms based on minimum tracking or desired signal presence probability). While this is true for speech signals for which a high noise PSD estimation accuracy can be obtained, severe noise PSD overestimation can occur if the desired signal contains only very few pauses or does not contain pauses at all, e.g., for music signals (cf. Section 5.4.5.2).

Although most noise reduction algorithms are based on methods that were originally designed to enhance speech signals, they are often successfully applied to diverse audio recordings [38, 39, 83]. However, the problem of noise PSD estimation for signals different from speech has only been treated marginally. In [97] an automatic method estimate the noise PSD in music signals is proposed that simultaneously performs signal activity detection and noise PSD estimation based on dynamic Bayesian networks. It is shown that the proposed algorithm outperforms an earlier algorithm [170] that was designed for speech applications, however, the evaluation is restricted to comparatively low SNRs around 15 dB. Other recently proposed noise reduction algorithms, e.g., [172], eliminate the need for a noise PSD estimate by performing a sparse approximation of the noisy input signal and taking into account the time-frequency structure of audio signals. However, in order to obtain optimal restoration results, crucial parameters of the algorithm need to be adjusted in dependence on the characteristics and the SNR of the input signal [173].

5.1.3 *Paper Structure*

This paper is structured as follows: Section 5.2 describes the signal model and the assumed distribution of the periodogram coefficients of the noise disturbance. In Section 5.3 the proposed noise PSD estimation algorithm is presented in detail. The evaluation of the proposed algorithm with a large test signal database comprising speech and music signals and different types of broadband noise is presented in Section 5.4.

5.2 Signal Model

We assume that the broadband noise disturbance is additive, i.e.,

$$x[n] = s[n] + d[n] \text{ for } 0 \leq n < L, \tag{5.1}$$

with n denoting the sample index, L denoting the length of the signal, $x[n]$ the disturbed signal, $s[n]$ the clean (unobservable) audio signal and $d[n]$ the broadband noise disturbance. We assume that the noise is stationary over the complete duration of the recording and is uncorrelated with the audio signal. These assumptions are motivated by the targeted audio archive application, in which technical noise disturbances are caused by shortcomings of storage media or recording equipment, e.g., a recording that has been digitized from a single reel of tape.

In [174] it has been shown that the real and imaginary parts of the discrete Fourier transform (DFT) coefficients of stationary noise approximately follow a Gaussian distribution. Although this requires a sufficiently long DFT, it has been shown in [174] that the assumption of a Gaussian distribution already holds for a DFT length of $N = 1024$ samples. In the short-time Fourier transform (STFT) domain, the signal model in Eq. (5.1) is given by

$$X[k, l] = S[k, l] + D[k, l], \text{ for } 0 \leq k < N, 0 \leq l < M, \tag{5.2}$$

with $X[k, l]$, $S[k, l]$ and $D[k, l]$ the STFT coefficients of the time-domain signals $x[n]$, $s[n]$ and $d[n]$, respectively, k the frequency index, N the DFT length, l the block index and M the number of blocks. The STFT coefficients are obtained by computing the DFT for each (non-overlapping) block of the time-domain input signal, i.e.,

$$X[k, l] = \sum_{n=0}^{N-1} w[n] x[lN + n] \cdot e^{-j2\pi kn/N}, \tag{5.3}$$

where $w[n]$ is an analysis window function that is used to alleviate spectral leakage between neighboring frequency bins. The real and imaginary parts of the STFT coefficients $D[k, l]$ of the noise disturbance are assumed to be Gaussian distributed, i.e.,

$$\mathrm{Re}\{D[k,\,l]\},\,\mathrm{Im}\{D[k,\,l]\} \sim \mathcal{N}\big(0,\,\sigma^2[k]\big),\ \text{ for } 0 \le k < N,$$

with $\sigma^2[k]$ the variance of the Gaussian distribution in the kth frequency bin. Assuming that the real and imaginary parts are uncorrelated, which holds for large values of N [174], the squared magnitudes of the STFT coefficients $D[k,\,l]$, i.e., the short-time periodograms, follow an exponential distribution [175, pp. 259–260]:

$$P_d[k,\,l] = |D[k,\,l]|^2 \sim \mathrm{Exp}\big(\sigma^2[k]\big),\ \text{ for } 0 \le k < N,$$

with the exponential distribution Exp(•) defined via its probability density function (PDF):

$$f_\mathrm{e}(x;\,\mu) = \begin{cases} \frac{1}{\mu}e^{-\frac{x}{\mu}} & \text{for } x \ge 0 \\ 0 & \text{else} \end{cases}, \tag{5.4}$$

with mean $\mu > 0$.

For the clean audio signal s, we assume that for some time-frequency points its short-time periodogram coefficients

$$P_s[k,\,l] = |S[k,\,l]|^2$$

are zero (or at least much smaller than the corresponding noise periodogram coefficients $P_d[k,\,l]$). Simulation results in Section 5.4.5.1 show that only a small number of zero coefficients per frequency are required to obtain a very good noise PSD estimation accuracy.

5.3 Noise PSD Estimation Algorithm

5.3.1 *Overall Procedure*

The proposed noise PSD estimation algorithm makes use of the assumed stationarity of the noise disturbance over the complete duration of the recording[1] and the assumption that the clean audio signal is zero for some time-frequency points, corresponding to a number of time-frequency points where only noise is present. A confidence value is computed to indicate unreliable estimation results if too few noise-only time-frequency points are present.

---

[1] In our experiments we used signals with a length of 30 s.

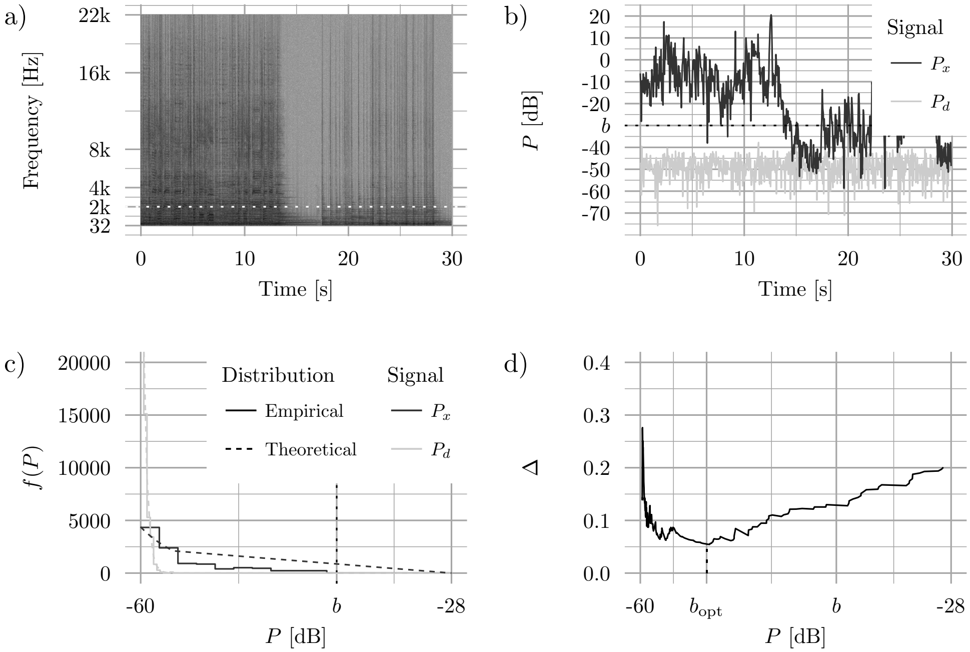


Figure 5.1: Intermediate steps of the proposed noise PSD estimation algorithm, exemplarily shown with an artificially disturbed music signal as input. The broadband SNR was set to 40 dB by adding white noise to a clean music recording. a) Spectrogram of the input signal x. The horizontal dotted line indicates a frequency of approximately 2 kHz. b) Power of the input signal (P_x) and the noise disturbance (P_d) at a frequency of approximately 2 kHz. The SNR at this frequency is approximately 47 dB. The horizontal dotted line indicates an exemplary truncation level $b = -30$ dB. c) Normalized histograms of the periodogram coefficients and PDFs of the assumed distributions. The vertical dotted line indicates an exemplary truncation level $b = -30$ dB. The y-axis has been limited to [0, 20000] to improve clarity. d) Distance measure Δ for different truncation levels b, indicating the optimal value b_{opt}.

Figure 5.1 shows four diagrams that illustrate the intermediate steps of the proposed algorithm. First, the short-time periodogram coefficients of the disturbed input signal

$$P_x[k,\, l] = |X[k,\, l]|^2$$

are computed. Subsequently, each frequency bin is analyzed separately and independently from all other frequency bins. Hence, in order to simplify the notation, from now on we will drop the frequency index k. For a music signal that has been artificially disturbed with white noise at a broadband SNR of 40 dB, Figure 5.1 a) depicts the spectrogram of the input signal. For an exemplary frequency of approximately 2 kHz, Figure 5.1 b) depicts the power of the input signal P_x and the (unobservable) power of the noise disturbance P_d.

The central idea of the proposed algorithm is to determine the power level for each frequency, below which the empirical distribution of the periodogram coefficients of the disturbed input signal is closest to the assumed distribution of the periodogram coefficients of the noise disturbance. The subset of the periodogram coefficients of the disturbed input signal that is smaller than or equal to a *truncation level* b is denoted as

$$\mathbf{P}_b = \{P_x[l] : 0 \leq l < M,\, P_x[l] \leq b\},$$

where we assume that the elements of $\mathbf{P}_b$ are sorted in ascending order. The size of this subset is denoted as Q, and the smallest truncation level b is selected such that $Q \geq M_{\min}$ (in this paper we use $M_{\min} = 10$). From this subset the normalized histogram $\mathbf{H}_b$ of the truncated periodogram coefficients is calculated, such that $\sum_{i=0}^{B-1} H_b[i] = 1$, with B the number of histogram bins (in this paper we use $B = 10$). The empirical cumulative distribution function (CDF) of the truncated periodogram coefficients F_b is given by [176]

$$F_b(x) = \frac{1}{Q} \sum_{i=0}^{Q-1} 1_{P_b[i] \leq x}, \tag{5.5}$$

with the indicator function

$$1_{P_b[i] \leq x} = \begin{cases} 1 & \text{for } P_b[i] \leq x \\ 0 & \text{else} \end{cases}.$$

As mentioned in Section 5.2, the periodogram coefficients of the noise disturbance are assumed to follow an exponential distribution, cf. Eq. (5.4). As a consequence, the truncated periodogram coefficients are assumed to follow a *truncated exponential distribution*, whose PDF f_{te} and CDF F_{te} are defined as

$$f_{\mathrm{te}}(x;\,\mu,\,b) \;=\; \begin{cases} \dfrac{\frac{1}{\mu}e^{-\frac{x}{\mu}}}{1-e^{-\frac{b}{\mu}}} & \text{for } 0 \le x \le b \\[2ex] 0 & \text{else} \end{cases} \tag{5.6}$$

$$F_{\mathrm{te}}(x;\,\mu,\,b) \;=\; \begin{cases} 0 & \text{for } x < 0 \\[1ex] \dfrac{e^{\frac{b}{\mu}}}{e^{\frac{b}{\mu}}-1}\left(1-e^{-\frac{x}{\mu}}\right) & \text{for } 0 \le x \le b \\[1ex] 1 & \text{for } x > b \end{cases} \;. \tag{5.7}$$

The corresponding normalized histogram is denoted as $\mathbf{H}_{\mathrm{te}}(\mu,\,b)$, using the same number of histogram bins as $\mathbf{H}_b$.

For each truncation level b, the optimal value $\hat{\mu}(b)$ of the truncated exponential distribution is then determined by minimizing the distance between the empirical distribution of the (truncated) periodogram coefficients and the assumed truncated exponential distribution. We have considered two different distance measures, cf. Section 5.3.2. The optimal truncation level leading to the minimum distance is denoted as b_{opt}. The corresponding parameter $\hat{\mu}(b_{\mathrm{opt}})$ of the truncated exponential distribution is used as the noise PSD estimate $\hat{\sigma}^2$.

For an exemplary value of the truncation level ($b = -30\,\mathrm{dB}$), Figure 5.1 c) depicts the normalized histogram of the truncated periodogram coefficients (black, solid line) together with the PDF of the exponential distribution using the optimal value $\hat{\mu}(b)$ estimated from $\mathbf{P}_b$ (black, dotted line). For reference, the normalized histogram and the PDF of the estimated exponential distribution of the truncated periodogram coefficients of the (unobservable) noise disturbance P_d are shown (gray, solid and dotted line, respectively). Figure 5.1 d) shows the distance measure (the normalized total absolute difference, cf. Section 5.3.2) as a function of the truncation level b, indicating the optimal value b_{opt}.

