M. Brandt, S. Doclo and J. Bitzer, "Automatic Noise PSD Estimation for Restoration of Archived Audio,"
J. Audio Eng. Soc., vol. 67, no. 1/2, pp. 38–53, (2019 January/February.).
DOI: https://doi.org/10.17743/jaes.2018.0073

Automatic Noise PSD Estimation for Restoration of Archived Audio

MATTHIAS BRANDT[1], **SIMON DOCLO**,[1] *AES Associate Member*, **AND**
(matthias.brandt@uni-oldenburg.de)

JOERG BITZER,[2] *AES Member*

[1] *University of Oldenburg, Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Oldenburg, Germany*
[2] *Jade University of Applied Sciences, Oldenburg, Germany*

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 that 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.

0 INTRODUCTION

The quality of audio recordings is often degraded by various types of disturbances, such as broadband noise, hum, clicks, and crackles [1–4]. Broadband noise is one of the most frequently occurring types of disturbance, especially in old recordings, and can be classified as having either a *technical* or *acoustic* origin [2, 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. 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. On the other hand, 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 a degradation of the original acoustic signal. Early recording media, such as wax cylinders and shellac discs, typically had SNRs of below 40 dB [5]. The vinyl disc represented a large improvement in the dynamic range, with SNRs of 55 dB to 60 dB [6]. The next improvement was marked by magnetic tape storage media, leading to SNRs between 60 dB and 70 dB [7], allowing for the first recordings that practically contain no perceivable noise disturbance. The compact disc, commercially introduced in 1982, made dynamic ranges above 90 dB possible [8]. 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 [9–11]. In doing so, a central problem is the estimation of the characteristics of the broadband noise disturbance. State-of-the-art audio restoration algorithms [2, 12, 13] often require the manual selection of one or more noise-only sections of a recording to determine a so-called *noise fingerprint* [1]. 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 [2], 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, the manual restoration of large numbers of recordings stored in audio archives is not feasible due to the required manual intervention for each individual recording, necessitating considerable time and effort. The number of audio recordings stored in archives around the globe is immense: the Library of Congress alone reports more than 3.5 million audio materials in 2014 [14], 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 and SNRs—from granular sounding noise at SNRs of approximately 30 dB for old wax cylinder recordings, via tape media at SNRs of approximately 60 dB, to practically noise-free recordings of today. If the restoration of large numbers of audio recordings 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., [15–18], have been proposed to enhance noisy speech signals in communication applications such as conferencing systems or hearing aids (cf., Sec. 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 generic audio archives the input signal may be 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, e.g., cars passing by or other people talking in the background. In contrast, the noise in audio archive recordings is usually assumed to be of a technical nature and rather constant for each individual recording. It is caused, e.g., by limited dynamic ranges of recording media or aging effects of the carrier material. In the audio archive context, acoustic noise is typically regarded as part of the audio 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.

0.1 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 that 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 that 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, therefore whether they should be processed by a broadband noise reduction algorithm or not.

0.2 Related Work

During the last decades the reduction of broadband noise has received steady research attention, mainly however for speech communication applications [19–22]. The earliest broadband noise reduction approach is probably described in a patent from 1965 [23]. 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) [24–26] to determine noise-only sections, based on which the noise PSD was estimated by averaging short-time periodograms, e.g., using the Welch method [27]. Obviously, VAD-based noise PSD estimation algorithms perform poorly when no pauses are detected in the desired signal over a longer period. This holds true especially for music signals where pauses are typically scarce. Furthermore, 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 [16]. 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 [15], which estimates the noise PSD by tracking minima of the noisy input PSD within a certain time window. Since the minimum of the noisy input PSD within the time window is smaller than the sought-after mean value of the PSD, a bias compensation factor is introduced for the minimum statistics algorithm. 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 [16, 17].

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 speech signals typically contain frequent pauses and a high noise PSD estimation accuracy can be obtained for these signals, severe noise PSD overestimation may occur if the desired signal contains only very few pauses or does not contain pauses at all, e.g., for music signals (cf., Sec. 4.5.2).

Although most noise reduction algorithms were originally designed to enhance speech signals, they are often successfully applied to diverse audio recordings [1, 2, 28]. However, the problem of noise PSD estimation for signals different from speech has only been treated marginally. In [29] an automatic method that estimates 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 [18] 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., [30], 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 according to the characteristics and the SNR of the input signal [31].

0.3 Paper Structure

This paper is structured as follows: Sec. 2 describes the signal model and the assumed distribution of the periodogram coefficients of the noise disturbance. In Sec. 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 Sec. 4.

1 SIGNAL MODEL

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

$$x[n] = s[n] + d[n], \quad \text{for } 0 \le n < L, \tag{1}$$

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 uncorrelated with the audio signal and stationary over the complete duration L of the recording. 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 [32] 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 [32] 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. (1) is given by

$$X[k,\, l] = S[k,\, l] + D[k,\, l], \quad \text{for } 0 \le k < N,\ 0 \le l < M, \tag{2}$$

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.,

$$X\,[k,\, l\,] = \sum_{n=0}^{N-1} w[n]x[lN + n] \cdot e^{-j2\pi kn/N}, \tag{3}$$

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.,

$$\mathrm{Re}\,\{D[k,\, l]\}\,,\ \mathrm{Im}\,\{D[k,\, l]\} \sim \mathcal{N}\left(0,\, \sigma^2\,[k]\right), \quad \text{for } 0 \le k < N, \tag{4}$$

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 [32], the squared magnitudes of the STFT coefficients $D[k, l]$, i.e., the short-time periodograms $P_d[k, l] = |D[k, l]|^2$ follow an exponential distribution, which is defined via its probability density function (PDF) [33, p. 85]:

$$f_{\mathrm{e}}(x;\, \mu) = \begin{cases} \frac{1}{\mu} e^{-\frac{x}{\mu}} & \text{for } x \ge 0 \\ 0 & \text{else} \end{cases}, \tag{5}$$

with mean $\mu > 0$.

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

$$P_s\,[k,\, l\,] = |S\,[k,\, l\,]|^2 \tag{6}$$

are zero or at least much smaller than the corresponding noise periodogram coefficients $P_d[k, l]$. Simulation results in Sec. 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.

2 NOISE PSD ESTIMATION ALGORITHM

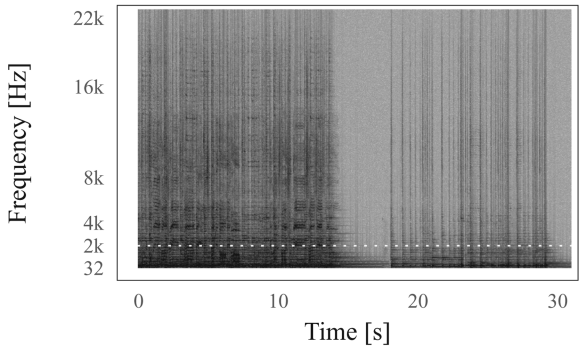
2.1 Overall Procedure

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.

