Journal of the Audio Engineering Society
Vol. 65, No. 10, October 2017 (© 2017)
DOI: https://doi.org/10.17743/jaes.2017.0032

Impulsive Disturbances in Audio Archives: Signal Classification for Automatic Restoration

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

TIMO GERKMANN,[2] *AES Member*, **AND JOERG BITZER**,[3] *AES Member*

[1] *University of Oldenburg, Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Oldenburg, Germany*
[2] *University of Hamburg, Dept. of Informatics, Signal Processing Group, Hamburg, Germany*
[3] *Jade University of Applied Sciences, Oldenburg, Germany*

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.

0 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 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 [1]. Millions of further recordings are stored in a multitude of archives in the United States alone [26]. 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 contain 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 Sec. 3.4.3 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.

0.1 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 ≈ 23 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 Fig. 1, each stage with its associated signals and notation.

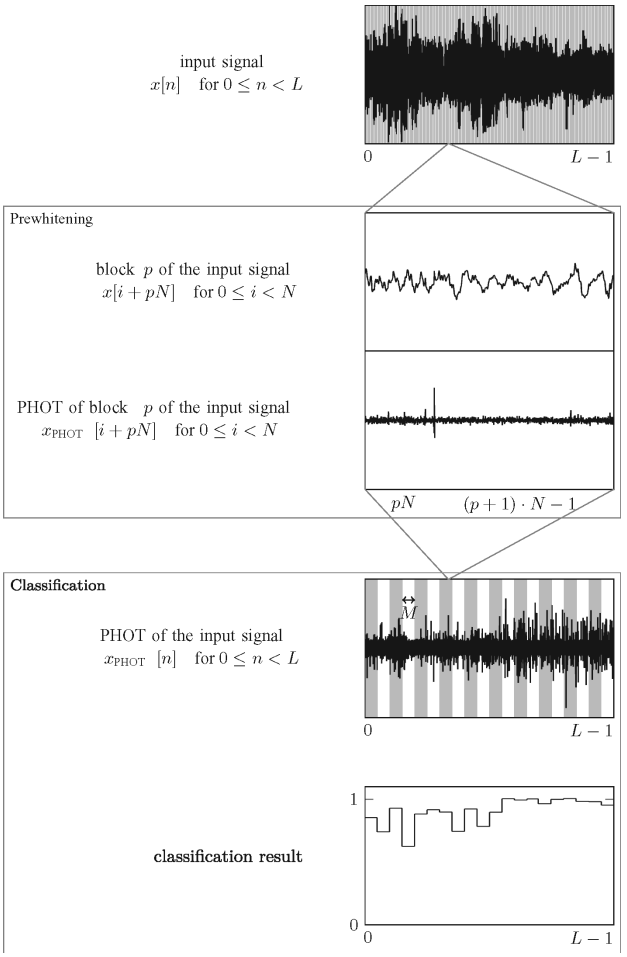

Fig. 1. Schematic overview of the classification algorithm. The prewhitening and classification stages are shown with associated signal notation, block, and frame lengths.

0.2 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 preemphasis [22] or on subtracting the median filtered version from the input signal [37]. Early interpolators consisted in replacing the damaged part of the signal with silence or linear interpolation of the neighboring sample values [37]. Restoration methods based on linear prediction, introduced in [18, 47, 45, 43], 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 [21] or frequency-warped prediction achieve high audio quality even for gap lengths of around 45 ms [20, 11]. 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-channel approach proposed in [14] gains advantage from using two signals obtained with a stereo replay cartridge, compared to only single-channel processing. Other recent methods use bidirectional processing [29] or click templates [30] to increase the detection accuracy. In [8, 6, 7] 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

[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.

audio signal processing tasks, e.g., audio tagging [51], or acoustic event detection [25]. 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.

0.3 Paper Structure

The structure of this paper is as follows. In Sec. 1 the characteristics of the signals to be processed are described. A thorough explanation of the proposed classification algorithm is given in Sec. 2. To analyze the performance of the proposed algorithm, the evaluation method and the results for a large number of test signals are given in Sec. 3.

1 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 44100 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.,

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

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 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 [4] 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 Sec. 1.1 the used model for the artificially generated disturbances will be reviewed, while in Sec. 1.2 two ways to set the SNR will be discussed.

1.1 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 [44]). 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 [49, 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.,

$$f(\tau; k, \Theta) = \frac{1}{\Theta^k \cdot \Gamma(k)} \cdot \tau^{k-1} e^{-\frac{\tau}{\Theta}} \quad \text{for } \tau > 0 \text{ and } k,\ \Theta > 0, \tag{2}$$

with shape parameter k, scale parameter Θ and $\Gamma(\cdot)$ the gamma function (see [32]). The magnitude A of the impulses is modeled with a log-normal distribution, i.e.,

$$f(A; \mu, \sigma) = \frac{1}{A\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(A) - \mu)^2}{2\sigma^2}\right) \quad \text{for } A > 0 \text{ and } \sigma > 0, \tag{3}$$

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 Eq. (2). The individual impulses are scaled according to the log-normal distribution in Eq. (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 [49]. 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.