5.3.2 *Distance Measures between Probability Distributions*

A crucial part of the proposed algorithm is to determine how well the empirical distribution of the truncated periodogram coefficients of the disturbed input signal fits the assumed truncated exponential distribution. On the one hand, well-known statistical hypothesis tests for goodness-of-fit measures could be used, as they aim at determining whether a sample follows a specific probability distribution. Possible tests comprise, e.g., the Kolmogorov-Smirnov test [177], the Anderson-Darling test [178] or the chi-squared test [177]. On the other hand, distance measures between probability distributions, such as the Kullback-Leibler (KL) or Jensen-Shannon (JS) divergence [179], could be used.

In the context of our application, the measure should fulfill two properties: 1) independence of the sample size, which is related to the length of the input signal, and 2) boundedness in order to be able to derive a confidence value. As the behavior of statistical hypothesis tests depends on the sample size [180], they will not be further considered: for large sample sizes, the statistical evidence that the samples have been produced by an assumed distribution tends to zero, since small deviations from the assumed distribution become statistically significant. Hence, we will only consider distance measures between probability distributions. Since the KL divergence is not bounded [179], we will consider the JS divergence as a first option for an appropriate distance measure. The JS divergence between the normalized histograms $\mathbf{H}_b$ and $\mathbf{H}_{\text{te}}(\mu,\, b)$ is defined as

$$\Delta_{\text{JS}}(\mathbf{H}_b,\, \mathbf{H}_{\text{te}}) = \frac{1}{2}\left[\Delta_{\text{KL}}\left(\mathbf{H}_b,\, \frac{1}{2}(\mathbf{H}_b + \mathbf{H}_{\text{te}})\right) + \Delta_{\text{KL}}\left(\frac{1}{2}(\mathbf{H}_b + \mathbf{H}_{\text{te}}),\, \mathbf{H}_{\text{te}}\right)\right],$$

with the KL divergence between two histograms $\mathbf{H}_1$ and $\mathbf{H}_2$ defined as [179]

$$\Delta_{\text{KL}}(\mathbf{H}_1,\, \mathbf{H}_2) = \sum_{i=0}^{B-1} H_1[i] \cdot \log_2 \frac{H_1[i]}{H_2[i]}. \tag{5.8}$$

As the JS divergence is bounded by one ($0 \leq \Delta_{\text{JS}} \leq 1$) if the binary logarithm is used as in Eq. (5.8) [179], a confidence value can easily be derived as

$$C_{\text{JS}} = 1 - \Delta_{\text{JS}}. \tag{5.9}$$

As a second option, we propose to use the normalized total absolute difference (AD) between the empirical CDF in Eq. (5.5) and the CDF of the truncated exponential distribution in Eq. (5.7) as a simple and intuitive distance measure. The (unnormalized) total AD is given by

$$\text{AD} = \sum_{i=0}^{Q-1} |F_b(P_b[i]) - F_{\text{te}}(P_b[i];\, \mu,\, b)|,$$

where both CDFs are evaluated at the periodogram coefficient values $P_b[i]$. A loose upper bound for AD is given by

$$\text{AD}_{\text{max}} = \frac{Q}{2},$$

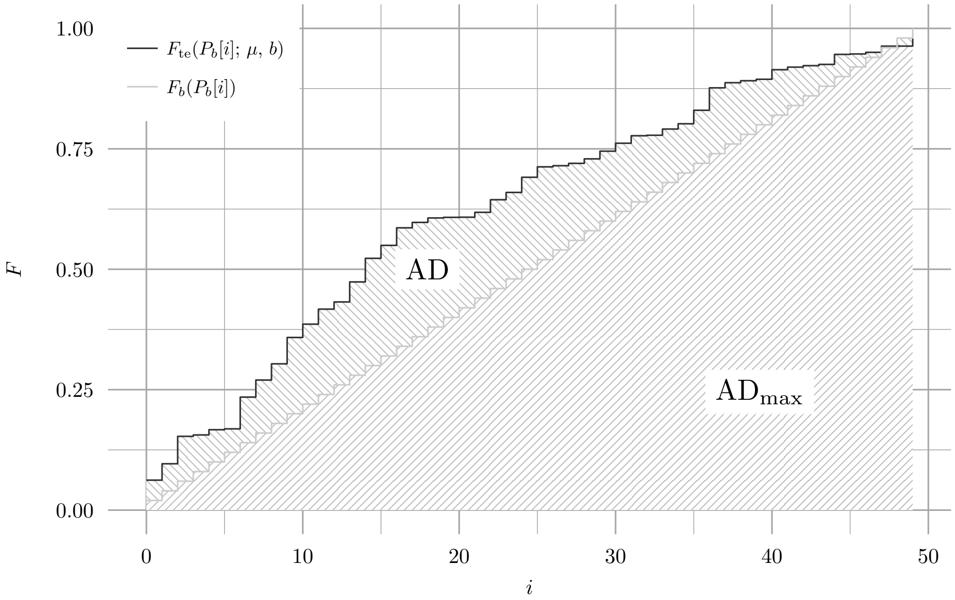


Figure 5.2: Absolute difference (AD) and $\mathrm{AD}_{\max}$ for exemplary values of $\mathbf{P}_b$, μ and b. In this example $Q = 50$.

which corresponds to $F_{\mathrm{te}}(P_b[i];\,\mu,\,b)$ being equal to either 0 or 1 for all i. Figure 5.2 shows AD and $\mathrm{AD}_{\max}$ for exemplary values of $\mathbf{P}_b$, μ and b, where AD is the area between $F_{\mathrm{te}}(P_b[i];\,\mu,\,b)$ and $F_b(P_b[i])$, and $\mathrm{AD}_{\max}$ is the area between 0 and $F_b(P_b[i])$. In this example, $F_{\mathrm{te}}(P_b[i];\,\mu,\,b) > F_b(P_b[i])$ for most i, which indicates that the periodogram coefficients are generally larger than assumed for this specific choice of μ and b.

Since AD is not bounded by one, we propose to normalize it by $\mathrm{AD}_{\max}$, i.e.,

$$\Delta_{\mathrm{AD}} = \frac{\mathrm{AD}}{\mathrm{AD}_{\max}},$$

such that similarly as in Eq. (5.9) a confidence value can be easily derived as

$$C_{\mathrm{AD}} = 1 - \Delta_{\mathrm{AD}}. \tag{5.10}$$

The confidence values in Eq. (5.9) and (5.10) are a measure for the reliability of the obtained noise PSD estimates. Although it will be shown in Section 5.4.5.2 that the proposed algorithm provides accurate PSD estimation results for a wide range of SNRs, the confidence values can be used to refrain from restoration when the confidence in the noise PSD estimate is too low. To this end, a minimum required confidence can be defined, and the noise PSD estimate is set to zero at frequencies where the confidence value is smaller than the minimum required confidence. This is especially relevant for signals without pauses or with high SNRs. In these cases, the proposed algorithm typically yields a noise PSD estimate which overestimates the true noise PSD, leading to signal distortion when used in a noise reduction algorithm. The complete noise PSD estimation algorithm is summarized in Algorithm 1.

Algorithm 1 The proposed noise PSD estimation algorithm. The MINIMIZEDISTANCE function determines the optimum value $\hat{\mu}(b)$ of the truncated exponential distribution as well as the distance Δ to that distribution.

```
procedure ESTIMATENOISEPSD($x$)                    ▷ Estimate the noise PSD in $x$
    for all frequency bins $k$ do
        $\Delta_{\min} \leftarrow \infty$                          ▷ Initialize the minimum distance
        for all block indices $l$ do       ▷ Compute the periodogram coefficients
            $P_x[l] \leftarrow \left|\sum_{n=0}^{N-1} w[n]x[lN+n] \cdot e^{-j2\pi kn/N}\right|^2$
        end for
        $\mathbf{P}_x \leftarrow$ SORT($\mathbf{P}_x$)
        for $l = M_{\min} \dots M$ do
            $\mathbf{P}_b \leftarrow [P_x[0], \dots, P_x[l-1]]$
            $b \leftarrow P_x[l-1]$                                  ▷ Truncation level
            $[\hat{\mu}, \Delta] \leftarrow$ MINIMIZEDISTANCE($\mathbf{P}_b$, $b$)
            if $\Delta < \Delta_{\min}$ then
                $\Delta_{\min} \leftarrow \Delta$
                $\hat{\sigma}^2[k] \leftarrow \hat{\mu}$
                $C[k] \leftarrow 1 - \Delta$                           ▷ Confidence value
            end if
        end for
    end for
    return $\hat{\boldsymbol{\sigma}}^2$, $\mathbf{C}$
end procedure
```

5.4 Evaluation & Results

We evaluate the proposed noise PSD estimation algorithm using a database of music and speech signals and different types of broadband noise (cf. Section 5.4.1). The noise PSD estimation accuracy is evaluated using several error measures (cf. Section 5.4.2). Furthermore, the perceptual quality of the restored audio signal is rated when the obtained noise PSD estimate is used in a high-quality noise reduction algorithm [77]. The evaluation results are presented in Section 5.4.5 and consist of three parts: First, we analyze the influence of the used distance measure on the estimation accuracy of the proposed algorithm. This analysis is based on a small test signal database in order to alleviate the computational requirements of the experiment. Second, for a large test signal database we compare the estimation accuracy using the optimum distance measure with a reference noise PSD estimation algorithm based on minimum statistics (cf. Section 5.4.3). Third, we evaluate the noise reduction performance when the obtained noise PSD estimate is used in a high-quality noise reduction algorithm and compare its performance to a recently proposed noise reduction algorithm that does not require a noise PSD estimate (cf. Section 5.4.4).

5.4.1 *Test Signals*

The used test signals are clean music and speech signals[2] to which we have added different types of broadband noise at different SNRs ranging from 20 dB to 60 dB. The clean signals are

- modern music recordings and
- high-quality speech recordings.

The noise disturbances are all technical in nature:

- artificially generated white noise,
- artificially generated pink noise,
- a recording of real tape noise [181] and
- a recording of real optical film soundtrack noise [182].

All clean signals and noise signals are single-channel[3] and sampled with $f_\mathrm{s} = 44.1\,\mathrm{kHz}$. The length of each recording is 30 s. For all experiments, we used a von-Hann window as the analysis window function for the computation of the short-time periodogram coefficients and a block length of $N = 2048$ samples such that the number of blocks is $M = 645$. The blocks do not overlap for the proposed algorithm[4] while an overlap of 50 % is used for the minimum statistics algorithm (cf. Section 5.4.3) as specified in [91].

5.4.2 *Performance Measures*

In order to evaluate the performance of the proposed noise PSD estimation algorithm, we use different instrumental measures that assess different properties of the algorithm.

---

[2] All signals, i.e., the clean signals and the noise signals, have either been published under a CC-BY license or are in the public domain and are available for download from the website accompanying this paper [112].

[3] The left channel was extracted if a recording had two channels.

[4] As no gain in estimation accuracy could be observed in informal experiments when using overlapping blocks, we use non-overlapping blocks to reduce the computational complexity of the algorithm.

5.4.2.1 *Noise PSD Estimation Errors*

To evaluate the PSD estimation accuracy of the proposed algorithm, we use the error measure proposed in [94], which equally weighs the logarithmic *overestimation* error (LogErrOver) and the logarithmic *underestimation* error (LogErrUnder), i.e.,

$$\text{LogErr} = \text{LogErrOver} + \text{LogErrUnder}, \tag{5.11}$$

with

$$\begin{aligned} \text{LogErrOver} &= \frac{1}{N_{\text{bins}}} \sum_{k=0}^{N_{\text{bins}}-1} \left| \min\left(0,\, 10 \cdot \log_{10}\left[\frac{\sigma^2[k]}{\hat{\sigma}^2[k]}\right]\right) \right| \\ \text{LogErrUnder} &= \frac{1}{N_{\text{bins}}} \sum_{k=0}^{N_{\text{bins}}-1} \left| \max\left(0,\, 10 \cdot \log_{10}\left[\frac{\sigma^2[k]}{\hat{\sigma}^2[k]}\right]\right) \right|, \end{aligned}$$

where the number of considered frequency bins $N_{\text{bins}} = \frac{N}{2} + 1$.

5.4.2.2 *Instrumental Measure for Audio Quality*

In order to evaluate the performance of a noise reduction algorithm using the obtained noise PSD estimate, we use an instrumental measure to rate the perceptual quality of the processed input signal. Specifically, we use the "perceptual evaluation of audio quality" (PEAQ) measure[5] [108, 109, 110]. This measure aims at determining the perceptual similarity between two audio signals by first computing a representation of each signal that takes the human hearing properties into account and then computing a similarity measure based on these representations. This similarity measure is called *Objective Difference Grade* (ODG) and ranges from -4 ("very annoying") to 0 ("imperceptible"), cf. Table 5.1. Although the PEAQ measure was devised to rate the perceptual quality of artifacts that were produced by audio coding algorithms, we believe that it also makes sense to use it to rate the quality of broadband noise restoration algorithms. This can be justified by the fact that additive noise disturbances were used during the development of the PEAQ measure (cf. [108]). In addition, in informal listening experiments we found that the ODG ratings obtained with the PEAQ measure generally correspond well with the subjective impression. The website accompanying this paper [112] makes the audio signals that were used for the evaluation available for listening, along with the corresponding ODG ratings.