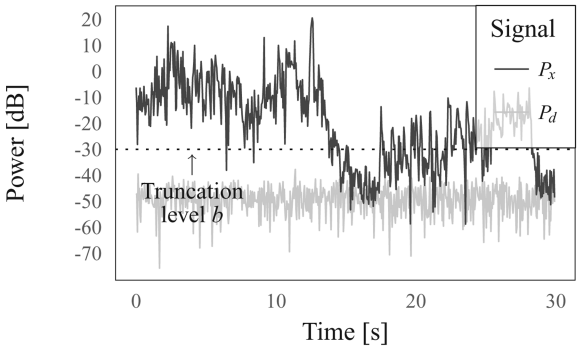
Fig. 1 shows four diagrams that illustrate the intermediate steps of the proposed algorithm. First, the short-time periodogram coefficients of the disturbed input signal

$$P_x[k,\, l\,] = |X\,[k,\, l\,]|^2 \tag{7}$$

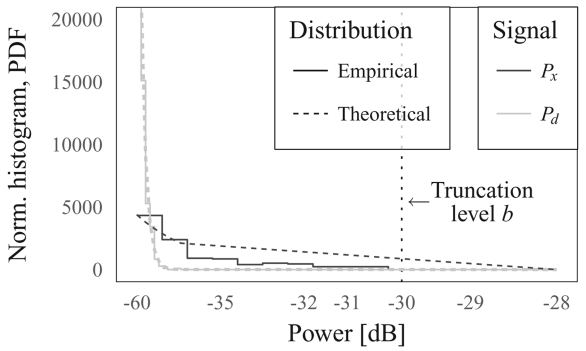
[1] In our experiments we used signals with a length of 30 s.



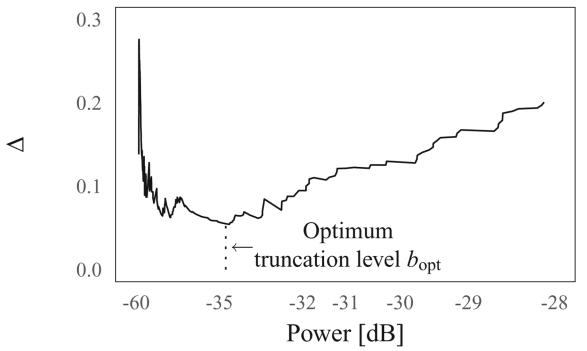
(a) Spectrogram of the input signal x. The horizontal dotted line indicates a frequency of approximately 2 kHz.



(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 47dB. The horizontal dotted line indicates an example truncation level $b = -30\,\mathrm{dB}$.



(c) Normalized histograms of the periodogram coefficients and PDFs of the assumed distributions. The vertical dotted line indicates an example truncation level $b = -30\,\mathrm{dB}$. The x-axis uses logarithmic ticks and 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}.

Fig. 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.

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, Fig. 1(a) depicts the spectrogram of the input signal. For an example frequency of approximately 2 kHz, Fig. 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, for each frequency, the power level 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

$$\mathbf{P}_b = \{P_x[l] : 0 \leq l < M,\ P_x[l] \leq b\}, \tag{8}$$

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 is greater than or equal to a minimum size $M_{\min}$, i.e., $Q \geqslant M_{\min}$ (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 [34]

$$F_b(x) = \frac{1}{Q} \sum_{i=0}^{Q-1} I_{P_b[i] \leq x}, \tag{9}$$

with the indicator function

$$I_{P_b[i] \leq x} = \begin{cases} 1 & \text{for } P_b[i] \leq x \\ 0 & \text{else} \end{cases}. \tag{10}$$

As mentioned in Sec. 2, the periodogram coefficients of the noise disturbance are assumed to follow an exponential distribution, cf., Eq. (5). 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

$$f_{\text{te}}(x;\, \mu,\, b) = \begin{cases} \frac{\frac{1}{\mu} e^{-\frac{x}{\mu}}}{1 - e^{-\frac{b}{\mu}}} & \text{for } 0 \leq x \leq b \\ 0 & \text{else} \end{cases} \tag{11}$$

$$F_{\text{te}}(x;\, \mu,\, b) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{e^{\frac{b}{\mu}}}{e^{\frac{b}{\mu}} - 1} \left(1 - e^{-\frac{x}{\mu}}\right) & \text{for } 0 \leq x \leq b \\ 1 & \text{for } x > b \end{cases}. \tag{12}$$

The corresponding normalized histogram is denoted as $\mathbf{H}_{\text{te}}(\mu,\, b)$ and has 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., Sec. 3.2. The optimal truncation level leading to the minimum distance is denoted as b_{opt}. The corresponding parameter $\hat{\mu}\left(b_{\text{opt}}\right)$ of the truncated exponential distribution is used as the noise PSD estimate $\hat{\sigma}^2$.

For an example value of the truncation level ($b = -30$ dB), Fig. 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). Fig. 1(d) shows the distance measure (the normalized total absolute difference, cf., Sec. 3.2) as a function of the truncation level b, indicating the optimal value b_{opt}.

2.2 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 the Kolmogorov-Smirnov test [35], the Anderson-Darling test [36], or the chi-squared test [35]. On the other hand, distance measures between probability distributions such as the Kullback-Leibler (KL) or Jensen-Shannon (JS) divergence [37] could be used.

In the context of our application, the measure should fulfill two properties: (1) independence of the sample size, which is proportional 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 [38], 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 [37], 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

$$\Delta_{\text{JS}}(\mathbf{H}_b,\, \mathbf{H}_{\text{te}}) = \frac{1}{2}\left[\Delta_{\text{KL}}\left(\mathbf{H}_b,\, \frac{1}{2}(\mathbf{H}_b + \mathbf{H}_{\text{te}})\right) + \Delta_{\text{KL}}\left(\frac{1}{2}(\mathbf{H}_b + \mathbf{H}_{\text{te}}),\, \mathbf{H}_{\text{te}}\right)\right], \tag{13}$$

with the KL divergence between two histograms $\mathbf{H}_1$ and $\mathbf{H}_2$ defined as [37]

$$\Delta_{\text{KL}}(\mathbf{H}_1,\, \mathbf{H}_2) = \sum_{i=0}^{B-1} H_1[i] \cdot \log_2 \frac{H_1[i]}{H_2[i]}. \tag{14}$$

As the JS divergence is bounded by one ($0 \leqslant \Delta_{\text{JS}} \leqslant 1$) if the binary logarithm is used as in Eq. (14) [37], a confidence value can easily be derived as

$$C_{\text{JS}} = 1 - \Delta_{\text{JS}}. \tag{15}$$

As a second option, we propose to use the normalized total absolute difference (AD) between the empirical CDF in Eq. (9) and the CDF of the truncated exponential distribution in Eq. (12) as a simple and intuitive distance measure. The (unnormalized) total AD is given by

$$\text{AD} = \sum_{i=0}^{Q-1} \left|F_b(P_b[i]) - F_{\text{te}}(P_b[i];\, \mu,\, b)\right|, \tag{16}$$