Table 1. Default parameters of the impulsive disturbance generation method from [49]. The values related to the inter-occurrence time hold for a sampling rate of $f_s = 44100$ Hz.

	Parameter		
	Symbol	Description	Value
Gamma distribution (inter-occurrence time)	k	Shape	0.2
	Θ	Scale	2433.8
Log-normal distribution (impulse magnitude)	μ	Location	-3.63
	σ	Scale	0.74

1.2 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 [40] 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 [49] (see Table 1), where only the gain is adjusted to obtain the desired SNR. In the second case, the scale parameter of the gamma distribution in Eq. (2) is adjusted to obtain the desired SNR. Changing the scale parameter has the effect of changing the average time between impulses. Signals demonstrating the two characteristics of the disturbances are available online on the website accompanying this article [4].

1.2.1 SNR via Gain Factor

In this case the default disturbance signal, $d_{\text{def}}[n]$, generated with the default parameters from [49], 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 $s[n]$.

1.2.2 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 $\text{SNR}_{\text{def}} = 30$ 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 Eq. (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 $\Delta_{\text{SNR}} = 0.1$ dB. The appendix at the end of this paper contains a table of the mean shape parameters required to obtain different SNRs.

2 CLASSIFICATION ALGORITHM

The complete impulsive disturbance classification algorithm is shown in Fig. 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., Sec. 2.1). 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., Sec. 2.2). Using these features as input data, a classifier is trained to determine the class of each frame of the input signal (cf., Sec. 2.3). 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.

2.1 Prewhitening

In many cases impulsive disturbances are audible even if their amplitude is very low. As a consequence, it may be a

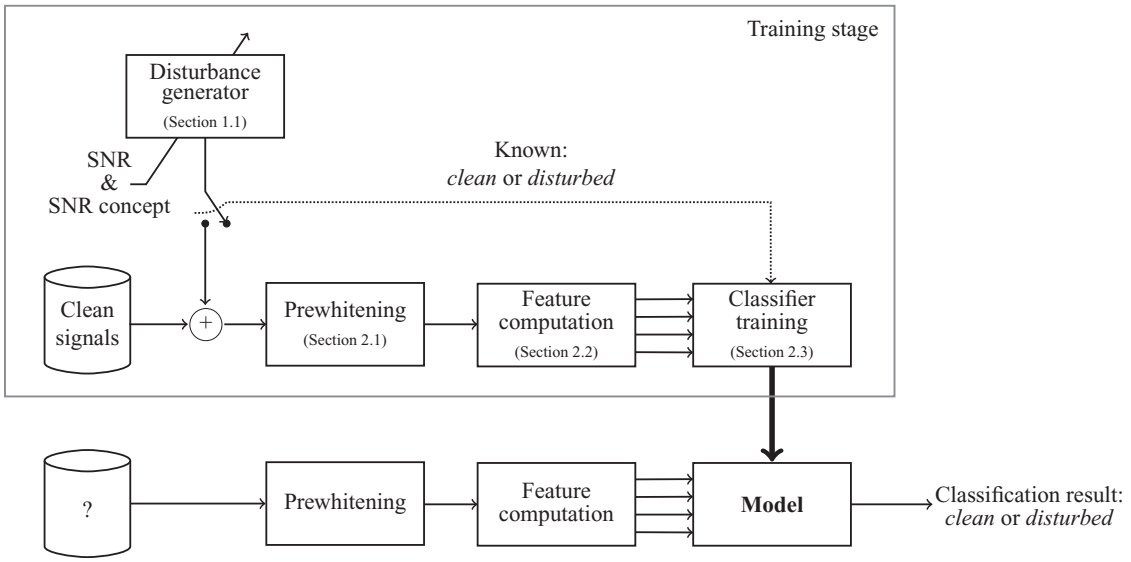

Fig. 2. Flow diagram of the impulsive disturbance classification system.

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., [44, Ch. 13]). The most common type of prewhitening is to use the prediction error signal of a *linear predictor*, which is briefly reviewed in Sec. 2.1.1. However, since for impulsive disturbance classification we found that prewhitening based on linear prediction performs only suboptimally (see the evaluation results in Sec. 3.4.1), we also investigated *phase-only transform (PHOT)* prewhitening, which is described in Sec. 2.1.2.

2.1.1 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 [45, 46, 36].