5.4.3 *Reference Noise PSD Estimation Algorithm*

In order to assess the performance of the proposed noise PSD estimation algorithm, we use the well-known noise PSD estimation algorithm based on minimum statistics

---

[5] As the PEAQ measure requires its input signals to have a sampling rate of 48 kHz, the analyzed signals were resampled accordingly.

Table 5.1: The ODG scale.

ODG	Impairment Description
0	Imperceptible
-1	Perceptible but not annoying
-2	Slightly annoying
-3	Annoying
-4	Very annoying

[91] (cf. Section 5.1.2) for reference. An important parameter of this algorithm is the length of the minimum search window. If this window is too short to capture a pause in the desired signal, the minimum value within the window no longer corresponds to the noise power but contains a certain amount of desired signal power, resulting in an overestimation of the noise PSD. In [91] a compensation mechanism has been proposed that accounts for the estimation bias that is caused by using the power minimum while the goal is to estimate the mean power. It has been shown that the bias compensation factor becomes larger for longer minimum search windows. As a consequence, long minimum search windows may even lead to an increased overestimation if no pause in the desired signal is captured, caused by the bias compensation factor. In this paper, we will consider two different window lengths: the standard value of $\approx 1.5\,\mathrm{s}$ (typically used in speech communication applications) and the maximum value of $3.7\,\mathrm{s}$ specified in [91].

5.4.4 *Reference Noise Reduction Algorithms*

Since the main objective is audio restoration, we also evaluate the performance of the proposed noise PSD estimation algorithm (and the reference noise PSD estimation algorithm) in combination with the frequently used minimum mean square error short-time spectral attenuation (MMSE STSA) noise reduction algorithm [77]. We use the implementation and parameter values from [62]. In addition, we use a recently proposed noise reduction algorithm based on structured sparsity [172], which takes the time-frequency structure of the input signal into account and does not require a noise PSD estimate. We use the default parameters but reduced the threshold level (λ) to 0.001 as this value allows for good restoration results for a wide range of SNRs. The two noise reduction algorithms will be denoted by *MMSE STSA* and *Struc. sparsity*, respectively.

In order to reduce artifacts produced by the noise reduction algorithms, i.e., musical noise and degradation of the desired signal, we restrict the maximum attenuation of each time-frequency coefficient, i.e., we set the spectral floor [79], to $-20\,\mathrm{dB}$ in all experiments.

5.4.5 *Results*

This section presents the results of three experiments to determine the optimum distance measure (Section 5.4.5.1), the noise PSD estimation accuracy with this optimum distance measure (Section 5.4.5.2) and the noise reduction performance when using the proposed noise PSD estimate in the MMSE STSA noise reduction algorithm (Section 5.4.5.3). In Section 5.4.5.1 we use 50 music and 50 speech recordings, while in Sections 5.4.5.2 and 5.4.5.3 we use 500 music and 500 speech recordings.

5.4.5.1 *Distance Measure*

In this section we analyze the noise PSD estimation accuracy of the proposed algorithm for both considered distance measures (cf. Section 5.3.2), i.e., the *JS divergence* and the *normalized total AD*. For both distance measures, Figure 5.3 shows the LogErr measure in Eq. (5.11) for different SNRs. The box plots represent the distribution of the LogErr measure for all combinations of the 100 clean music and speech signals and the four noise types (cf. Section 5.4.1). It can be observed that the LogErrs are very similar for both distance measures.

For all of the following experiments we selected the normalized total AD as the distance measure as it leads to a good overall estimation accuracy for all SNRs, and it is easier to compute than the JS divergence.

5.4.5.2 *Noise PSD Estimation Accuracy*

Using the optimum distance measure determined in the previous section, in this section we analyze the noise PSD estimation accuracy in more detail and compare it to the reference noise PSD estimation algorithm based on minimum statistics (cf. Section 5.4.3).

Figure 5.4 shows the noise PSD estimation errors of the proposed algorithm and the reference minimum statistics algorithm (for two search window lengths) for different SNRs and noise types. First, it can be observed that the estimation errors for all algorithms depend strongly on the input SNR, i.e., the larger the input SNR, the larger the estimation errors. Furthermore, increasing the length of the search window for the minimum statistics algorithm leads to lower estimation errors. This is due to the fact that a longer search window increases the probability of capturing parts of the desired signal with pauses inside the search window, hence reducing the noise PSD overestimation error. For most SNRs and noise types, the proposed algorithm yields lower estimation errors than the minimum statistics algorithm. We therefore conclude that the assumption of exponentially distributed

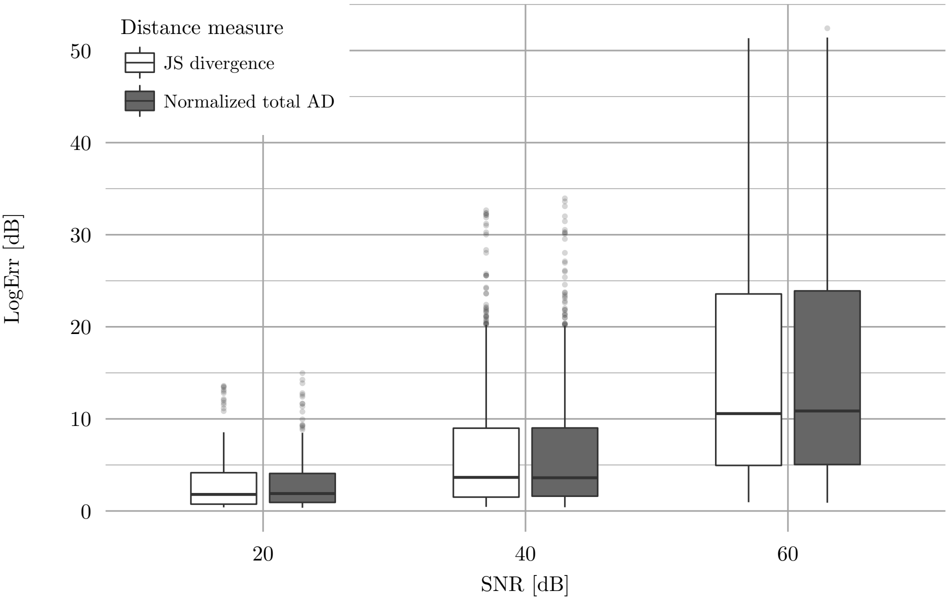


Figure 5.3: Noise PSD estimation errors of the proposed algorithm for different SNRs for both distance measures. The lower and upper edge of each box indicates the first and the third quartile of the data, respectively, while the horizontal line in each box corresponds to the median value. Vertical lines extending from the boxes extend to the smallest and largest data point, respectively, within 1.5 times the inter-quartile range (IQR), with IQR the distance between the first and the third quartile. All data outside these intervals are considered outliers and are represented by dots.

periodogram coefficients is valid not only for artificially generated noise but also for real-world noise recordings. Only for an SNR of 20 dB and film noise, the minimum statistics algorithm yields a smaller error than the proposed algorithm. This can probably be explained by the fact that the assumption regarding the distribution of the noise periodogram coefficients is violated: in addition to broadband noise, the used film noise also contains a certain amount of hum and impulsive disturbances. The proposed algorithm only estimates the PSD of the stationary noise part of the disturbance, while the PSD estimate obtained using the minimum statistics algorithm includes the PSD of the hum. As a consequence, the proposed algorithm underestimates the noise PSD, leading to an increased LogErr. This specific result indicates that it is important to detect and remove hum and impulsive disturbances before noise reduction (e.g., [113, 114]).

It should be noted that the confidence value C_{AD} was not taken into account in this experiment, i.e., for each frequency the noise PSD is determined from the truncation level that leads to the maximum confidence (corresponding to the smallest distance between the empirical distribution of the truncated periodogram coefficients and the truncated exponential distribution, cf. Eq. (5.10)), however low that confidence is.

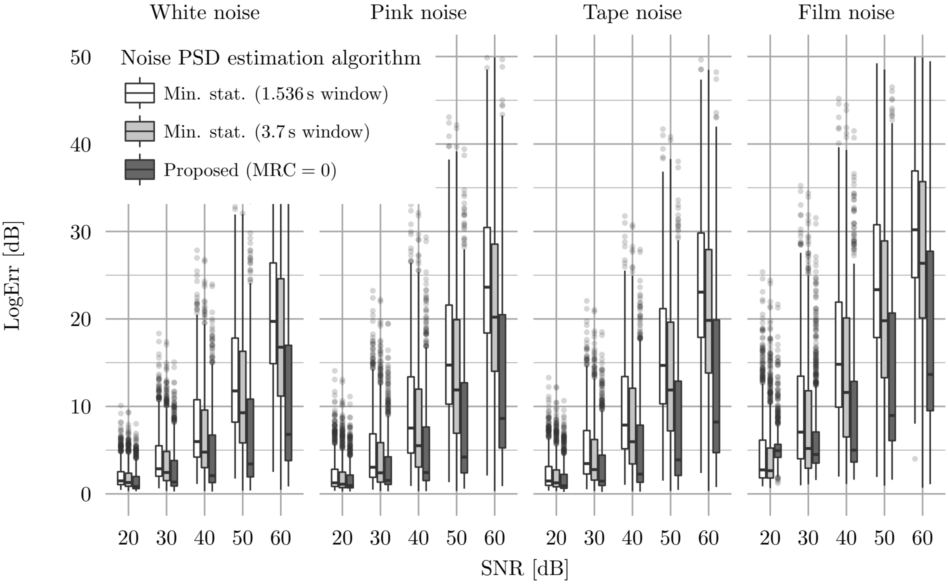


Figure 5.4: Noise PSD estimation errors of the proposed algorithm using the normalized total AD distance measure and the minimum statistics algorithm for two search window lengths. This figure integrates the results for all speech and music signals. The results are separately shown for each noise type.

Figure 5.5 shows the confidence values for each noise type and SNR, averaged over all frequencies for each disturbed signal. It can be observed that the confidence values depend on the input SNR, i.e., the larger the input SNR the smaller the confidence. The confidence values are similar for white noise, pink noise and tape noise, while they are generally lower for film noise. Similarly, as for the noise PSD estimation errors in Figure 5.4, this can probably be explained by a certain amount of hum and impulsive disturbances in the film noise. The results in Figure 5.5 indicate that the confidence value allows to distinguish severely disturbed, i.e., SNR ≤ 30 dB, from weakly disturbed input signals, i.e., SNR ≥ 50 dB, in most cases.

5.4.5.3 *Noise Reduction*

This section presents the results of the proposed noise PSD estimation algorithm combined with the MMSE STSA noise reduction algorithm in terms of perceptual audio quality. First, the influence of using a minimum required confidence (MRC), cf. Section 5.3.2, on the perceptual audio quality is investigated. Second, the performance of the MMSE STSA noise reduction algorithm using the proposed and the reference noise PSD estimate is compared to a noise reduction algorithm based on structured sparsity.

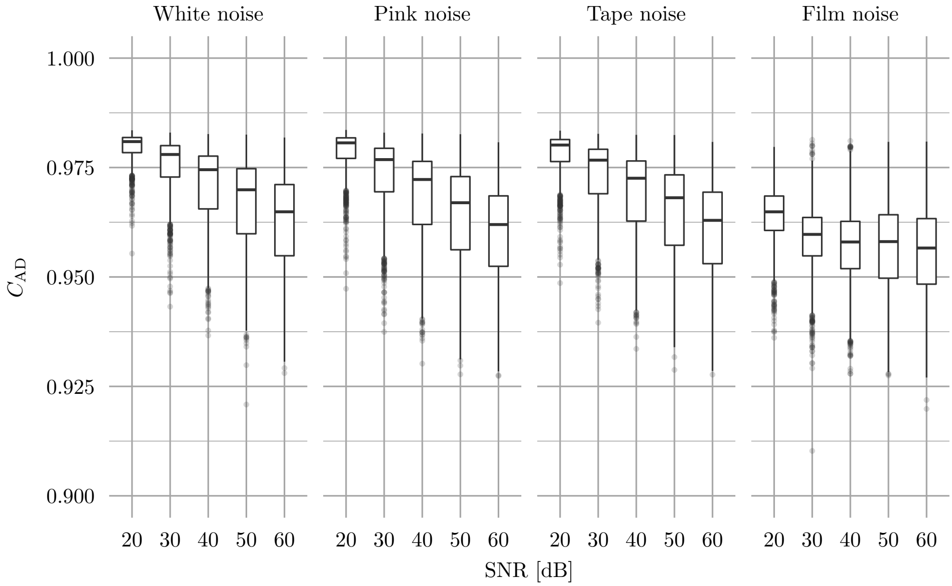


Figure 5.5: Confidence values of the proposed algorithm, averaged over all frequencies for each disturbed signal. This figure integrates the results for all speech and music signals. The results are separately shown for each noise type.