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
    $\mathbf{P}_x \leftarrow \text{SORT}(\mathbf{P}_x)$
    for $l = M_{\min} \ldots M$ **do**
      $\mathbf{P}_b \leftarrow [P_x[0], \ldots, P_x[l-1]]$
      $b \leftarrow P_x[l-1]$ ▷ Truncation level
      $[\hat{\mu}, \Delta] \leftarrow \text{MINIMIZEDISTANCE}(\mathbf{P}_b, b)$
      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 $\hat{\sigma}^2$, $\mathbf{C}$
end Procedure

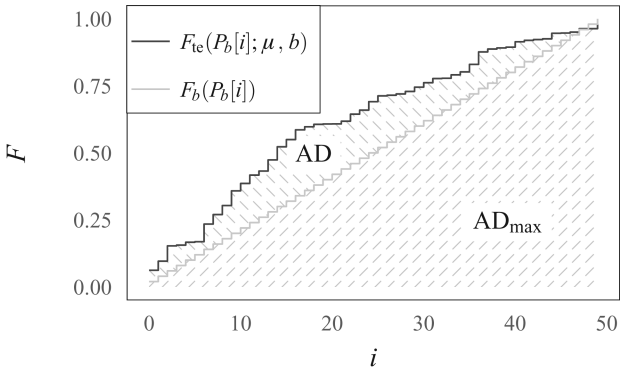


Fig. 2. Absolute difference (AD) and AD_{max} for example values of $\mathbf{P}_b$, μ and b. In this example $Q = 50$.

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

$$\text{AD}_{\max} = \frac{Q}{2}, \tag{17}$$

which corresponds to $F_{\text{te}}(P_b[i]; \mu, b)$ being equal to either 0 or 1 for all i. Fig. 2 shows AD and AD_{max} for example values of $\mathbf{P}_b$, μ, and b where AD is the area between $F_{\text{te}}(P_b[i]; \mu, b)$ and $F_b(P_b[i])$, and AD_{max} is the area between 0 and $F_b(P_b[i])$. In this example, $F_{\text{te}}(P_b[i]; \mu, b) > F_b(P_b[i])$ 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 AD_{max}, i.e.,

$$\Delta_{\text{AD}} = \frac{\text{AD}}{\text{AD}_{\max}}, \tag{18}$$

such that similarly as in Eq. (15) a confidence value can be easily derived as

$$C_{\text{AD}} = 1 - \Delta_{\text{AD}}. \tag{19}$$

The confidence values in Eqs. (15) and (19) are a measure for the reliability of the obtained noise PSD estimates. Although it will be shown in Sec. 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 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 that 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.

3 EVALUATION AND RESULTS

We evaluate the proposed noise PSD estimation algorithm using a database of music and speech signals and different types of broadband noise (cf., Sec. 4.1). The noise PSD estimation accuracy is evaluated in a first step (cf., Sec. 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 [20]. The evaluation consists 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 best distance measure with a reference noise PSD estimation algorithm based on minimum statistics (cf., Sec. 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 that of a recently proposed noise reduction algorithm that does not require a noise PSD estimate (cf., Sec. 4.4). The results of the evaluation are presented in Sec. 4.5.

3.1 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 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[3] and
- A recording of real optical film soundtrack noise[4].

All clean signals and noise signals are single-channel[5] and sampled at $f_s = 44.1$ kHz. The length of each recording is 30 s. For all experiments, we used a 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[6] while an overlap of 50 % is used for the minimum statistics algorithm (cf., Sec. 4.3) as specified in [15].

3.2 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.

Table 1. The ODG scale.

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

3.2.1 Noise PSD Estimation Errors

To evaluate the PSD estimation accuracy of the proposed algorithm, we use the error measure proposed in [17], which equally weighs the logarithmic *overestimation* error (LogErrOver) and the logarithmic *underestimation* error (LogErrUnder), i.e.,

$$\text{LogErr} = \text{LogErrOver} + \text{LogErrUnder}, \tag{20}$$

with

$$\text{LogErrOver} = \frac{1}{N_{\text{bins}}} \sum_{k=0}^{N_{\text{bins}}-1} \left| \min\left(0,\ 10 \cdot \log_{10}\left[\frac{\sigma^2[k]}{\hat{\sigma}^2[k]}\right]\right)\right| \tag{21}$$

$$\text{LogErrUnder} = \frac{1}{N_{\text{bins}}} \sum_{k=0}^{N_{\text{bins}}-1} \left| \max\left(0,\ 10 \cdot \log_{10}\left[\frac{\sigma^2[k]}{\hat{\sigma}^2[k]}\right]\right)\right|, \tag{22}$$

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

3.2.2 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[7] [39–41]. The PEAQ measure aims at determining the perceptual similarity between two audio signals by computing a representation of each signal that takes the human hearing properties into account. From this representation, a so-called *Objective Difference Grade* (ODG) is computed which ranges from –4 ("very annoying") to 0 ("imperceptible"), cf., Table 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 [41]. In addition, the results of a subjective quality evaluation (cf., Sec. 4.2.3) indicate that the ODG ratings obtained with the PEAQ measure generally correspond well with the subjective impression. A selection of the audio signals that were used for the evaluation and the

[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 at https://matbra.github.io/noise_psd_estimation.

[3] Available online at https://freesound.org/people/monotraum/sounds/242209.

[4] Available online at https://freesound.org/people/Yuval/sounds/197795.

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

[6] 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.

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

corresponding ODG ratings are available on the website accompanying this paper (please refer to the corresponding footnote in 3.1).

3.2.3 Subjective Quality Evaluation

As a supplement to the instrumental PEAQ measure, we conducted a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) listening test [42] to evaluate the perceived audio quality of the processed input signal. Participants of the listening test were 10 normal-hearing and trained listeners. In order to obtain comparability with the PEAQ ratings, they were asked to rate the similarity of the processed signals to the undisturbed signal on a scale from 0 to 100. The signals used in the test were based on three randomly selected speech signals and three randomly selected music signals from the test signals described in Sec. 4.1, each at SNRs of 20 dB, 40 dB, and 60 dB. In order to limit the duration of the listening test and avoid fatigue, we only used the recording of real tape noise as the noise disturbance.

The statistical evaluation of the listening test ratings consists in a Repeated Measures Analysis of Variance (ANOVA) [43] to determine whether the differences between the algorithm ratings are statistically significant followed by Wilcoxon signed-rank tests [44] to investigate rating differences between all pairs of algorithms.

3.3 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 [15] (cf., Sec. 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 [15] 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 s (typically used in speech communication applications) and the maximum value of 3.7 s specified in [15].