In forward linear prediction (see, e.g., [35]) the current sample is modeled as a linear combination of previous samples, i.e.,

$$\hat{x}\,[n] = -\sum_{i=1}^{P_{\mathrm{LP}}} a\,[i]\,x\,[n-i] + e\,[n]\,, \tag{4}$$

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:

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

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.

2.1.2 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 [23, 9] and for surface defect detection in images [2]. It is computed for the pth block of the input signal x defined in Eq. (1) as follows:

$$X^{(p)}[k] = \sum_{i=0}^{N-1} x[i+pN]\cdot e^{-j2\pi ki/N} \tag{5a}$$

$$X^{(p)}_{\mathrm{PHOT}}[k] = \frac{X^{(p)}[k]}{\left|X^{(p)}[k]\right|} \tag{5b}$$

$$x_{\mathrm{PHOT}}[n+pN] = \frac{1}{N}\sum_{i=0}^{N-1} X^{(p)}_{\mathrm{PHOT}}[i]\cdot e^{j2\pi ni/N} \tag{5c}$$

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 [5].

The reason why the PHOT enhances transients can be illustrated intuitively. The spectral magnitude of music signals usually decays with higher frequencies [48]—Fig. 3 shows the mean power spectral density of music signals from several decades of the 20th century. The PHOT in

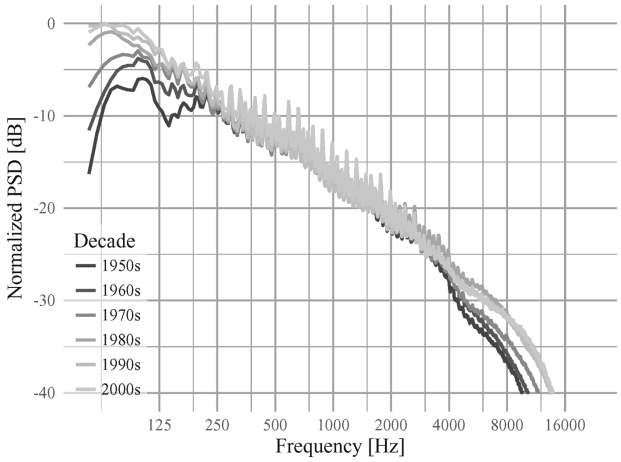

Fig. 3. Mean power spectral density of music signals from several decades of the 20th century. This figure has been generated from the database described in Sec. 3.1. The PSDs were estimated using the Welch method and were normalized such that the overall maximum value is 0 dB. The PSD axis is clipped at –40 dB for reasons of clarity.

Eq. (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 impulses compared to the audio signal. A more thorough examination of why the phase only transform makes irregularities stand out is given in [2].

2.2 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:

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

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 $M = 44100$ samples at a sampling rate of $f_s = 44100$ 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* [17]:

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

where $\sigma_{x_{\text{pre}}^{(q)},\,5\%}$ is the standard deviation of $x_{\text{pre}}^{(q)}$ 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 Sec. A.2). Using recursive feature elimination [13], we found that good performance can be achieved using the *crest factor*, i.e.,

$$C^{(q)} = \frac{\max\limits_{0 \le i < M} \left| x_{\text{pre}}'^{(q)}[n] \right|}{\sqrt{\frac{1}{M} \sum_{i=0}^{M-1} \left(x_{\text{pre}}'^{(q)}[i] \right)^2}}, \tag{6}$$

and the *sample kurtosis*,

$$\text{Kurt}^{(q)} = \frac{\frac{1}{M} \sum_{i=0}^{M-1} \left(x_{\text{pre}}'^{(q)}[i] - \overline{x_{\text{pre}}'^{(q)}} \right)^4}{\left(\frac{1}{M} \sum_{i=0}^{M-1} \left(x_{\text{pre}}'^{(q)}[i] - \overline{x_{\text{pre}}'^{(q)}} \right)^2 \right)^2}, \tag{7}$$

which are both relatively easy to compute.

2.3 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., Sec. 1.1) 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 [33]. The optimal hyperparameters (amount of regularization for logistic regression and SVM and kernel coefficient for SVM) were determined via 5-fold cross-validation [3]. Depending on the specific evaluation goal (cf., Sec. 3.4) 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.

3 EVALUATION METHOD AND 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., Sec. 3.4.1, 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., Sec. 3.4.2, 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., Sec. 3.4.3, 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 $M = 44100$ samples, corresponding to a frame duration of 1 s at the used sampling rate of $f_s = 44100$ Hz.

In Sec. 3.1 we describe the database of test signals. After that, Sec. 3.2 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 Sec. 3.3 we briefly describe the three reference impulse restoration algorithms that are used for the evaluation. Sec. 3.4, finally, presents and discusses the results of the four experiments.