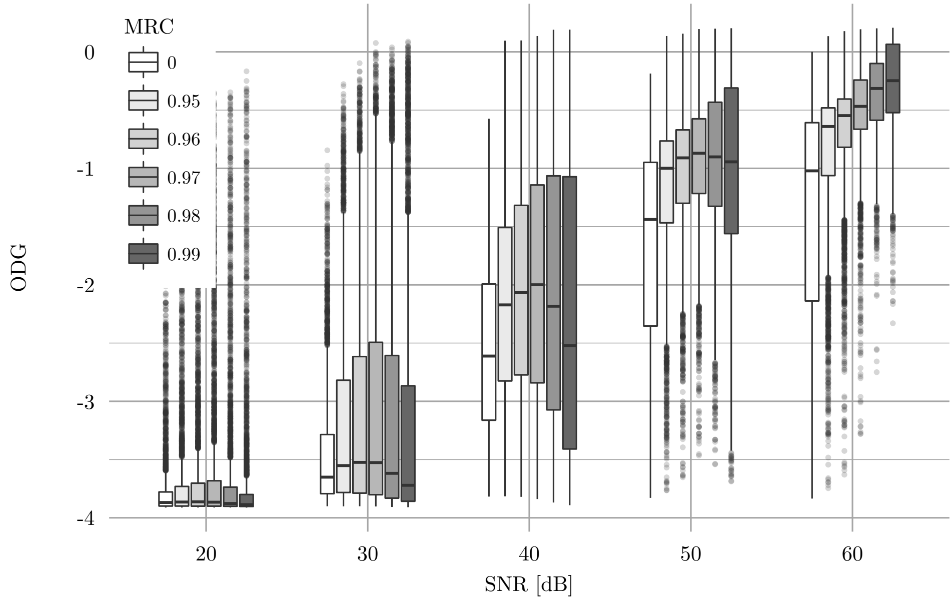


Figure 5.6: Instrumental audio quality evaluation of the processed signals obtained by combining the proposed noise PSD estimate with the MMSE STSA noise reduction algorithm for different MRC values. This figure integrates the results for all speech and music signals and all noise types.

Figure 5.6 shows the PEAQ ratings of the processed signals for different SNRs and for different values of the MRC: if the confidence value C_{AD} is lower than the MRC for a certain frequency, the noise PSD estimate is set to zero, i.e., no noise reduction is performed at this frequency. Using MRC = 0 corresponds to accepting all noise PSD estimates, however low the confidence value is. While MRC = 0 leads to the maximum amount of noise reduction, it can be expected that this will lead to a degradation of the desired signal in frequency bands with a high SNR or no pauses in the clean signal (both leading to an overestimation of the noise PSD). In contrast, MRC = 1 leads to no noise reduction because the empirical CDF of the periodogram coefficients always deviates from the theoretical CDF, at least by a small amount, and the confidence never reaches exactly 1. Hence, the MRC can be used as a trade-off between maximum noise reduction and maximum preservation of the desired signal. The optimal value of the MRC depends on the audio restoration task at hand. From Figure 5.6 it can be observed that the achieved restoration quality depends on the MRC, and the MRC for which the best PEAQ rating is achieved highly depends on the SNR. Especially for high SNRs, using MRC > 0 is important to protect the clean signal. While all considered MRC values yield similar median PEAQ ratings for an SNR of 20 dB, MRC = 0.97 yields the best median PEAQ rating for SNRs of 30 dB to 50 dB, and MRC = 0.99 yields the best median PEAQ rating for an SNR of 60 dB. As MRC = 0.98 yields PEAQ ratings that lie between those obtained with MRC = 0.97 and MRC = 0.99, it represents a trade-off between high noise reduction at low SNRs and preservation of the clean signal at high SNRs. Hence, in the following experiment we will consider two values for the MRC that work well for all SNRs, namely MRC $\in \{0.97, 0.98\}$.

Figure 5.7 shows the PEAQ ratings of the unprocessed disturbed input signals ("None") and of the signals processed by the MMSE STSA noise reduction algorithm using the proposed noise PSD estimate (for two MRC values), the minimum-statistics-based noise PSD estimate (for two search window lengths) and the oracle noise PSD estimate. In addition, the PEAQ results of a noise reduction algorithm based on structured sparsity (cf. Section 3.3) are shown. It can be observed that a broadband noise disturbance may lead to a severe degradation of the overall audio quality ("None"). The rightmost boxes ("Oracle") indicate that the MMSE STSA algorithm using the true noise PSD is able to increase the PEAQ rating for all SNRs except for SNR = 60 dB. As the true noise PSD is obviously unknown in practice, these results can only serve as a reference. The results obtained with the minimum statistics algorithm (for both search window lengths), show that an improvement of the PEAQ rating is only achieved for an SNR of 20 dB. For SNR > 30 dB the audio quality is severely reduced, due to a degradation of the desired signal as the noise PSD is overestimated for these SNRs. As music signals usually do not contain as many pauses as speech signals, the quality degradation is especially severe for music signals. The choice of a longer search window alleviates this problem to a certain extent—the median PEAQ ratings for the minimum statistics algorithm with a search window length of 3.7 s are higher than those with a search window length of 1.536 s. Furthermore, the noise reduction algorithm based on structured sparsity achieves an improvement of the PEAQ rating compared to the unprocessed signal for SNRs of 20 dB to 40 dB. For SNR > 40 dB, the application of this algorithm also leads to a considerable decrease in audio quality. This is caused by a degradation of the desired signal which is presumably the result of fixed algorithm parameters that are not adjusted in dependence on the SNR.

For SNRs of 20 dB and 30 dB the PEAQ ratings obtained by the proposed algorithm are comparable to those obtained by the structured sparsity algorithm. For an SNR of 40 dB, the PEAQ ratings obtained by the proposed algorithm are a bit worse than those obtained by the structured sparsity algorithm, but better than the unprocessed input signal. For SNRs of 50 dB and 60 dB, the PEAQ ratings obtained by the proposed algorithm with MRC = 0.98 are much higher than those obtained by all other considered algorithms, reaching quality ratings close to those of the unprocessed input signal.

In conclusion, the proposed algorithm with MRC = 0.98 is the only algorithm which improves the audio quality of noisy signals over the wide range of considered SNRs and input signals typically encountered in audio archives, while only leading to a small amount of signal degradation for practically clean input signals.

For most SNRs, the difference between the PEAQ ratings of the unprocessed signal and the processed signal using the proposed algorithm (MRC = 0.98) are, in general, rather small. Nevertheless, an informal subjective evaluation of the processed signals indicates that the application of the proposed noise reduction algorithm in many cases leads to a substantial increase in audio quality, also in cases in which this is not indicated by the PEAQ rating. However, the informal subjective evaluation also indicates that the overall quality rating trend obtained with PEAQ corresponds well with the subjective impression—especially considering relative quality differences between different algorithms. A selection of the audio files can be found on the website accompanying this paper [112].

5.5 Summary & Conclusions

In this paper we presented a novel algorithm to estimate the PSD of stationary broadband noise disturbances in audio signals. The proposed algorithm assumes that the noise periodogram coefficients are exponentially distributed and estimates the noise PSD as the mean value of an exponential distribution which corresponds to the truncated periodogram coefficients of the disturbed input signal. In addition, a confidence value is computed reflecting the reliability of the noise PSD estimate. This confidence value is used to reject noise PSD estimates with a low confidence in order to avoid degradation of the desired signal when the obtained noise PSD estimate is used in a noise reduction algorithm. Based on experiments with a large database of clean speech and music signals and different artificial and real-world broadband noise disturbances, we have shown that the proposed algorithm yields reduced PSD estimation errors compared to the state-of-the-art minimum statistics algorithm for a large range of SNRs. When using the proposed noise PSD estimate in the MMSE STSA noise reduction algorithm with an MRC of 0.98, we showed that an unsupervised restoration is possible for a large variety of test signals at

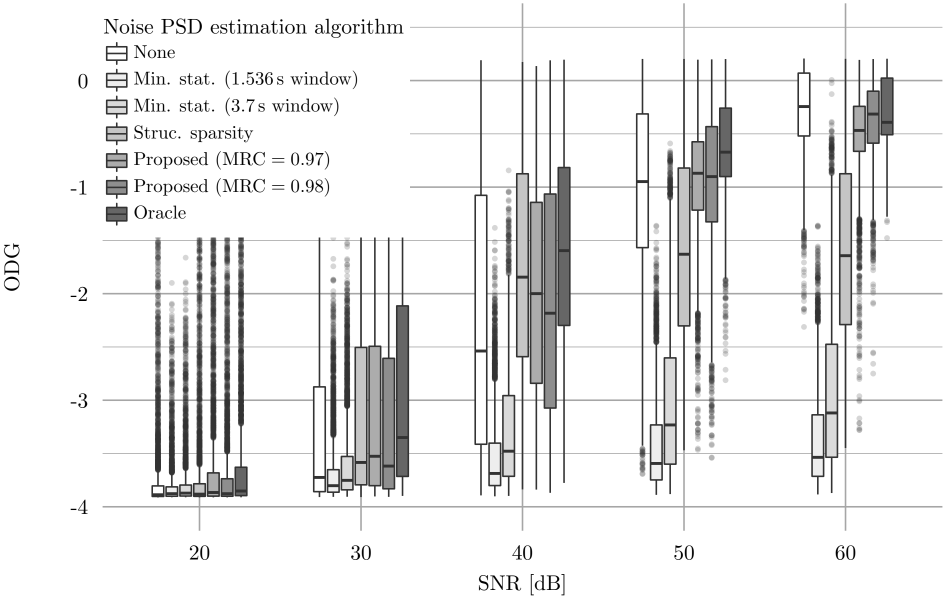


Figure 5.7: Instrumental audio quality evaluation of the processed signals obtained by several algorithms: 1) no processing, 2) MMSE STSA noise reduction algorithm using the minimum-statistics-based noise PSD estimate, the proposed PSD estimate and the oracle noise PSD estimate, 3) noise reduction algorithm based on structured sparsity.

a wide range of SNRs, leading to a median PEAQ improvement for SNRs below 60 dB and very little signal degradation at an SNR of 60 dB. In contrast, restoration results with minimum-statistics-based noise PSD estimates and a noise reduction algorithm based on structured sparsity lead to a severe decrease in PEAQ rating for SNRs above 30 dB and 40 dB, respectively. In conclusion, the presented algorithm constitutes a crucial step for automatic broadband noise restoration over a wide range of SNRs and input signals, which are typically encountered in large audio archives.

6

SUMMARY, CONCLUSIONS AND FURTHER RESEARCH

In this chapter we summarize the main contributions of this thesis in Section 6.1 and give suggestions for further research in Section 6.2.

6.1 Summary and Conclusions

Most existing solutions for the high-quality restoration of audio material have been designed for supervised operation and usually require the manual adjustment of one or more parameters for each individual recording. Due to the large number of recordings stored in media archives, a supervised restoration is typically not feasible. As the diversity of recordings stored in archives is generally very high, regarding the desired signals, the disturbance types and their intensities, the unsupervised application of existing audio restoration algorithms may lead to a severe quality degradation, especially for undisturbed signals.

The main objective of this thesis was to develop algorithms that allow for an unsupervised restoration of large numbers of very diverse audio recordings, eliminating the need for manual parameter adjustment for each recording. Key elements in the design of these algorithms were to achieve robustness against a high variety of input signals with regard to the type of the desired signal and the intensity of the disturbance on the one hand and the desire to keep the risk of signal degradations as low as possible on the other hand. Typical audio restoration algorithms comprise two stages: the *estimation of disturbance parameters* and the actual *disturbance reduction.* To automate the restoration process, it has turned out that the estimation of disturbance parameters is crucial, including the detection of signal portions (or complete signals) which are disturbance free. To increase the robustness against highly diverse audio material, in this thesis we therefore proposed novel algorithms for the classification of impulsive disturbances, for the detection and parameter estimation of hum disturbances and for the PSD estimation of broadband noise.