3.4 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 [20]. We use the implementation and parameter values from [13]. In addition, we use a recently proposed noise reduction algorithm based on structured sparsity [30], 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 to –20 dB in all experiments [19].

3.5 Results

This section presents the results of three experiments to determine the best distance measure (Sec. 4.5.1), the noise PSD estimation accuracy with this distance measure (Sec. 4.5.2), and the noise reduction performance when using the proposed noise PSD estimate in the MMSE STSA noise reduction algorithm (Sec. 4.5.3). In Sec. 4.5.1 we use 50 music and 50 speech recordings. In Secs. 4.5.2 and 4.5.3 we use 500 music and 500 speech recordings for the evaluation of the noise PSD estimation error and the noise reduction quality with the PEAQ measure, and three music and three speech signals for the subjective noise reduction quality evaluation.

3.5.1 Distance Measure

In this section we analyze the noise PSD estimation accuracy of the proposed algorithm for both considered distance measures (cf., Sec. 3.2), i.e., the *JS divergence* and the *normalized total AD*. Fig. 3 shows the LogErr measure in Eq. (20) for both distance measures and 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., Sec. 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.

3.5.2 Noise PSD Estimation Accuracy

Using the optimum distance measure determined in the previous 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., Sec. 4.3).

Fig. 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

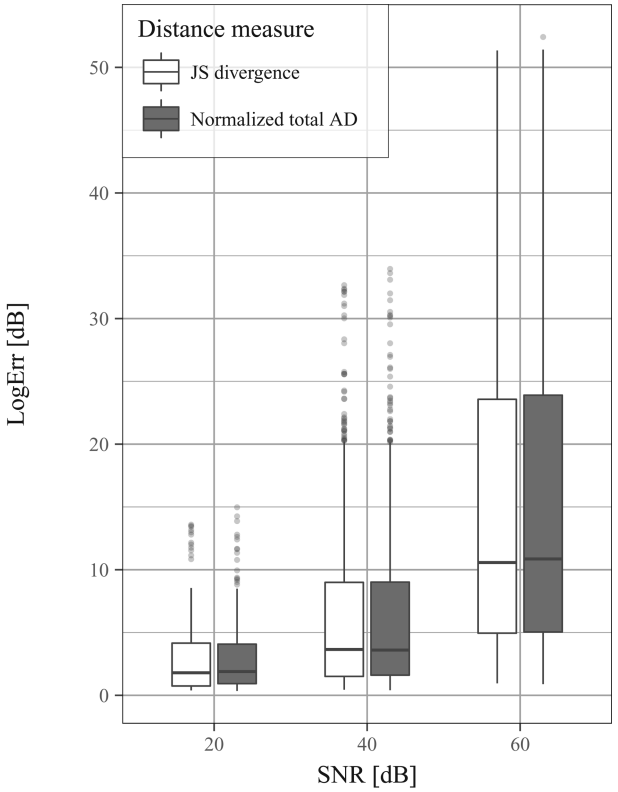


Fig. 3. Noise PSD estimation errors of the proposed algorithm for different SNRs for both distance measures. The lower and upper edge of each box indicates the first and the third quartile of the data, respectively, while the horizontal line in each box corresponds to the median value. Vertical lines 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.

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 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 [3, 4].

It should be noted that the confidence value C_{AD} was not considered further 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. (19)), however low that confidence is. Fig. 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 Fig. 4, this can probably be explained by the amount of hum and impulsive disturbances in the film noise. The results in Fig. 5 indicate that the confidence value allows to distinguish severely disturbed from weakly disturbed input signals in most cases, i.e., SNR $\leqslant$ 30 dB from SNR $\geqslant$ 50 dB.

3.5.3 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 on the perceptual audio quality of using a minimum required confidence (MRC), cf., Sec. 3.2, 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.

Fig. 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 Fig. 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

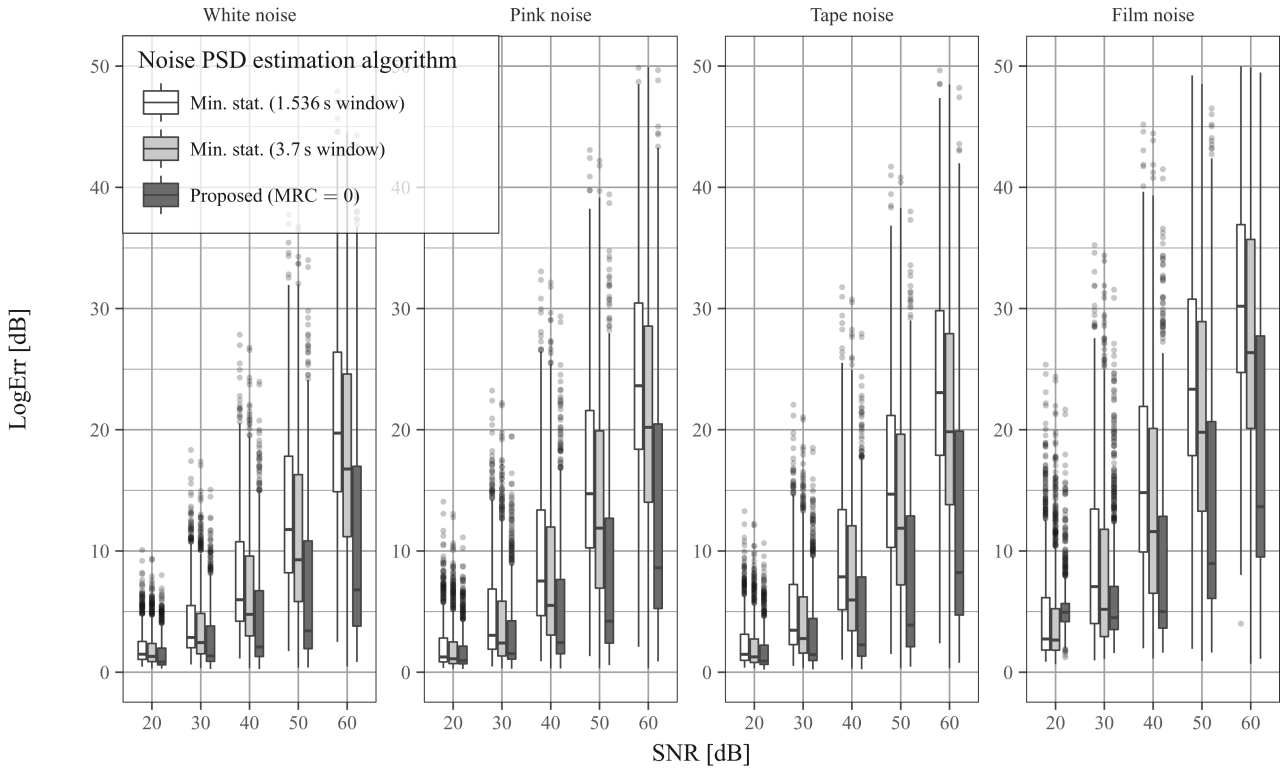


Fig. 4. Noise PSD estimation errors of the proposed algorithm using the normalized total AD distance measure and the minimum statistics algorithm for two search window lengths. This figure integrates the results for all speech and music signals. The results are shown separately for each noise type.