3.1 Test Signals

For the development and evaluation of the classification algorithm, we used a database of clean music recordings [42] 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., [28]). 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 Sec. 1. 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 $6200 \cdot 20\text{s} \approxeq 34\text{h}$. 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 that 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.

3.2 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.

3.2.1 Classification Performance

The performance of a classification system is typically rated based on three measures: *accuracy*, *precision,* and *recall* [12, 34]. 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:

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

3.2.2 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 al-

Table 2. The ODG scale.

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

gorithm, 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 [16, 19, 41]. 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]

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 [38] 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 [4]).

3.3 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 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 [31].
- DT-LSAR – An impulse restoration algorithm that uses an improved detection stage by using a double-threshold based approach [10]. 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 [10] and is reported to achieve good restoration performance for a wide range of input material without manual parameter adjustment [24].

3.4 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.

3.4.1 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., Sec. 3.1) 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 Sec. 3.2.1, 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

[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.

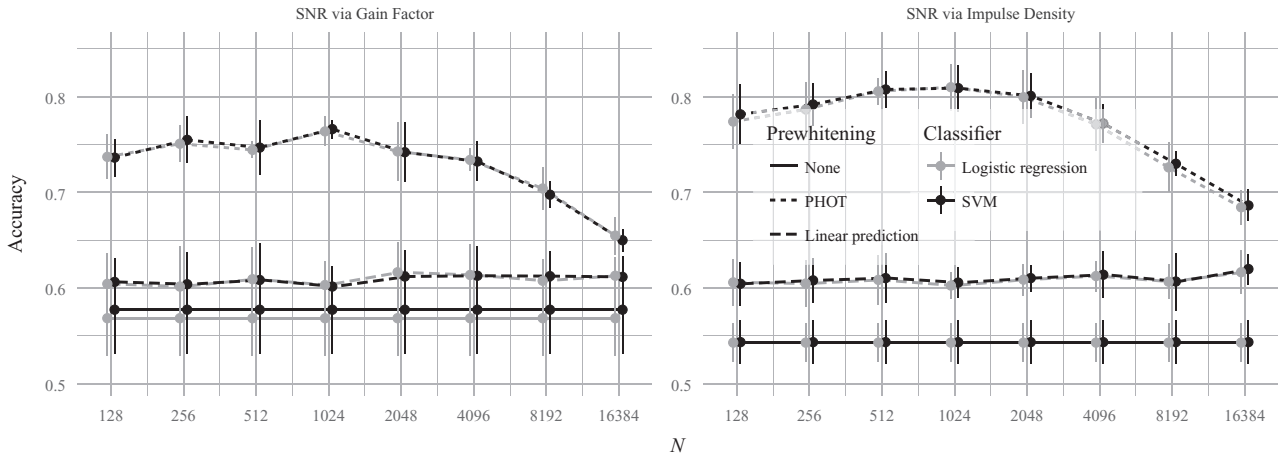

Fig. 4. Classification accuracy for all analyzed types of prewhitening with varying block lengths. The two figures show classification results averaged over all SNRs (20, 30, 40, and 50 dB). For the linear predictor, for each block length the optimum prediction order was selected. To enhance the clarity, 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.

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.

Fig. 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 Fig. 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. Fig. 4 shows that the achieved overall accuracy is higher if the SNR is set by modifying the impulse density (cf., Sec. 1.2.2). 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.

3.4.2 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 Sec. 2.3. The classification performance was then evaluated based only on the test data. Table 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 3). At an SNR of 40 dB the accuracy decreases to $\approx$67%, however note the precision and recall values: The recall value drops to $\approx$45% for both SNR concepts whereas the precision value indicates that $\approx 76\% - 78\%$ 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.

3.4.3 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., Sec. 3.1) with the LSAR, DT-LSAR, and Auto-LSAR algorithms (cf., Sec. 3.3) and rated the restoration capabilities in terms of the perceptual quality of the restored signal. As described in Sec. 3.2.2 the perceived quality is determined using an instrumental measure, namely the PEAQ algorithm. This

Table 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

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 [27] in Fig. 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 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 $1.5 \cdot \mathrm{IQR}$ 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). 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 [10]). 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.

3.4.4 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 Fig. 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. Fig. 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 Fig. 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 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

[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.