The restoration of impulsive disturbances greatly depends on properly adjusting a threshold parameter for each individual signal. Inadequate parameter values either lead to only minor disturbance reduction or a degradation of the desired signal, e.g., caused by the reduction of transient elements in the desired signal. These transient elements are, e.g., certain drum sounds or attacks of brass instruments. Unsupervised application of existing impulse restoration algorithms therefore may lead to unacceptable sound quality of the processed signal, especially for high input SNRs and undisturbed signals. The classification algorithm proposed in Chapter 2 allows to classify frames of the input signal as either clean or disturbed. The proposed classification algorithm is based on a supervised learning approach, using a logistic regression model and selected features of the appropriately prewhitened input signal. The model parameters were trained using a large number of test signals that contain artificial but plausible impulsive disturbances. The proposed classification algorithm benefits from using comparatively long frames of 1 s length, compared to detection stages of typical impulse restoration algorithms that work on the sample-by-sample level. In doing so, the aim of the proposed algorithm is not to replace detection stages of impulse restoration algorithms but rather to determine which signal frames actually contain impulsive disturbances and, thus, should be processed with one of the existing impulse restoration algorithms. We presented evaluation results for a large number of test signals and a large range of SNRs, using the PEAQ algorithm to rate the perceptual quality, that indicate on the one hand that a state-of-the-art AR-model-based impulse restoration algorithm improves the perceptual quality of signals that contain perceivable impulsive disturbances at SNRs below 40 dB. On the other hand, the results indicate that the application of this impulse restoration algorithm leads to a degradation of the signal quality for SNRs larger than 40 dB. This was explained by false alarms in the algorithm's detection stage and subsequent reduction of transient elements in the desired signal. Furthermore, the evaluation results indicate that the application of the proposed impulsive disturbance classification algorithm to determine those frames of a recording that actually contain impulsive disturbances, and subsequent impulse restoration of only those frames leads to an overall increase in restoration quality, especially for large SNRs and undisturbed signals.

Existing algorithms for the detection of hum disturbances in recordings make assumptions regarding the input signal that impede their applicability for archive audio restoration applications. These algorithms have either been designed for speech signals and exploit pauses in the input signal to estimate the hum disturbance parameters or it is assumed that the input signal always contains a hum disturbance. If a hum restoration algorithm is applied to a hum-free recording, a degradation of the signal quality can be expected. Therefore, in Chapter 3 we presented a hum detection algorithm that allows to determine the presence of hum in a recording and to estimate all parameters that are required to remove the disturbance with one of the existing hum reduction algorithms. The presented algorithm is based on a quantile-based statistical analysis of the short-time PSD estimates of the input signal in order to detect the presence of steady hum tones. It is therefore assumed that the power of the hum tones changes less than the desired signal power at the hum frequencies. A measure for the power fluctuation is computed as the ratio of a low and a high quantile of the power in each frequency bin. A post-processing step increases the detection accuracy by discarding short detections and allowing for hum tones to drop out for a certain duration. As the short-time PSD estimates are calculated via the DFT, the frequency resolution of the detection is limited. In order to increase the frequency estimation accuracy, notch filters are placed at the detected hum frequencies and an accurate hum frequency estimate is determined from the filter coefficients after convergence. We presented an evaluation with different types of desired signals and artificial hum disturbances that shows that the optimum low and high quantiles, leading to the minimum number of missed detections and false alarms, are the 10 % and 55 % quantiles. In addition, using these optimum quantiles, we presented evaluation results based on artificially disturbed test signals that indicate that more than 90 % of all hum disturbances with an ODG of -2 (corresponding to “slightly annoying” disturbances) or less were correctly detected as disturbed while none of the hum-free signals were erroneously detected as disturbed. Furthermore, the absolute hum frequency estimation error in most cases was below 0.1 Hz, allowing for an efficient hum restoration. Finally, an evaluation with 24 h of real-world signals showed that approximately 65 % of the disturbing hum disturbances and approximately 26 % of the just audible hum disturbances were detected with a false alarm rate of below 5 % percent.

Following the description of the hum detection algorithm, Chapter 4 contained an overview and a comparison of different hum reduction algorithms. We reviewed well-known filters that are often used for hum reduction applications, namely comb filters, subband comb filters and notch filters. While comb filters allow for an efficient reduction of hum tones by placing notches at integer multiples of the fundamental frequency of the hum tone complex up to the Nyquist frequency, the processing often introduces artifacts which may be perceived as an echo effect. This is due to the fact that comb filtering is achieved by adding a delayed, and possibly scaled, version of the signal to itself. While the notch depth can be adjusted by scaling the delayed version of the signal, we restricted our analysis to a scaling factor of one, leading to periodic nulls in the frequency response. Subband comb filters take the low-pass character of typical hum disturbances into account and reduce the amount of artifacts by only processing the frequency range that actually contains hum tones. This is achieved with a bandsplitting algorithm to split the input signal into a low and a high frequency band, applying a comb filter to the low frequency band, and combining both bands again to obtain the output signal. Notch filters offer the largest flexibility of the three algorithms by allowing to place notches on the individual hum tones. We presented an evaluation with different types of desired signals and artificial and real hum disturbances that showed that comb filters lead to the largest amount of hum reduction of the three analyzed algorithms of approximately 33 dB, but also to the largest amount of degradation of the desired signal of approximately 3 dB. Compared to comb filters, subband comb filters lead to a lower hum reduction of approximately 18 dB and a lower degradation of the desired signal of 1.5 dB. Finally, the evaluation showed that a large amount of hum reduction of up to 40 dB is possible by using notch filters, with only minor degradation of the desired signal of -1.2 dB if the frequencies of the hum partial tones are known. This is probably due to the fact that unnecessary notches at integer multiples of the hum fundamental frequency are avoided, leading reduced degradation of the desired signal.

It was shown in Chapter 5 that a state-of-the-art noise PSD estimation algorithm based on minimum statistics yields significant estimation errors if it is applied to diverse audio material in an unsupervised manner, especially for SNRs larger than 40 dB. This is mainly caused by the fact that this noise PSD estimation algorithm has been developed for speech signals, based on the assumption of frequent pauses in the desired signal. As music signals typically only contain scarce pauses or no pauses at all, this leads to an overestimation of the noise PSD. We presented an evaluation, using a large number of artificially disturbed test signals and the PEAQ algorithm to rate the perceptual quality of the restored signals, that shows that noise PSD estimation errors may result in a severe quality degradation when the noise PSD estimation algorithm is combined with the state-of-the-art MMSE STSA noise reduction algorithm. The algorithm proposed in Chapter 5 allows for a robust estimation of the noise PSD for a large diversity of input signals and SNRs. It is based on the assumption that the noise PSD is constant in a recording, which holds for many archive recordings that have been digitized from a single carrier. The noise PSD is estimated as the mean value of an exponential distribution that corresponds to the truncated short-time periodogram coefficients of the input signal. The optimum truncation level is determined as the one that minimizes a distance measure between the empirical distribution of the truncated periodogram coefficients and the corresponding truncated exponential distribution. The evaluation results indicated that the proposed algorithm achieves significantly lower noise PSD estimation errors than the algorithm based on minimum statistics under most conditions, for SNRs of 20 dB to 60 dB, and for different types of broadband noise, i.e., artificial white and pink noise and real tape and film noise. In addition, the proposed algorithm provides a confidence measure that indicates the reliability of the PSD estimate. This measure was used to reject noise PSD estimates with a low confidence to avoid a degradation of the desired signal if the proposed noise PSD estimation algorithm is used in combination with a broadband noise reduction algorithm. In order to do so, we defined a minimum required confidence (MRC) and the noise PSD estimate was set to zero at frequencies where the confidence was smaller than the MRC. While the presented simulation showed that a noise PSD estimation algorithm based on minimum statistics, in combination with the MMSE STSA broadband noise reduction algorithm, can increase the perceptual quality of severely disturbed signals at an SNR of 20 dB, the quality is generally reduced for SNRs of 30 dB to 60 dB. The evaluation results for an algorithm based on structured sparsity that does not require a noise PSD estimate showed that the quality is increased for SNRs of 20 dB to 40 dB, but that the quality is reduced for SNRs above 40 dB. The evaluation results showed that the combination of the proposed noise PSD estimation algorithm with the MMSE STSA broadband noise reduction algorithm yields output signals with an improved perceptual quality for SNRs of 20 dB to 60 dB, increasing the quality for SNRs of 20 dB to 50 dB and leading to very little quality degradation at an SNR of 60 dB.

Although the algorithms presented in this thesis represent an important step towards the automatic restoration of highly diverse audio material, they are still not perfect. The evaluation results showed that the proposed algorithms significantly improve the quality for the majority of signals compared to the unsupervised operation of state-of-the-art restoration algorithms. Nevertheless, the interaction with a skilled professional is still required for the most demanding audio restoration projects of particularly valuable recordings, and care has to be taken if whole media archives are processed in an unsupervised fashion. Still, the impulse probability of the proposed impulse classification algorithm and the confidence measures of the proposed broadband noise PSD estimation algorithm can be used to trigger a manual inspection of signals that yield unclear classification and estimation results.

6.2 Suggestions for Further Research

In this section we give some suggestions for further research—regarding the individual restoration of the three considered disturbance types in Sections 6.2.1 to 6.2.3 and regarding more general topics in Sections 6.2.4 and 6.2.5. As the three considered disturbances (impulsive disturbances, hum and broadband noise) differ fundamentally regarding the signal model and the algorithms to determine the disturbance parameters and to reduce the disturbances, possibilities for further research depend on the disturbance.

6.2.1 *Restoration of Impulsive Disturbances*

The impulsive disturbance classification algorithm presented in Chapter 2 achieves high classification accuracy for the majority of signals. However, informal experiments have shown that certain transient elements of the desired signal lead to an erroneous classification. In many cases, these transient elements are attack sounds of drums or brass instruments or picked guitars which appear rhythmically, i.e., in a regular pattern. This information can be used to increase the robustness of the classification algorithm. One approach is to perform beat detection in a preprocessing step, possibly based on an existing beat tracking algorithm [183, 184]. Short signal portions around the estimated beat positions can then be excluded from the feature computation and, hence, will probably not influence the classification result. Another approach is to inspect the positions of the potential impulsive disturbances by analyzing the result of the detection stage of an impulse restoration algorithm. Potential impulses appearing in a regular pattern can then be assumed to be related to the desired signal, whereas impulses that appear unregularly can be assumed to be related to the disturbance.

6.2.2 *Restoration of Hum Disturbances*

The results presented in Chapter 3 indicate that detecting hum disturbances and estimating their parameters is possible with high accuracy. Furthermore, the results in Chapter 4 indicate that a large amount hum reduction is possible with low signal distortion if the frequencies of the hum harmonics are known or well estimated. Therefore, the combination of the proposed hum detection algorithm with a notch filter-based hum reduction algorithm is expected to allow for an unsupervised operation, resulting in high-quality hum restoration for most signals. Nevertheless, the results also indicate that the hum detection algorithm misses a certain amount of hum disturbances if their power is low compared to the desired signal. As many hum disturbances are harmonic tone complexes with multiple partial tones, an increased detection accuracy may be obtained by integrating all partial tones to determine a single detection result instead of analyzing each partial tone individually.

A different approach to reduce hum disturbances could be based on *cancelling* the tone complex that represents the hum disturbance. Obviously, this requires very precise knowledge about the hum disturbance itself, with large parameter estimation errors potentially having catastrophic consequences, i.e., possibly even increasing the hum power. While a cancellation approach possibly does not lend itself to a fully unsupervised restoration, it may be advantageous if manual operation and parameter adjustment is feasible. This is because the cancellation approach in principle allows for a reduction of hum signal power without any degradation of the desired signal. In contrast, state-of-the-art filter-based algorithms as summarized in Chapter 4 always represent a compromise between hum power reduction and the degradation of the desired signal.

6.2.3 *Broadband Noise*

The proposed noise PSD estimation algorithm in Chapter 5 relies on the assumption that the periodogram coefficients of the broadband noise follow an exponential distribution. Furthermore, it is assumed that the spectrogram of the desired signal exhibits a small number of signal-free time-frequency coefficients, such that the smallest short-time periodogram coefficients of the noisy input signal for each frequency are assumed to be related to the noise disturbance only. The presented simulation results indicate that this assumption may be assumed to be valid, as accurate noise PSD estimation results are obtained with the proposed algorithm. The basic idea of determining a *subset* of the short-time periodogram coefficients to estimate the noise PSD can, however, be extended. For example, a different distribution of the periodogram coefficients of the disturbance can be assumed. This may lead to improved results, e.g., for certain types of film noise that can coarsely be described as a combination of continuous and impulsive noise. Due to outliers caused by impulses, in this case the periodogram coefficients related to the disturbance are no longer concentrated at the smallest periodogram coefficients. The challenge is then to come up with an efficient means to determine the subset of periodogram coefficients that minimizes the distance to the assumed distribution. Approaches to solve this problem could be based on efficient algorithms for discrete optimization, e.g., Branch-and-Bound [185]. In order to do so, however, an efficient way to compute a lower bound for the minimum distance is required to prune the search space of possible subsets. Another extension of the proposed algorithm idea could be to incorporate *cyclostationarity* of the broadband noise into the signal model. Cyclostationary noise may occur, e.g., with cylinder or disc media if dust and dirt is distributed unevenly across the carrier. The resulting broadband noise is then modulated depending on the playback position. Due to the geometrical shape of these carriers, the modulation is typically periodic and the noise can be described as a cyclostationary process.

The proposed noise PSD estimation algorithm assumes that the noise PSD is constant over the complete duration of a recording. This is the case for many signals whose broadband noise is caused by insufficiencies of the original carrier material and, therefore, can be assumed to be rather stationary for the complete recording. Nevertheless, the range of application of the proposed algorithm can be extended to time-varying broadband noise characteristics that may occur, e.g., if a recording contains a sequence of clips from different original media, or simply at different levels. In order to do so, information about sections with stationary broadband noise is required. Therefore, an interesting research topic is to develop an algorithm to detect these sections. Informal experiments based on analyzing high-frequency information above the maximum frequency of the desired signal have yielded promising results.