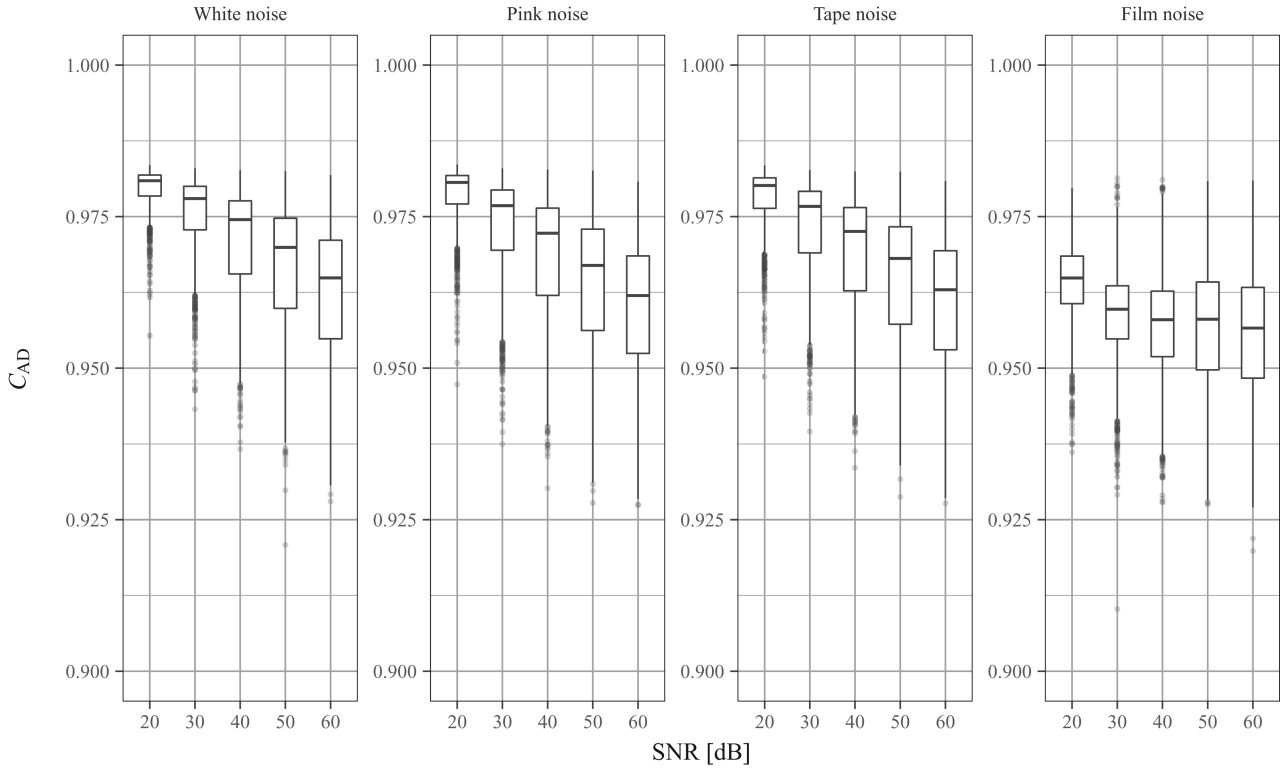


Fig. 5. Confidence values of the proposed algorithm, averaged over all frequencies for each disturbed signal. This figure integrates the results for all speech and music signals. The results are shown separately for each noise type.

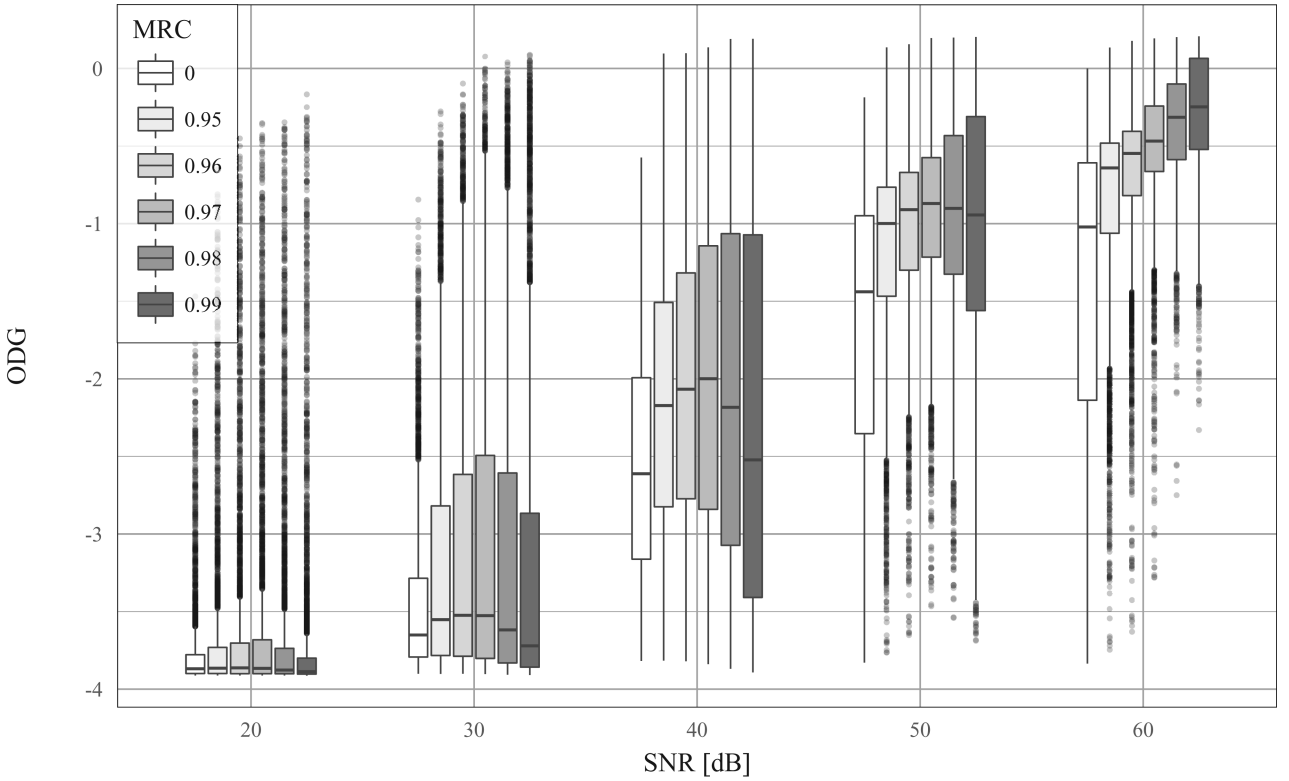


Fig. 6. Instrumental audio quality evaluation of the processed signals obtained by combining the proposed noise PSD estimate with the MMSE STSA noise reduction algorithm for different MRC values. This figure integrates the results for all speech and music signals and all noise types.