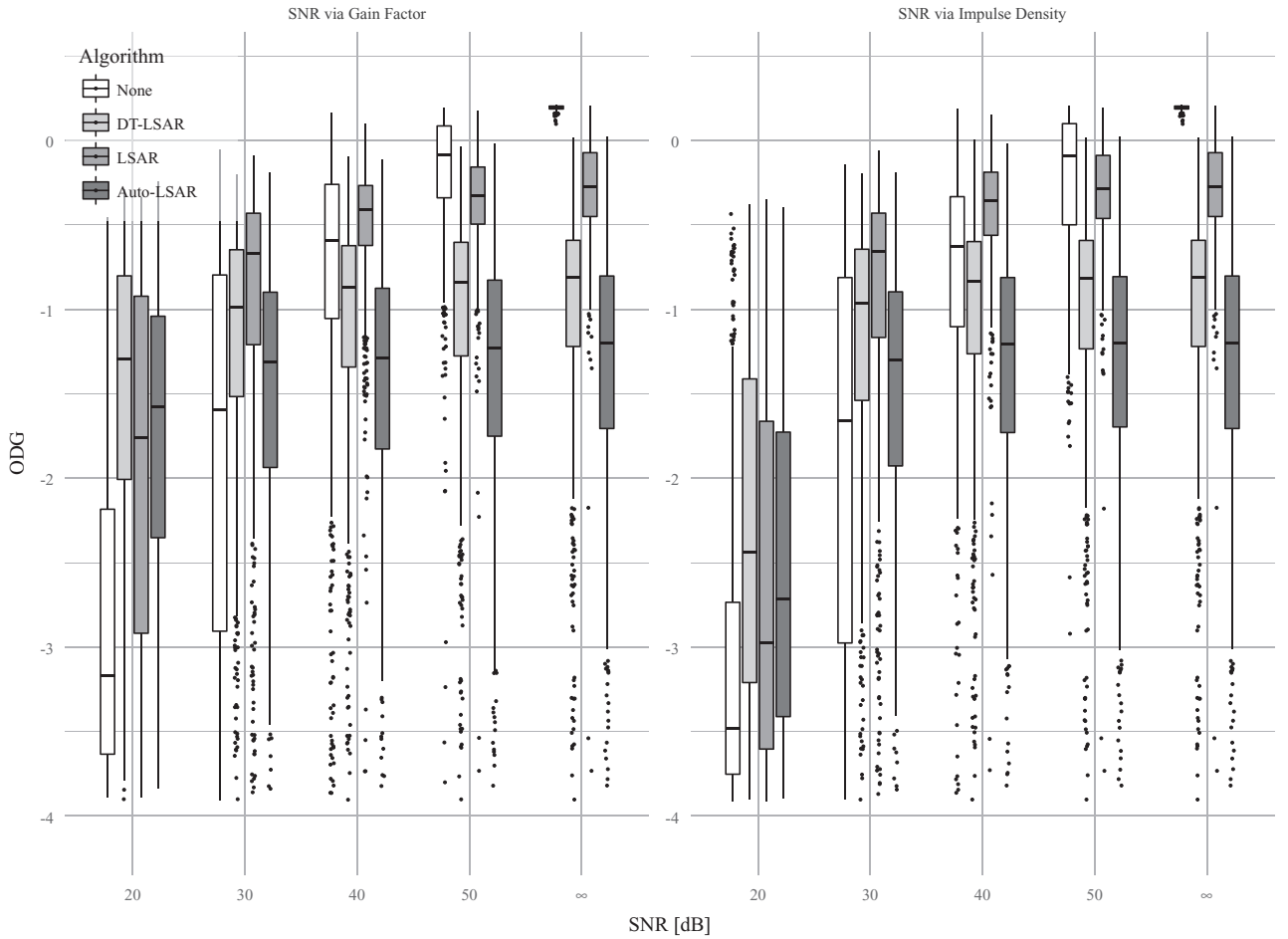

Fig. 5. Results of the instrumental audio quality evaluation of the three impulse restoration algorithms described in Sec. 3.3. The ODG scores were obtained with the PEAQ algorithm, and for each SNR concept ("Gain" or "Impulse Density") and SNR all 620 signals from the test signal database were used. 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 Sec. 3.4.3 for the interpretation of this box plot.

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.

4 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 [4] makes a number of disturbed and restored signals available for listening, including their ODG scores.