6.2.4 *Automatic Restoration Processing Chain*

Chapters 2 to 5 concentrate on the detection, parameter estimation and reduction of individual disturbances. In combination, the proposed algorithms can be used to determine which disturbance types are present in recording, or in a section of a recording. A surely reasonable next step is to combine the algorithms and create an automatic restoration system for all three disturbance types, e.g., implementing the processing workflow described in Chapter 1. As mentioned before, it is usually sensible to treat the disturbance types in a specific order, e.g., to reduce masking effects and ease the disturbance parameter estimation. The automatic restoration system first may detect and remove impulsive disturbances, followed by hum disturbances and, finally, broadband noise. While the implementation of this system is most probably straightforward, its evaluation under realistic conditions that are typical for archive audio would be very interesting.

6.2.5 *Automatic Information Retrieval*

While the focus of this thesis is the restoration of recordings in media archives, an interesting application of the proposed algorithms is automatic information retrieval from the audio signals. For many recordings, only little information is available other then the recording itself. This information may regard the original carrier, the age of a recording or the recording location, and can be useful for the management of an archive. In addition, information retrieved from the signals may allow to draw conclusions about the history of a recording, possibly in combination with further information. For example, the presence of certain disturbance types indicates a certain original carrier, potentially giving indications on the age of a recording. The presence of a hum disturbance with a specific fundamental frequency may give hints on the country where the recording was made. Altogether, the proposed algorithms may help to increase the accessibility and management of media archives that get bigger and bigger every day.

REFERENCES

[1] G. Bryan, *Edison – The Man and His Work*, London, England: Alfred A. Knopf (1930).

[2] C. R. Mc Queary, "The Early Years of the Acoustic Phonograph: Its Developmental Origins and Fall From Favor 1877–1929," Thesis, Lubbock, USA: Texas Tech University (Mar. 1990).

[3] L. J. Newville, "Development of the Phonograph at Alexander Graham Bell's Volta Laboratory," *Contributions from The Museum of History and Technology*, pp. 69–79 (1959).

[4] A. Millard, *America on Record: A History of Recorded Sound*, 2nd ed., Cambridge/New York, USA: Cambridge University Press (2005).

[5] I. McNeil, ed., *An Encyclopaedia of the History of Technology*, London/New York: Routledge (1990).

[6] N. P. da Costa, *Off the Record: Performing Practices in Romantic Piano Playing*, New York, USA: Oxford University Press (2012).

[7] W. Hinz, "Audio Technology in Berlin to 1943: Recording and Playing Equipment," *Proceedings of the 94th Audio Engineering Society Convention*, Berlin, Germany (1993 Mar.).

[8] J. Nichols, "A High-Performance, Low-Cost Wax Cylinder Transcription System," *Proceedings of the 20th Audio Engineering Society Conference* (2001 Oct.).

[9] W. R. Isom, "Record Materials, Part II: Evolution of the Disc Talking Machine," *Journal of the Audio Engineering Society*, vol. 25, no. 10, pp. 718–723 (1977).

[10] J. B. Minter, "Recent Developments in Precision Master Recording Lathes," *Journal of the Audio Engineering Society*, vol. 4, no. 2, pp. 50–55 (1956 Apr.).

[11] E. D. Daniel, C. D. Mee, M. H. Clark, *Magnetic Recording: The First 100 Years*, New York, USA: Wiley-IEEE Press (1999).

[12] F. Engel, P. Hammar, R. Hess, "A Selected History of Magnetic Recording" (2006).

[13] F. Hirsch, "Techniques for Improving the Quality of High Speed Tape Duplication," *Proceedings of the 80th Audio Engineering Society Convention*, Montreux, Switzerland (1986 Mar.).

[14] R. Dolby, "An Audio Noise Reduction System," *Journal of the Audio Engineering Society*, vol. 15, no. 4, pp. 383–388 (1967 Oct.).

[15] R. Dolby, "A Noise Reduction System for Consumer Tape Recording," *Proceedings of the Audio Engineering Society Convention 2ce*, Cologne, Germany (1972 Mar.).

[16] J. Mosely, "Motion Picture Sound in Record-Industry Perspective," *Journal of the Audio Engineering Society*, vol. 29, no. 3, pp. 114–125 (1981 Mar.).

[17] D. Lynskey, "How The Compact Disc Lost its Shine," *The Guardian* (2015 May). `http://www.theguardian.com/music/2015/may/28/how-the-compact-disc-lost-its-shine`

[18] T. Fine, "The Dawn of Commercial Digital Recording," *ARSC Journal*, vol. 39, no. 1, pp. 1–17 (2008).

[19] T. T. Doi, T. Itoh, H. Ogawa, "A Long-Play Digital Audio Disk System," *Journal of the Audio Engineering Society*, vol. 27, no. 12, pp. 975–981 (1979 Dec.).

[20] *Synclavier Early History.* `http://www.500sound.com/synclavierhistory.html`

[21] N. B. M. Inc, *Billboard*, London, England: Nielsen Business Media, Inc. (Oct. 23, 1982).

[22] International Organization for Standardization, *ISO/IEC 11172-3:1993 – Information Technology – Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1,5 Mbit/s – Part 3: Audio* Geneva, Switzerland, (Aug. 1993).

[23] Library of Congress, *Annual Report of the Librarian of Congress*, Washington, D.C., USA (2016).

[24] The National Archives and Records Administration, *About the National Archives of the United States*. `https://www.archives.gov/publications/general-info-leaflets/1-about-archives.html`

[25] S. Barteleit, B. Kolev, T. Menzel, A. Kunz, R. Jacobs, G. Eckes, *Forum – Das Fachmagazin des Bundesarchivs*, Koblenz, Germany: Bundesarchiv (2014).

[26] D. Callahan, *A History of Birdwatching in 100 Objects*, ed. by D. Mitchell, London, England: Helm (2014).

[27] National Film and Sound Archive of Australia, *Fanny Cochrane Smith's Tasmanian Aboriginal Songs*, ID: 500445.

[28] F. Densmore, *The American Indians and Their Music*, New York, USA: The Womans Press (1926).

[29] E. Grieg, "The Piano Music in Historic Interpretations," *Simax PSC 1809* (1992).

[30] P. Shambarger, "Cylinder Records: An Overview," *Journal of the Association for Recorded Sound Collections (ARSC)*, vol. 2, no. 26, pp. 133–161 (1995).

[31] A. Assmann, *Erinnerungsräume: Formen und Wandlungen des kulturellen Gedächtnisses*, 5th ed., Munich, Germany: C.H.Beck (2010).

[32] *Memory of the World Programme – Safeguarding the Documentary Heritage*, ed. by G. Boston, M. Keynes, Paris, France: United Nations Educational, Scientific and Cultural Organization (UNESCO) (Apr. 1998).

[33] G. Boston, G. Brock-Nannestad, L. Gaustad, A. Häfner, D. Schüller, T. Sjöberg, "The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy," *International Association of Sound and Audiovisual Archives* (2005 Dec.).

[34] C. A. Paton, “Preservation Re-Recording of Audio Recordings in Archives: Problems, Priorities, Technologies, and Recommendations,” *The American Archivist*, vol. 61, no. 1, pp. 188–219 (1998).

[35] F. Bressan, S. Canazza, “A Systemic Approach to the Preservation of Audio Documents: Methodology and Software Tools,” *Journal of Electrical and Computer Engineering* (2013). `https://doi.org/10.1155/2013/489515`

[36] Panel on Removal of Noise From a Speech/Noise Signal, *Removal of Noise From Noise-Degraded Speech Signals*, Washington, D.C., USA: National Academies Press (1989). `https://doi.org/10.17226/19048`

[37] J. Benesty, S. Makino, J. Chen, *Speech Enhancement*, Berlin/Heidelberg, Germany: Springer (2005).

[38] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 4th ed., Chichester/West Sussex, England: Wiley (2008).

[39] S. J. Godsill, P. J. Rayner, *Digital Audio Restoration—A Statistical Model-Based Approach*, London, England: Springer (1998). `https://doi.org/10.1007/978-1-4471-1561-8`

[40] S. C. Bartholomew, “Power Circuit Interference with Telegraphs and Telephones,” *Journal of the Institution of Electrical Engineers*, vol. 62, no. 334, pp. 817–858 (1924 Oct.). `https://doi.org/10.1049/jiee-1.1924.0111`

[41] J. A. Moorer, “DSP Restoration Techniques for Audio,” *Proceedings of the IEEE International Conference on Image Processing*, San Antonio, USA, vol. 4, pp. IV-5–IV-8 (2007 Sept.). `https://doi.org/10.1109/ICIP.2007.4379940`

[42] J. Bitzer, J. Houpert, “Azimuth-Correction: Digital Solutions in the Time- and Frequency-Domain,” *Proceedings of the 106th Audio Engineering Society Convention*, Munich, Germany (1999 May).

[43] N. Ikonomov, “Preservation and Digital Restoration of Audio Archives,” *Review of the National Center for Digitization*, no. 2, pp. 40–45 (2003).

[44] C. Landone, J. Harrop, J. Reiss, “Enabling Access to Sound Archives Through Integration, Enrichment and Retrieval: The EASAIER Project,” *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, pp. 159–160 (2007 Sept.).

[45] I. Damnjanovic, C. Landone, P. Kudumakis, J. Reiss, “Intelligent Infrastructure for Accessing Sound and Related Multimedia Objects,” *Proceedings of the International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution*, Florence, Italy, pp. 121–126 (2008 Nov.). `https://doi.org/10.1109/AXMEDIS.2008.18`

[46] W. D. Storm, “A Proposal for the Establishment of International Re-Recording Standards,” *Journal of the Association for Recorded Sound Collections (ARSC)*, vol. 15, no. 2–3, pp. 26–37 (1983).

[47] D. Schüller, “The Ethics of Preservation, Restoration, and Re-Issues of Historical Sound Recordings,” *Journal of the Audio Engineering Society*, vol. 39, no. 12, pp. 1014–1017 (1991 Dec.).

[48] R. Lagadec, D. Pelloni, “Signal Enhancement via Digital Signal Processing,” *Proceedings of the 74th Audio Engineering Society Convention*, New York, USA (1983).

[49] A. Czyzewski, B. Kostek, A. Kupryjanow, "Automatic Sound Restoration System – Concepts and Design," *Proceedings of the International Conference on Signal Processing and Multimedia Applications (SIGMAP)*, Seville, Spain, pp. 207–211 (2011 July). `https://doi.org/10.5220/0003527702070211`

[50] G. R. Kinzie Jr., D. W. Gravereaux, "Automatic Detection of Impulse Noise," *Journal of the Audio Engineering Society*, vol. 21, no. 3, pp. 181–184 (1973 Apr.).

[51] J. M. Sacks, B. Isenberg, S. Klynas, "Reduction of Impulse Noise in Audio Signals," *Proceedings of the 57th Audio Engineering Society Convention*, Los Angeles, USA (1977 May).

[52] S. Vaseghi, P. Rayner, "Detection and Suppression of Impulsive Noise in Speech Communication Systems," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 137, no. 1, pp. 38–46 (1990 Feb.). `https://doi.org/10.1049/ip-i-2.1990.0007`

[53] M. Niedźwiecki, M. Ciołek, "Elimination of Impulsive Disturbances From Archive Audio Signals Using Bidirectional Processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1046–1059 (2013 May). `https://doi.org/10.1109/TASL.2013.2244090`

[54] M. Ciołek, M. Niedźwiecki, "Detection of Impulsive Disturbances in Archive Audio Signals," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, pp. 671–675 (2017 Mar.). `https://doi.org/10.1109/ICASSP.2017.7952240`

[55] M. Niedźwiecki, M. Ciołek, "Localization of Impulsive Disturbances in Archive Audio Signals using Predictive Matched Filtering," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, pp. 2888–2892 (2014 May). `https://doi.org/10.1109/ICASSP.2014.6854128`

[56] A. Czyżewski, C. Suproń, "Learning Algorithms for the Cancellation of Old Recordings Noise," *Proceedings of the 96th Audio Engineering Society Convention*, Amsterdam, Netherlands (1994 Feb.).

[57] A. Czyżewski, "Artificial Intelligence-Based Processing of Old Audio Recordings," *Proceedings of the 97th Audio Engineering Society Convention*, San Francisco, USA (1994 Nov.).

[58] A. Czyżewski, "Learning Algorithms for Audio Signal Enhancement, Part 1: Neural Network Implementation for the Removal of Impulse Distortions," *Journal of the Audio Engineering Society*, vol. 45, no. 10, pp. 815–831 (1997 Oct.).