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 $\in \{0.97, 0.98\}$.

Fig. 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., Sec. 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 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. 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 that is presumably the result of fixed algorithm parameters that are not adjusted in dependence on the SNR, as already pointed out in [31].

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.

Table 2 summarizes the PEAQ results and lists the number of signals whose quality was improved, remained unchanged, or was worsened by the noise reduction processing using different noise PSD estimation algorithms. It can be observed that at low SNRs all algorithms lead to a general quality improvement. Nevertheless, it is evident that the noise PSD estimation algorithm based on minimum statistics leads to a quality reduction for a substantial fraction of

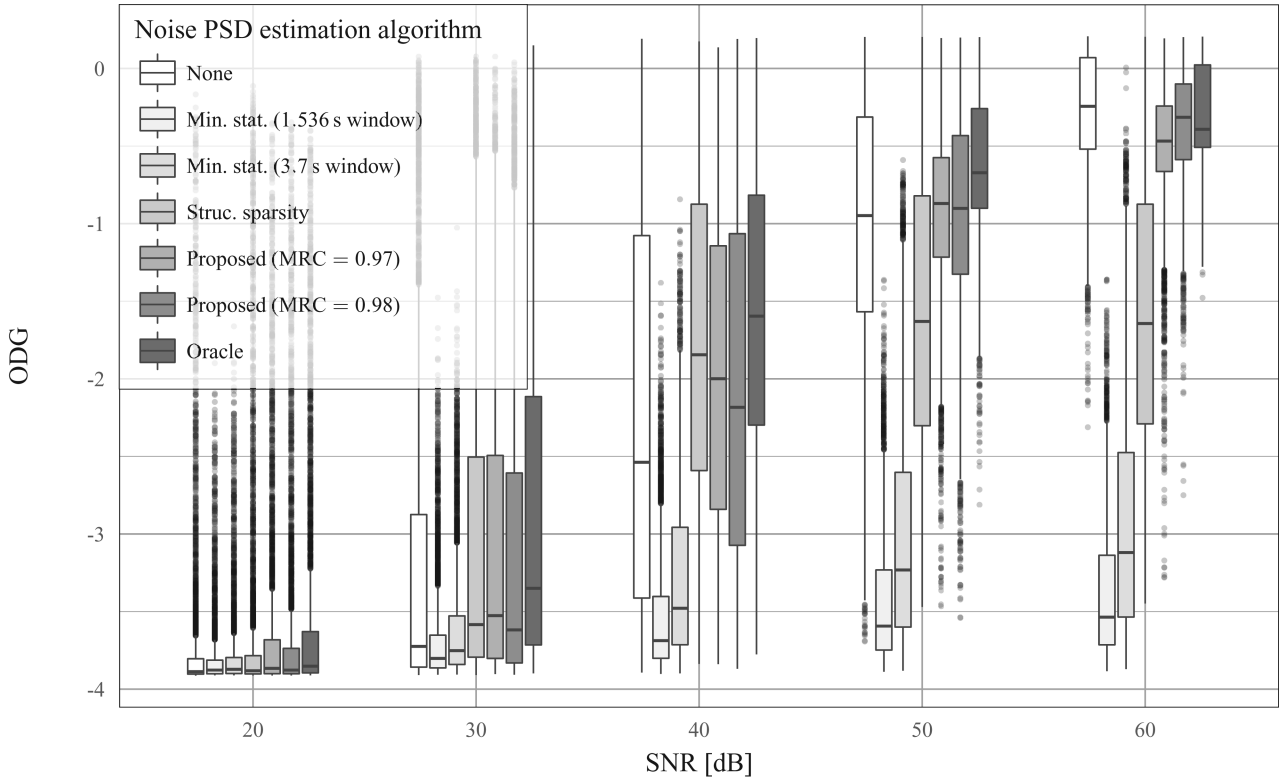


Fig. 7. Instrumental audio quality evaluation of the processed signals obtained by several algorithms: (1) no processing; (2) MMSE STSA noise reduction algorithm using the minimum-statistics-based noise PSD estimate, the proposed PSD estimate and the oracle noise PSD estimate; (3) noise reduction algorithm based on structured sparsity.

the signals. The results also suggest that at an SNR of 20 dB the noise reduction algorithm based on structured sparsity performs no change of the signal at all. (It should be noted that we used fixed algorithm parameters that were adjusted for an overall good performance across all SNRs.) Of all analyzed algorithms, the proposed algorithm with MRC = 0.98 leads to the highest number of signals with improved quality at SNRs of 20 dB, 50 dB, and 60 dB, i.e., at low as well as high SNRs.

The results of the subjective quality evaluation are shown in Fig. 8. In addition to the signals processed with the noise reduction algorithms, a lowpass anchor signal was used that was obtained by filtering the clean signal with a 3.5 kHz lowpass filter. It can be observed that the variation in the ratings of the signals obtained with each noise PSD estimation algorithm is large, covering almost the full range of rating scores for the lowpass anchor signal and the signals obtained with the noise PSD estimation algorithm based on minimum statistics. Furthermore, the median ratings of the unprocessed disturbed input signals ("None") are very high at SNRs of 40 dB and 60 dB, indicating that the majority of signals used in the listening test are perceived as undisturbed at these SNRs. This may be explained by the fact that only the tape noise disturbance was used in the listening test, which is masked by the desired signal at higher SNRs due to its lowpass characteristic. In comparison, the results in Fig. 7 show that the PEAQ ratings of the unprocessed disturbed input signals are generally much lower, especially at SNRs of 20 dB and 40 dB, suggesting that the PEAQ measure penalizes the broadband noise disturbance much more than the listeners. It should be noted that the subjective quality evaluation is able to evaluate the efficiency of the

Table 2. Change of the PEAQ ratings compared to the unprocessed, disturbed input signal for different noise PSD estimation algorithms. For each algorithm and SNR, the number of signals whose rating are improved by the processing (↑), remained unchanged (=), and are worsened (↓) are shown.