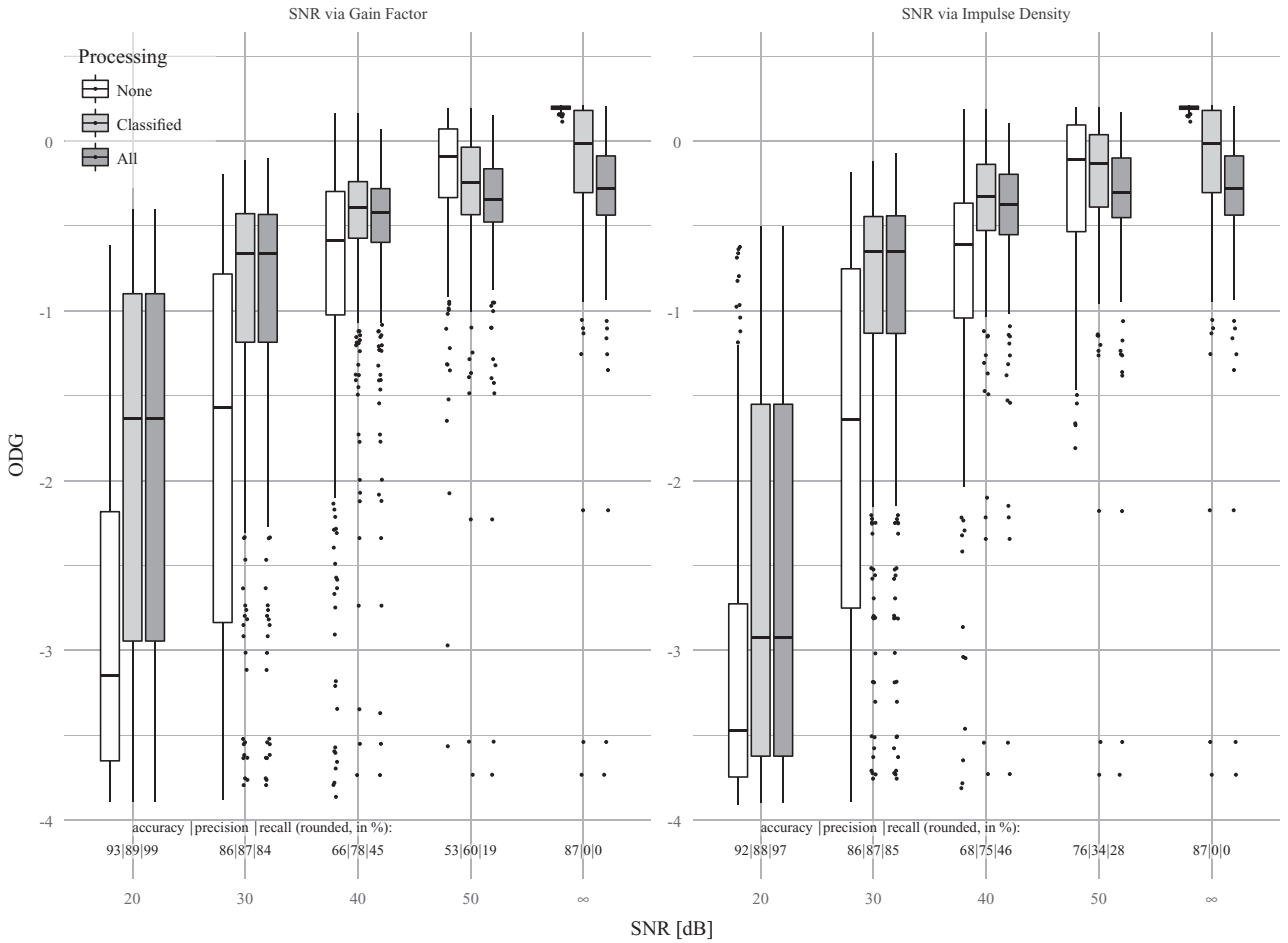

Fig. 6. Results of the instrumental audio quality evaluation of the full restoration processing chain. The ODG scores were obtained with the PEAQ algorithm, and for each SNR concept ("Gain" or "Impulse Density"), SNR, and type of processing only the 248 signals from the test set, previously unknown to 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 that 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 3 for convenience.

5 ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their very valuable and helpful comments that greatly contributed to improving this version of the paper. Also, we thank Dr. Esquef for providing an implementation of the impulse restoration algorithm in [10].

6 REFERENCES

[1] Annual Report of the Librarian of Congress, Washington: Library of Congress (2015 Sep.).

[2] D. Aiger, H. Talbot, "The Phase Only Transform for Unsupervised Surface Defect Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, San Francisco, pp. 295–302 (2010 June). https://doi.org/10.1109/CVPR.2010.5540198

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

[4] M. Brandt, *Impulsive Disturbances in Audio Archives: Signal Classification for Automatic Restoration – Demonstration Signals Accompanying the Article*. http://www.matbra.org (visited on 06/17/2017)

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

[6] A. Czyżewski, "Artificial Intelligence-Based Processing of Old Audio Recordings," *Proc. 97th Audio Eng. Soc. Conv.*, San Francisco (1994 Nov.).

[7] A. Czyżewski, "Learning Algorithms for Audio Signal Enhancement, Part 1: Neural Network Implementation for the Removal of Impulse Distortions," *J. Audio Eng. Soc.*, vol. 45, pp. 815–831 (1997 Oct.).

[8] A. Czyżewski, C. Suproń, "Learning Algorithms for the Cancellation of Old Recordings Noise," *Proc. 96th Audio Eng. Soc. Conv.*, Amsterdam (1994 Feb.).