[59] J. K. Kauppinen, P. E. Saarinen, "True Linear Prediction by Use of a Theoretical Impulse Response," *Journal of the Optical Society of America B*, vol. 11, no. 9, pp. 1631–1638 (1994 Sept.). `https://doi.org/10.1364/JOSAB.11.001631`

[60] I. Kauppinen, J. Kauppinen, P. Saarinen, "A Method for Long Extrapolation of Audio Signals," *Journal of the Audio Engineering Society*, vol. 49, no. 12, pp. 1167–1180 (2001 Dec.).

[61] P. A. A. Esquef, V. Välimäki, K. Roth, I. Kauppinen, "Interpolation of Long Gaps in Audio Signals using the Warped Burg's Method," *Proceedings of*

the 6th International Conference on Digital Audio Effects (DAFx), London, England, pp. 8–11 (2003 Sept.).

[62] J. Nuzman, *Audio Restoration: An Investigation of Digital Methods for Click Removal and Hiss Reduction* (Jan. 2004). `http://jnuzman.github.io/audio-restoration-2004/`

[63] Glenn Miller and his Orchestra, Ray Eberle, Lewis, Stock, Rose, Internet Archive, *Blueberry Hill*, Bluebird, B-10768 (1940). `http://archive.org/details/78_blueberry-hill_glenn-miller-and-his-orchestra-ray-eberle-lewis-stock-rose_gbia0012035a`

[64] Jonny Bla5t, *Power (Original Mix)*, Jamendo (2013).

[65] Silver Aeroplane, *Tide Change*, Jamendo (2015).

[66] H. R. Pfitzinger, "Removing Hum From Spoken Language Resources," *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, vol. 3, pp. 618–621 (2000).

[67] R. A. Dobre, V. A. Niţă, A. Ciobanu, C. Negrescu, D. Stanomir, "A Hum Removal Algorithm Used for Audio Restoration Purposes," *Proceedings of the International Symposium on Signals, Circuits and Systems (ISSCS)*, Iaşi, Romania, pp. 1–4 (2015 July). `https://doi.org/10.1109/ISSCS.2015.7204000`

[68] M. R. Schroeder, *Apparatus for Suppressing Noise and Distortion in Communication Signals*, US Patent 3,180,936 (Apr. 1965).

[69] R. E. Crochiere, "A Weighted Overlap-Add Method of Short-time Fourier Analysis/Synthesis," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 1, pp. 99–102 (1980 Feb.). `https://doi.org/10.1109/TASSP.1980.1163353`

[70] J. G. Proakis, D. G. Manolakis, *Digital Signal Processing*, 4th ed., Upper Saddle River, USA: Prentice Hall, pp. 823–879 (2006).

[71] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120 (1979 Apr.). `https://doi.org/10.1109/TASSP.1979.1163209`

[72] J. Benesty, Jingdong Chen, E. A. Habets, *Speech Enhancement in the STFT Domain*, Berlin/Heidelberg, Germany: Springer (2012). `https://doi.org/10.1007/978-3-642-23250-3`

[73] J. Chen, J. Benesty, Y. A. Huang, S. Doclo, "New Insights Into the Noise Reduction Wiener Filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234 (2006 July). `https://doi.org/10.1109/TSA.2005.860851`

[74] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed., Boca Raton, USA: CRC Press (2013).

[75] R. C. Hendriks, T. Gerkmann, J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*, Morgan & Claypool (2013). `https://doi.org/10.2200/S00473ED1V01Y201301SAP011`

[76] R. McAulay, M. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Transactions on Acoustics, Speech, and Signal*

Processing, vol. 28, no. 2, pp. 137–145 (1980 Apr.). `https://doi.org/10.1109/TASSP.1980.1163394`

[77] Y. Ephraim, 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 (1984 Dec.). `https://doi.org/10.1109/TASSP.1984.1164453`

[78] Y. Ephraim, D. Malah, "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 (1985 Apr.). `https://doi.org/10.1109/TASSP.1985.1164550`

[79] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Washington, D. C., USA, vol. 4, pp. 208–211 (1979 Apr.). `https://doi.org/10.1109/ICASSP.1979.1170788`

[80] S. Gustafsson, P. Jax, P. Vary, "A Novel Psychoacoustically Motivated Audio Enhancement Algorithm Preserving Background Noise Characteristics," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, USA, vol. 1, 397–400 vol.1 (1998 May). `https://doi.org/10.1109/ICASSP.1998.674451`

[81] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137 (1999 Mar.). `https://doi.org/10.1109/89.748118`

[82] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349 (1994 Apr.). `https://doi.org/10.1109/89.279283`

[83] O. Cappé, J. Laroche, "Evaluation of Short-Time Spectral Attenuation Techniques for the Restoration of Musical Recordings," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 84–93 (1995 Jan.). `https://doi.org/10.1109/89.365378`

[84] G. Yu, S. Mallat, E. Bacry, "Audio Denoising by Time-Frequency Block Thresholding," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1830–1839 (2008). `https://doi.org/10.1109/TSP.2007.912893`

[85] K. Siedenburg, M. Dörfler, "Audio Denoising by Generalized Time-Frequency Thresholding," *Proceedings of the 45th International Audio Engineering Society Conference*, Helsinki, Finland (2012 Mar.).

[86] E. V. Harinarayanan, A. Ferreira, S. Saeed, D. Sinha, "A Novel Automatic Noise Removal Technique for Audio and Speech Signals," *Proceedings of the 123rd Audio Engineering Society Convention*, New York, USA (2007 Oct.).

[87] J. Benesty, M. M. Sondhi, Y. Huang, *Springer Handbook of Speech Processing*, Berlin/Heidelberg, Germany: Springer (2008).

[88] F. Heese, M. Niermann, P. Vary, "Speech-Codebook Based Soft Voice Activity Detection," *Proceedings of the IEEE International Conference on Acous-*

tics, Speech and Signal Processing (ICASSP), Brisbane, Australia, pp. 4335–4339 (2015 Apr.). `https://doi.org/10.1109/ICASSP.2015.7178789`

[89] X. Li, R. Horaud, L. Girin, S. Gannot, “Voice Activity Detection Based on Statistical Likelihood Ratio With Adaptive Thresholding,” *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, Xi’an, China (2016 Sept.).

[90] R. Martin, “Spectral Subtraction Based on Minimum Statistics,” *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, Lisbon, Portugal (1994 Sept.).

[91] 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 (2001 July). `https://doi.org/10.1109/89.928915`

[92] I. Cohen, B. Berdugo, “Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement,” *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15 (2002 Jan.). `https://doi.org/10.1109/97.988717`

[93] 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 (2003 Sept.). `https://doi.org/10.1109/TSA.2003.811544`

[94] T. Gerkmann, R. C. Hendriks, “Unbiased MMSE-Based Noise Power Estimation With Low Complexity and Low Tracking Delay,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1383–1393 (2012 May). `https://doi.org/10.1109/TASL.2011.2180896`

[95] V. Stahl, A. Fischer, R. Bippus, “Quantile Based Noise Estimation for Spectral Subtraction and Wiener Filtering,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, vol. 3, pp. 1875–1878 (2000).

[96] N. W. D. Evans, J. S. Mason, *Time-Frequency Quantile-Based Noise Estimation* (2002).

[97] G. García, “Automatic Denoising for Musical Audio Restoration,” PhD thesis, Stanford, USA: Stanford University (2009).

[98] D. Van Compernolle, W. Ma, F. Xie, M. Van Diest, “Speech Recognition in Noisy Environments with the Aid of Microphone Arrays,” *Speech Communication*, vol. 9, no. 5, pp. 433–442 (1990 Dec.). `https://doi.org/10.1016/0167-6393(90)90019-6`

[99] The League, *The Soundtrack Of Our Summer*, Jamendo (2009).

[100] T. Lorenz, “High-End Audio Restoration,” *Proceedings of the 26th VDT International Convention*, Leipzig, Germany (2010 Nov.).

[101] S. Bech, N. Zacharov, *Perceptual Audio Evaluation—Theory, Method and Application*, Chichester/West Sussex, England: John Wiley & Sons (2006).

[102] M. G. Kendall, B. B. Smith, “On the Method of Paired Comparisons,” *Biometrika*, vol. 31, no. 3, pp. 324–345 (1940 Mar.). `https://doi.org/10.2307/2332613`

[103] R. A. Bradley, M. E. Terry, “Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons,” *Biometrika*, vol. 39, no. 3, pp. 324–345 (1952 Dec.). `https://doi.org/10.2307/2334029`

[104] K. Tsukida, M. R. Gupta, *How to Analyze Paired Comparison Data* (May 2011).

[105] ITU-R, “Recommendation BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems” (2015 Oct.).

[106] A. Gray, J. Markel, “Distance Measures for Speech Processing,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391 (1976 Oct.). `https://doi.org/10.1109/TASSP.1976.1162849`

[107] F. Itakura, S. Saito, “Analysis Synthesis Telephony Based on the Maximum Likelihood Method,” *Proceedings of the 6th International Congress on Acoustics*, Tokyo, Japan, vol. 2, pp. C-17–C-20 (1968 Aug.).

[108] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, “PEAQ—The ITU Standard for Objective Measurement of Perceived Audio Quality,” *Journal of the Audio Engineering Society*, vol. 48, no. 1, pp. 3–29 (2000 Feb.).

[109] International Telecommunication Union, *Method for Objective Measurements of Perceived Audio Quality*, Recommendation BS.1387-1, Geneva, Switzerland: ITU-R (Nov. 2001)

[110] P. Kabal, “An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality,” Technical Report, Montreal, Canada: Dept. Electrical & Computer Engineering, McGill University (May 2002).

[111] M. Brandt, *Impulsive Disturbances in Audio Archives: Signal Classification for Automatic Restoration – Demonstration Signals Accompanying the Article*. `http://www.matbra.org`

[112] M. Brandt, *Automatic Noise PSD Estimation for Archive Audio Restoration —Website Accompanying the Paper*. `https://matbra.github.io/noise_psd_estimation`

[113] M. Brandt, S. Doclo, T. Gerkmann, J. Bitzer, “Impulsive Disturbances in Audio Archives: Signal Classification for Automatic Restoration,” *Journal of the Audio Engineering Society*, vol. 65, no. 10, pp. 826–840 (2017 Oct.). `https://doi.org/10.17743/jaes.2017.0032`

[114] M. Brandt, J. Bitzer, “Automatic Detection of Hum in Audio Signals,” *Journal of the Audio Engineering Society*, vol. 62, no. 9, pp. 584–595 (2014 Oct.). `https://doi.org/10.17743/jaes.2014.0034`

[115] M. Brandt, J. Bitzer, “Hum Removal Filters: Overview and Analysis,” *Proceedings of the 132nd Audio Engineering Society Convention*, Budapest, Hungary (2012 Apr.).

[116] M. Brandt, S. Doclo, J. Bitzer, “Automatic Noise PSD Estimation for Archive Audio Restoration,” *Submitted to the Journal of the Audio Engineering Society* (2018 Mar.).

[117] B. Lyons, R. Chandler, C. Lacinak, *Quantifying the Need: A Survey of Existing Sound Recordings in Collections in the United States*, New York, USA: AVPreserve (May 2015)

[118] A. J. E. M. Janssen, R. N. J. Veldhuis, L. B. Vries, “Adaptive Interpolation of Discrete-Time Signals That Can Be Modeled as Autoregressive Processes,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 2, pp. 317–330 (1986 Apr.). `https://doi.org/10.1109/TASSP.1986.1164824`

[119] S. Vaseghi, P. Rayner, “A New Application of Adaptive Filters for Restoration of Archived Gramophone Recordings,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New York, USA, vol. 5, pp. 2548–2551 (1988 Apr.). `https://doi.org/10.1109/ICASSP.1988.197163`

[120] S. V. Vaseghi, “Algorithms for Restoration of Archived Gramophone Recordings,” PhD thesis, Cambridge, England: Cambridge University (1988).

[121] R. Veldhuis, *Restoration of Lost Samples in Digital Signals*, Prentice Hall International Series in Acoustics, Speech and Signal Processing, New York, USA: Prentice Hall (1990).

[122] C. Hicks, S. Godsill, “A Two-Channel Approach to the Removal of Impulsive Noise from Archived Recordings,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Adelaide, Australia, vol. 2, pp. II-213–II-216 (1994 Apr.). `https://doi.org/10.1109/ICASSP.1994.389681`

[123] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. J. B. Jackson, M. D. Plumbley, “Unsupervised Feature Learning Based on Deep Models for Environmental Audio Tagging,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1230–1241 (2017 June). `https://doi.org/10.1109/TASLP.2017.2690563`

[124] Y. Lavner, R. Cohen, D. Ruinskiy, H. Ijzerman, “Baby Cry Detection in Domestic Environment using Deep Learning,” *Proceedings of the IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, Eilat, Israel, pp. 1–5 (2016 Nov.). `https://doi.org/10.1109/ICSEE.2016.7806117`

[125] V. Välimäki, S. González, O. Kimmelma, J. Parviainen, “Digital Audio Antiquing—Signal Processing Methods for Imitating the Sound Quality of Historical Recordings,” *Journal of the Audio Engineering Society*, vol. 56, no. 3, pp. 115–139 (2008 Mar.).