Noise PSD	20 dB			30 dB			40 dB			50 dB			60 dB		
Estimation Algorithm	↓	=	↑	↓	=	↑	↓	=	↑	↓	=	↑	↓	=	↑
Min. stat. (1.536 s window)	439	75	486	695	19	286	897	0	103	983	0	17	1000	0	0
Min. stat. (3.7 s window)	371	75	554	565	14	421	732	2	266	900	0	100	987	0	13
Struc. sparsity	8	585	407	13	14	973	39	1	960	539	3	458	942	0	58
Proposed (MRC=0.98)	68	129	803	77	8	915	86	2	912	150	8	842	317	46	637
Proposed (MRC=0.97)	190	110	700	105	9	886	115	0	885	211	1	788	609	9	382
Oracle	90	39	871	23	0	977	52	1	947	122	0	878	623	7	370

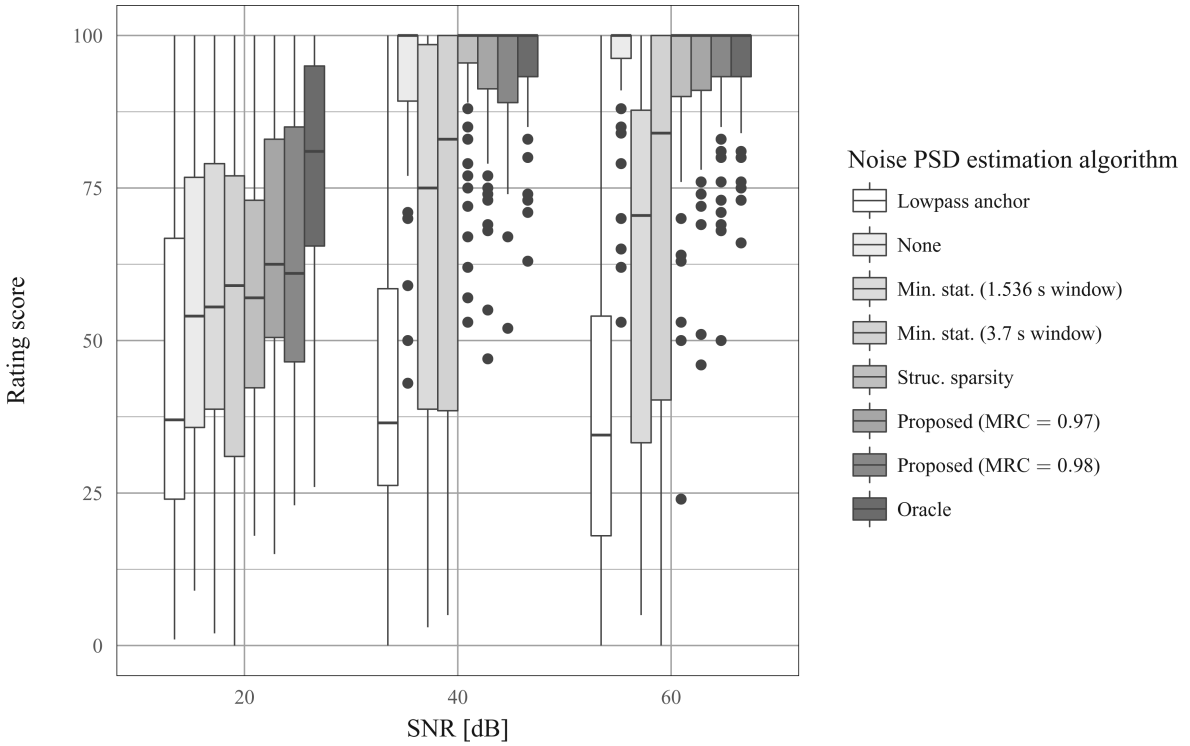


Fig. 8. Subjective audio quality evaluation of the processed signals obtained by several algorithms: (1) no processing; (2) MMSE STSA noise reduction algorithm using the minimum-statistics-based noise PSD estimate, the proposed PSD estimate and the oracle noise PSD estimate; (3) noise reduction algorithm based on structured sparsity.

noise PSD estimation algorithms to a certain extent only. This is due to the fact that only six clean signals were used, compared to 1000 clean signals used for the instrumental evaluation using the PEAQ measure. The results of the subjective quality evaluation are expected to vary significantly in dependence on the test signals, analogous to the large variations observed in the PEAQ ratings in Fig. 7. Still, the results of the subjective quality evaluation in general correspond well with the PEAQ ratings, confirming the validity of the instrumental evaluation: at SNRs above 20 dB, the results obtained with the noise PSD estimation algorithm based on minimum statistics are significantly lower than the results obtained with the noise reduction algorithm based on structured sparsity and the proposed algorithm. The high ratings obtained with the true noise PSD show that efficient noise reduction is possible if an exact estimate of the noise PSD is available.

We analyzed the results of the subjective quality evaluation statistically to investigate rating differences between the algorithms. In doing so, we integrated the subjective ratings for all SNRs. The results of a repeated-measures ANOVA indicate in a first step that the ratings for all algorithms are significantly different, with $p < 0.001$. In a second step, the results of Wilcoxon signed-rank tests shown in Table 3 indicate the statistical significance of differences between all algorithm pairs. The numbers indicate the significance levels at which the null hypothesis can be rejected that the ratings for a pair of algorithms follow the same distribution. It can be observed that the rating differences between almost all pairs of algorithms are significant

Table 3. Results of the Wilcoxon signed-rank tests. The table shows the significance levels at which the null hypothesis that the two respective samples follow the same distribution may be rejected. Bold entries highlight p values >0.05 that suggest that the two respective samples follow the same distribution.

	Lowpass Anchor	None	Min. stat. (1.536 s window)	Min. stat. (3.7 s window)	Struc. sparsity	Proposed (MRC=0.97)	Proposed (MRC=0.98)	Oracle
Lowpass Anchor	–	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
None	–	–	<0.001	<0.001	>**0.05**	<0.05	<0.05	<0.001
Min. stat. (1.536 s window)	–	–	–	<0.01	<0.001	<0.001	<0.001	<0.001
Min. stat. (3.7 s window)	–	–	–	–	<0.001	<0.001	<0.001	<0.001
Struc. sparsity	–	–	–	–	–	<0.05	<0.05	<0.001
Proposed (MRC = 0.97)	–	–	–	–	–	–	>**0.05**	<0.001
Proposed (MRC = 0.98)	–	–	–	–	–	–	–	<0.001

at least at the 5% level. Only the ratings for the unprocessed disturbed input signal ("None") and the noise reduction algorithm based on structured sparsity ("Struc. sparsity"), and the ratings for the MMSE STSA algorithm combined with the proposed noise PSD estimation algorithm with both MRC values are not significantly different. However, in multiple cases, the difference between multiple pairs of noise reduction types is significant although their overall ratings seem similar in Fig. 8. For example, the ratings obtained for the noise reduction algorithm based on structured sparsity and for the MMSE STSA algorithm using the proposed noise PSD estimation algorithm are significantly different with $p < 0.05$ although the overall ratings seem similar. This shows that the signals processed with these two algorithms are rated inconsistently by the listeners, confirming that perceptual quality is a matter of personal taste.

In conclusion, the proposed algorithm with MRC $= 0.98$ is the only algorithm that 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.

4 SUMMARY AND 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 that 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. Noise PSD estimates with a low confidence are rejected 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 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.