[9] 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

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

[11] P. A. A. Esquef, V. Välimäki, K. Roth, I. Kauppinen, "Interpolation of Long Gaps in Audio Signals Using the Warped Burg's Method," *Proc. 6th Int. Conf. Digital Audio Effects (DAFx)*, London, pp. 8–11 (2003 Sep.).

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

[13] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, vol. 46, nos. 1–3, pp. 389–422 (2002 Jan.). https://dx.doi.org/10.1023/A:1012487302797

[14] C. M. Hicks, S. J. Godsill, "A Two-Channel Approach to the Removal of Impulsive Noise from Archived Recordings," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Adelaide, vol. 2, pp. II-213–II-216 (1994 Apr.). https://doi.org/10.1109/ICASSP.1994.389681

[15] P. O. Hoyer, "Non-Negative Matrix Factorization with Sparseness Constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469 (2004 Nov.).

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

[17] 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: Stanford Univ. Press (1960 Jan.).

[18] A Janssen, R. Veldhuis, L. Vries, "Adaptive Interpolation of Discrete-Time Signals That Can Be Modeled as Autoregressive Processes," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 2, pp. 317–330 (1986 Apr.). https://doi.org/10.1109/TASSP.1986.1164824

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

[20] I. Kauppinen, J. Kauppinen, P. Saarinen, "A Method for Long Extrapolation of Audio Signals," *J. Audio Eng. Soc.*, vol. 49, pp. 1167–1180 (2001 Dec.).

[21] J. K. Kauppinen, P. E. Saarinen, "True Linear Prediction by Use of a Theoretical Impulse Response," *J. Optical Soc. Amer. B*, vol. 11, no. 9, pp. 1631–1638. https://doi.org/10.1364/JOSAB.11.001631

[22] G. R. Kinzie, Jr., D. W. Gravereaux, "Automatic Detection of Impulse Noise," *J. Audio Eng. Soc.*, vol. 21, pp. 181–184 (1973 Apr.).

[23] C. Knapp, G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327 (1976 Aug.). https://doi.org/10.1109/TASSP.1976.1162830

[24] 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

[25] Y. Lavner, R. Cohen, D. Ruinskiy, H. Ijzerman, "Baby Cry Detection in Domestic Environment Using Deep Learning," *Proc. IEEE Int. Conf. the Science of Electrical Engineering (ICSEE)*, pp. 1–5 (2016 Nov.). https://dx.doi.org/10.1109/ICSEE.2016.7806117

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

[27] 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/10.2307/2683468

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

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

[30] M. Niedźwiecki, M. Ciołek, "Localization of Impulsive Disturbances in Archive Audio Signals Using Predictive Matched Filtering," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, 2888–2892 (2014 May). https://doi.org/10.1109/ICASSP.2014.6854128

[31] J. Nuzman, Audio Restoration: An Investigation of Digital Methods for Click Removal and Hiss Reduction (2004 Jan.). http://jnuzman.github.io/audio-restoration-2004 (visited on 01/09/2015)

[32] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., New York: McGraw-Hill, p. 79 (1991).

[33] 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," *J. Machine Learning Res.*, vol. 12, pp. 2825–2830 (2011 Oct.).

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

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

[36] P. J. W. Rayner, S. J. Godsill, "The Detection and Correction of Artifacts in Degraded Gramophone Recordings," *Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, pp. 151–152 (1991 Oct.). https://doi.org/10.1109/ASPAA.1991.634139

[37] J. M. Sacks, B. Isenberg, S. Klynas, "Reduction of Impulse Noise in Audio Signals,", *Proc. 57th Audio Eng. Soc. Conv*., Los Angeles (1977 May).

[38] K. Siedenburg, M. Dörfler, "Audio Denoising by Generalized Time-Frequency Thresholding," presented at the *Proc. 45th Int. Audio Eng. Soc. Conference*, Helsinki (2012 Mar.).

[39] S. J. Godsill, P. J. W. Rayner, *Digital Audio Restoration – A Statistical Model-Based Approach*, London: Springer, pp. 191-204 (1998). https://doi.org/10.1007/978-1-4471-1561-8

[40] A. K. Southey, M. Fox, T. Yeomans, "A Comparison of the Characteristics of ISO Fine Test Dust versus Real House Dust," *Proc. 12th Int. Conf. Indoor Air Quality and Climate*, Austin, vol. 868 (2011 June).

[41] 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," *J. Audio Eng. Soc.*, vol. 48, pp. 3–29 (2000 Feb.).

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

[43] S. V. Vaseghi *Algorithms for Restoration of Archived Gramophone Recordings*. PhD Thesis, Cambridge University (1988).