[126] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., New York, USA: McGraw-Hill (1991).

[127] A. K. Southey, M. Fox, T. Yeomans, “A Comparison of the Characteristics of ISO Fine Test Dust versus Real House Dust,” *Proceedings of the 12th International Conference on Indoor Air Quality and Climate*, Austin, USA, vol. 868 (2011 June).

[128] P. Rayner, S. Godsill, “The Detection and Correction of Artefacts in Degraded Gramophone Recordings,” *Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, pp. 151–152 (1991 Oct.). `https://doi.org/10.1109/ASPAA.1991.634139`

[129] C. Knapp, G. C. Carter, “The Generalized Correlation Method for Estimation of Time Delay,” *IEEE Transactions on Acoustics, Speech, and Signal*

Processing, vol. 24, no. 4, pp. 320–327 (1976 Aug.). `https://doi.org/10.1109/TASSP.1976.1162830`

[130] K. D. Donohue, J. Hannemann, H. G. Dietz, "Performance of Phase Transform for Detecting Sound Sources with Microphone Arrays in Reverberant and Noisy Environments," *Signal Processing*, vol. 87, no. 7, pp. 1677–1691 (2007 July). `https://doi.org/10.1016/j.sigpro.2007.01.013`

[131] D. Aiger, H. Talbot, "The Phase Only Transform for Unsupervised Surface Defect Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, USA, pp. 295–302 (2010 June). `https://doi.org/10.1109/CVPR.2010.5540198`

[132] R. F. Voss, J. Clarke, "'$1/f$ noise' in Music and Speech," *Nature*, vol. 258, no. 5533, pp. 317–318 (1975 Nov.). `https://doi.org/10.1038/258317a0`

[133] J. W. Tukey, "A Survey of Sampling from Contaminated Distributions," in: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pp. 448–485, Stanford, USA: Stanford Univ. Press (1960 Jan.)

[134] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1-3, pp. 389–422 (2002 Jan.). `https://doi.org/10.1023/A:1012487302797`

[135] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830 (2011 Oct.).

[136] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York, USA: Springer, pp. 32–33 (2007).

[137] Various, *50 Jahre Popmusik. Ein Jahr und seine 20 besten Songs, 1955–2005*, Munich, Germany: Süddeutsche Zeitung (2005).

[138] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, ROC Analysis in Pattern Recognition vol. 27, no. 8, pp. 861–874 (2006 June). `https://doi.org/10.1016/j.patrec.2005.10.010`

[139] D. M. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63 (2011 Dec.).

[140] P. A. A. Esquef, L. W. P. Biscainho, P. S. R. Diniz, F. P. Freelanci, "A Double-Threshold-Based Approach to Impulsive Noise Detection in Audio Signals," *Proceedings of the 10th European Signal Processing Conference (EUSIPCO)*, Tampere, Finland, pp. 1–4 (2000 Sept.).

[141] L. Oudre, "Automatic Detection and Removal of Impulsive Noise in Audio Signals," *Image Processing On Line*, vol. 5, pp. 267–281 (2015 Nov.). `https://doi.org/10.5201/ipol.2015.64`

[142] R. McGill, J. W. Tukey, W. A. Larsen, "Variations of Box Plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16 (1978 Feb.). `https://doi.org/http://dx.doi.org/10.2307/2683468`

[143] R. R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, 3rd ed., Amsterdam, Netherlands: Academic Press (2012).

[144] P. O. Hoyer, "Non-Negative Matrix Factorization with Sparseness Constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469 (2004 Nov.).

[145] C. Grigoras, "Digital Audio Recording Analysis: The Electric Network Frequency (ENF) Criterion," *International Journal of Speech Language and the Law*, vol. 12, no. 1, pp. 63–76 (2005).

[146] A. Czyzewski, A. Ciarkowski, A. Kaczmarek, J. Kotus, M. Kulesza, P. Maziewski, "DSP Techniques for Determining 'Wow' Distortion," *Journal of the Audio Engineering Society*, vol. 55, no. 4, pp. 266–284 (2007 Apr.).

[147] Y.Z. Liu, S. Chen, "A Wavelet Based Model for On-line Tracking of Power System Harmonics using Kalman Filtering," *Proceedings of the Power Engineering Society Summer Meeting*, Vancouver, Canada, vol. 2, pp. 1237–1242 (2001). `https://doi.org/10.1109/PESS.2001.970245`

[148] A. Klapuri, M. Davy, *Signal Processing Methods for Music Transcription*, New York, USA: Springer (2006).

[150] D. Manquen, "A Wideband Tape and Transport Diagnostic Method," *Proceedings of the 66th Audio Engineering Society Convention*, Los Angeles, USA (1980 May).

[151] S. J. Godsill, P. J. Rayner, "The Restoration of Pitch Variation Defects in Gramophone Recordings," *Final Program and Paper Summaries of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, USA, pp. 148–151 (1993 Oct.). `https://doi.org/10.1109/ASPAA.1993.379975`

[152] J. Howarth, P. J. Wolfe, "Correction of Wow and Flutter Effects in Analog Tape Transfers," *Proceedings of the 117th Audio Engineering Society Convention*, San Francisco, USA (2004 Oct.).

[153] P. Vary, R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, Chichester, England: John Wiley & Sons (2006).

[154] A. E. Beaton, J. W. Tukey, "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," *Technometrics*, vol. 16, no. 2, pp. 147–185 (1974 May). `https://doi.org/10.2307/1267936`

[155] P. K. Dash, B. R. Mishra, R. K. Jena, A. C. Liew, "Estimation of Power System Frequency Using Adaptive Notch Filters," *Proceedings of the International Conference on Energy Management and Power Delivery*, Singapore, Singapore, vol. 1, pp. 143–148 (1998 Mar.). `https://doi.org/10.1109/EMPD.1998.705491`

[156] M. R. Schroeder, "Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement," *The Journal of the Acoustical Society of America*, vol. 43, no. 4, pp. 829–834 (1968). `https://doi.org/10.1121/1.1910902`

[157] Wikipedia spoken article, *Bird* (May 1, 2008).

[158] J. B. Vanhal, *Symphony in C Minor – Menuetto Moderato* (1760).

[159] J. Osborne, *Relish* (Mar. 21, 1995).

[160] S. Gade, H. Herlufsen, "Windows to FFT Analysis (Part I)," *Brüel & Kjær Technical Review*, no. 3 (1987).

[161] S. Gade, H. Herlufsen, “Windows to FFT Analysis (Part II),” *Brüel & Kjær Technical Review*, no. 4 (1987).

[162] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Upper Saddle River, USA: Prentice Hall (1993).

[163] S. J. Orfanidis, *Introduction to Signal Processing* (2010).

[164] P. A. Regalia, S. K. Mitra, P. P. Vaidyanathan, “The Digital All-Pass Filter: A Versatile Signal Processing Building Block,” *Proceedings of the IEEE*, vol. 76, pp. 19–37 (1988 Jan.). `https://doi.org/10.1109/5.3286`

[165] P. Jarske, S. K. Mitra, Y. Neuvo, “Signal Processor Implementation of Variable Digital Filters,” *IEEE Transactions on Instrumentation and Measurement*, vol. 37, no. 3, pp. 363–367 (1988 Sept.). `https://doi.org/10.1109/19.7456`

[166] NoiseCollector, *Sound No. 57101*, freesound.org (July 12, 2008).

[167] J. Eargle, *Handbook of Recording Engineering*, 4th ed., Boston, USA: Kluwer Academic Publishers (2003).

[168] B. Fries, M. Fries, *Digital Audio Essentials: A Comprehensive Guide to Creating, Recording, Editing, and Sharing Music and Other Audio*, Sebastopol, USA: O’Reilly Media, Inc. (2005).

[169] F. Bressan, S. Canazza, D. Salvati, “The Vicentini Sound Archive of the *Arena di Verona* Foundation: A Preservation and Restoration Project,” *Workshop on Exploring Musical Information Spaces (WEMIS)*, Corfu, Greece (2009 Oct.).

[170] S. Rangachari, P. C. Loizou, “A Noise-Estimation Algorithm for Highly Non-Stationary Environments,” *Speech Communication*, vol. 48, no. 2, pp. 220–231 (2006 Feb.). `https://doi.org/10.1016/j.specom.2005.08.005`

[171] P. D. Welch, “The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, pp. 70–73 (1967 June). `https://doi.org/10.1109/TAU.1967.1161901`

[172] K. Siedenburg, M. Dörfler, “Persistent Time-Frequency Shrinkage for Audio Denoising,” *Journal of the Audio Engineering Society*, vol. 61, no. 1, pp. 29–38 (2013 Mar.).

[173] V. Mach, “Denoising phonogram cylinders recordings using Structured Sparsity,” *2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), pp. 314–319 (2015 Oct.). `https://doi.org/10.1109/ICUMT.2015.7382449`

[174] J. Schoukens, J. Renneboog, “Modeling the Noise Influence on the Fourier Coefficients After a Discrete Fourier Transform,” *IEEE Transactions on Instrumentation and Measurement*, vol. IM-35, no. 3, pp. 278–286 (1986 Sept.). `https://doi.org/10.1109/TIM.1986.6499210`

[175] A. Papoulis, S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed., New York, USA: McGraw-Hill (2002).

[176] A. W. v. d. Vaart, *Asymptotic Statistics*, Cambridge, England: Cambridge University Press (1998). `https://doi.org/10.1017/{CBO}9780511802256`

[177] J. D. Gibbons, S. Chakraborti, *Nonparametric Statistical Inference*, 4th ed., Basel, Switzerland: Marcel Dekker Inc. (2003).

[178] T. W. Anderson, D. A. Darling, “Asymptotic Theory of Certain ‘Goodness of Fit’ Criteria Based on Stochastic Processes,” *The Annals of Mathematical Statistics*, vol. 23, no. 2, pp. 193–212 (1952 June). `https://doi.org/10.1214/aoms/1177729437`

[179] J. Lin, “Divergence Measures Based on the Shannon Entropy,” *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151 (1991 Jan.). `https://doi.org/10.1109/18.61115`

[180] N. M. Razali, Y. B. Wah, “Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests,” *Journal of Statistical Modeling and Analytics*, vol. 2, no. 1 (2011 June).

[181] monotraum, *Sound No. 242209*, freesound.org (2014).

[182] Yuval, *Sound No. 197795*, freesound.org (2013).

[183] D. P. W. Ellis, “Beat Tracking by Dynamic Programming,” *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60 (2007 Mar.). `https://doi.org/10.1080/09298210701653344`

[184] M. E. P. Davies, M. D. Plumbley, “Context-Dependent Beat Tracking of Musical Audio,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1009–1020 (2007 Mar.). `https://doi.org/10.1109/TASL.2006.885257`

[185] A. H. Land, A. G. Doig, “An Automatic Method of Solving Discrete Programming Problems,” *Econometrica*, vol. 28, no. 3, pp. 497–520 (1960). `https://doi.org/10.2307/1910129`

ACKNOWLEDGMENTS

This thesis would not have been possible without the sustained support of several people. I am very thankful to have had great mentors, colleagues and friends who accompanied my life as a PhD candidate.

First of all, I would like to express my deep gratitude to Prof. Dr. ir. Simon Doclo and Prof. Dr.-Ing. Joerg Bitzer who accepted me as a PhD candidate. I am honoured to have had the opportunity to work with both of them and I am very thankful for many fruitful discussions improving my work in numerous ways.

I would like to thank Joerg Bitzer who gave me the opportunity to work on this interesting research topic and for letting me benefit from his great amount of knowledge and his experience with audio restoration algorithms. I always felt very lucky for being able to combine my interest in signal processing with my passion for music.

I would like to thank Simon Doclo for his valuable suggestions on the developed algorithms, their evaluation and in particular his very helpful advice regarding scientific writing.

I also would like to thank Prof. Dr. Joshua Reiss and Prof. Dr. Steven van de Par for their kind participation in my thesis committee.

I thank Dr. Paulo Esquef, Joseph Nuzman, Robert Rehr and Dr. Kai Siedenburg in particular for providing implementations of some of the reference algorithms used in this thesis.

Thanks to my colleagues at the Institute for Hearing Technology and Audiology for creating a stimulating and fun working environment, with lots of amusing coffee breaks and lighthearted discussions. It felt great to be with so many people who are sharing the same passion for signal processing and music.

I also would like to thank Dr.-Ing. Uwe Simmer for proofreading this thesis.

The financial support by the Bundesministerium für Bildung und Forschung (grant no. 17N3008) is greatly acknowledged.

Most importantly, I want to thank my wonderful girlfriend, my fantastic friends and my dear family for always being there for me and supporting me in innumerable ways. I would not have been able to do this without you guys. Love.