5 REFERENCES

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

[2] S. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 4th ed. (Chichester/West Sussex, UK: Wiley, 2008). https://doi.org/10.1002/9780470740156

[3] M. Brandt and J. Bitzer, "Automatic Detection of Hum in Audio Signals," *J Audio Eng. Soc.*, vol. 62, pp. 584–595 (2014 Sep.). https://doi.org/10.17743/jaes.2014.0034

[4] M. Brandt, S. Doclo, T. Gerkmann, and J. Bitzer, "Impulsive Disturbances in Audio Archives: Signal Classification for Automatic Restoration," *J. Audio Eng. Soc.*, vol. 65, pp. 826–840 (2017 Oct.). https://doi.org/10.17743/jaes.2017.0032

[5] J. B. Minter, "Recent Developments in Precision Master Recording Lathes," *J. Audio Eng. Soc.*, vol. 4, no. 2, pp. 50–55 (1956 Apr.).

[6] W. R. Isom, "Record Materials, Part II: Evolution of the Disc Talking Machine," *J. Audio Eng. Soc.*, vol. 25, pp. 718–723 (1977 Oct./Nov.).

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

[8] B. Fries and 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).

[9] D. Schüller, "The Ethics of Preservation, Restoration, and Re-Issues of Historical Sound Recordings," *J. Audio Eng. Soc.*, vol. 39, pp. 1014–1017 (1991 Dec.).

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

[11] F. Bressan, S. Canazza, and 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.).

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

[13] J. Nuzman, "Audio Restoration: An Investigation of Digital Methods for Click Removal and Hiss Reduction" (2004 Jan.). https://github.com/jnuzman/audio-restoration-2004

[14] Library of Congress, *Annual Report of the Librarian of Congress*, Washington, D.C., USA (2014).

[15] 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

[16] I. Cohen and 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

[17] T. Gerkmann and 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

[18] S. Rangachari and 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

[19] M. Berouti, R. Schwartz, and 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

[20] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121 (1984 Dec.). https://doi.org/10.1109/TASSP.1984.1164453

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

[22] R. C. Hendriks, T. Gerkmann, and 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

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

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

[25] F. Heese, M. Niermann, and P. Vary, "Speech-Codebook Based Soft Voice Activity Detection," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, pp. 4335–4339 (2015 Apr.). https://doi.org/10.1109/ICASSP.2015.7178789

[26] X. Li, R. Horaud, L. Girin, and 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 Sep.).

[27] 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

[28] O. Cappé and 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

[29] G. García, *Automatic Denoising for Musical Audio Restoration*, Ph.D. thesis, Stanford University (2009).

[30] K. Siedenburg and M. Dörfler, "Persistent Time-Frequency Shrinkage for Audio Denoising," *J. Audio Eng. Soc.*, vol. 61, pp. 29–38 (2013 Jan./Feb.).

[31] V. Mach, "Denoising Phonogram Cylinders Recordings Using Structured Sparsity," *Proceedings of the International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Brno, Czech Republic, pp. 314–319 (2015 Oct.). https://doi.org/10.1109/ICUMT.2015.7382449

[32] J. Schoukens and 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 Sep.). https://doi.org/10.1109/TIM.1986.6499210

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

[34] A. W. van der Vaart, *Asymptotic Statistics* (Cambridge, UK: Cambridge University Press, 1998). https://doi.org/10.1017/CBO9780511802256

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

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

[37] 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

[38] N. M. Razali and Y. B. Wah, "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests," *J. Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21–33 (2011 June).

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

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

[41] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "PEAQ—The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.*, vol. 48, pp. 3–29 (2000 Jan./Feb.).

[42] International Telecommunication Union, ″Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," Recommendation BS.1534-3, ITU-R (2015 Oct.).

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

[44] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83 (1945 Dec.). https://doi.org/10.2307/3001968

THE AUTHORS

Matthias Brandt

Simon Doclo

Joerg Bitzer

Matthias Brandt was born in Bremen, Germany, in 1980. He received his diploma in electrical engineering in 2008 from the University of Bremen, Germany. From 2009 to 2012 he was employed at the Jade University of Applied Sciences Oldenburg, Germany. From 2013 to 2018 he was an early-stage researcher at the University of Oldenburg, Germany, and received his Ph.D. in the field of audio restoration in 2018. His research focuses on the processing of music signals, in particular their automatic restoration, and related information retrieval.

•

Prof. Dr. Simon Doclo received the M.Sc. degree in electrical engineering and the Ph.D. degree in applied sciences from the Katholieke Universiteit Leuven, Belgium, in 1997 and 2003. From 2003 to 2007 he was a Postdoctoral Fellow with the Research Foundation – Flanders at the Electrical Engineering Department (Katholieke Universiteit Leuven) and the Cognitive Systems Laboratory (McMaster University, Canada). From 2007 to 2009 he was a Principal Scientist with NXP Semiconductors at the Sound and Acoustics Group in Leuven, Belgium. Since 2009 he is a full professor at the University of Oldenburg, Germany, and scientific advisor for the project group Hearing, Speech and Audio Technology of the Fraunhofer Institute for Digital Media Technology. His research activities center around signal processing for acoustical and biomedical applications, more specifically microphone array processing, speech enhancement, active noise control, acoustic sensor networks, and hearing aid processing. Prof. Doclo received the Master Thesis Award of the Royal Flemish Society of Engineers in 1997, the Best Student Paper Award at the International Workshop on Acoustic Echo and Noise Control in 2001, the EURASIP Signal Processing Best Paper Award in 2003, and the IEEE Signal Processing Society 2008 Best Paper Award. He is member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing, the EURASIP Special Area Team on Acoustic, Speech and Music Signal Processing, and the EAA Technical Committee on Audio Signal Processing. Prof. Doclo was and is involved in several large-scale national and European research projects (ITN DREAMS, Cluster of Excellence Hearing4All, CRC Hearing Acoustics). He was Technical Program Chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2013 and Chair of the ITG Conference on Speech Communication in 2018. In addition, he served as guest editor for several special issues (*IEEE Signal Processing Magazine, Elsevier Signal Processing*) and is associate editor for I*EEE/ACM Transactions on Audio, Speech and Language Processing* and *EURASIP Journal on Advances in Signal Processing*.

•

Joerg Bitzer was born in Bremen in 1970. He received his diploma in 1995 and his doctorate in electrical engineering in 2002 from the University of Bremen where he also worked as a research assistant until 1999. From 2000 to 2003 he was head of the algorithm development team at Houpert Digital Audio, a company specialized in audio signal processing. Since September 2003 he is a professor for audio signal processing at the Jade University of Applied Science Wilhelmshaven/Oldenburg/Elsfleth. In 2010 he joined the Fraunhofer project group for hearing, speech, and audio technology in Oldenburg as a scientific supervisor. His current research interests include all forms of single- and multichannel speech enhancement, audio restoration, audio effects for musical applications, and information retrieval for large media archives.