[44] S. V Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 4th ed., Chichester, West Sussex: Wiley, pp. 341-358 (2008). https://doi.org/10.1002/9780470740156

[45] S. V. Vaseghi, P. J. W. Rayner, "A New Application of Adaptive Filters for Restoration of Archived Gramophone Recordings," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP), New York* , vol. 5, pp. 2548–2551 (1988 Apr.). https://doi.org/10.1109/ICASSP.1988.197163

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

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

[48] 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

[49] V. Välimäki, S. González, O. Kimmelma, J. Parviainen, "Digital Audio Antiquing—Signal Processing Methods for Imitating the Sound Quality of Historical Recordings," *J. Audio Eng. Soc.*, vol. 56, pp. 115–139 (2008 Mar.).

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

[51] 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

APPENDIX

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 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.

A.2 List of Features

Table 5 lists all statistical measures that have been investigated as features of the prewhitened and normalized signal $x_{\text{pre}}^{\prime(q)}[n]$. As described in Sec. 2.2, recursive feature

Table 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	2434	0	(default)
40 dB	24083	6374	
50 dB	227976	99481	

Table 5. Features of the prewhitened signal that have been investigated.

Feature	Computation
Crest factor	See (6)
Crest factor – l% trimmed mean, ($l \in \{1, 2, 5, 10\}$ (see [50, Ch. 3.3])	$C_{l\%}^{(q)} = \frac{\max\limits_{0\le i<M}\left\lvert x_{\text{pre}}^{\prime(q)}[i]\right\rvert}{\sqrt{\text{Trimmean}_{l\%}\left\{\left(x_{\text{pre}}^{\prime(q)}[n]\right)^2\right\}}}$
Peak-to-Root-Median-Squared ratio	$\text{PRMedS}^{(q)} = \frac{\max\limits_{0\le i<M}\left\lvert x_{\text{pre}}^{\prime(q)}[i]\right\rvert}{\sqrt{\text{Med}\left\{\left(x_{\text{pre}}^{\prime(q)}[n]\right)^2\right\}}}$
Kurtosis	See (7)
Kurtosis of absolute value	$\text{Kurt}_{\text{abs}}^{(q)} = \frac{\frac{1}{M}\sum_{i=0}^{M-1}\left(\left\lvert x_{\text{pre}}^{\prime(q)}[i]\right\rvert - \overline{\left\lvert x_{\text{pre}}^{\prime(q)}\right\rvert}\right)^4}{\left(\frac{1}{M}\sum_{i=0}^{M-1}\left(\left\lvert x_{\text{pre}}^{\prime(q)}[i]\right\rvert - \overline{\left\lvert x_{\text{pre}}^{\prime(q)}\right\rvert}\right)^2\right)^2}$
Skewness	$\text{Skew}^{(q)} = \frac{\frac{1}{M}\sum_{i=0}^{M-1}\left(x_{\text{pre}}^{\prime(q)}[i] - \overline{x_{\text{pre}}^{\prime(q)}}\right)^3}{\left(\frac{1}{M}\sum_{i=0}^{M-1}\left(x_{\text{pre}}^{\prime(q)}[i] - \overline{x_{\text{pre}}^{\prime(q)}}\right)^2\right)^{3/2}}$
Sparseness (see [15])	$\text{Sparseness}^{(q)} = \frac{\sqrt{M} - \sum_{i=0}^{M-1}\left\lvert x_{\text{pre}}^{\prime(q)}[i]\right\rvert / \sqrt{\sum_{i=0}^{M-1}\left(x_{\text{pre}}^{\prime(q)}[i]\right)^2}}{\sqrt{M}-1}$

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 [50, Ch. 3.3].

THE AUTHORS


Matthias Brandt


Simon Doclo


Timo Gerkmann


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. Since 2013, he is a Ph.D. student at the University of Oldenburg, Germany, in the field of audio restoration. His research focus is on the processing of music signals – from developing methods to extract parameters required for automatic denoising to creating electronic music in his spare time.

•

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, 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 (with Erik De Clippel), 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 (with Marc Moonen) and the IEEE Signal Processing Society 2008 Best Paper Award (with Jingdong Chen, Jacob Benesty, Arden Huang). 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. He was secretary of the IEEE Benelux Signal Processing Chapter (1998–2002) and Technical Program Chair for the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) in 2013. Prof. Doclo served as guest editor for several special issues (*IEEE Signal Processing Magazine, Elsevier Signal Processing*) and is associate editor for *IEEE/ACM Transactions on Audio, Speech and Language Processing* and *EURASIP Journal on Advances in Signal Processing*.

•

Timo Gerkmann studied electrical engineering at the universities of Bremen and Bochum, Germany. He received his Dipl.-Ing. degree in 2004 and his Dr.-Ing. degree in 2010 both at the Institute of Communication Acoustics (IKA) at the Ruhr-Universität Bochum, Bochum, Germany. In 2005, he spent six months with Siemens Corporate Research in Princeton, NJ, USA. During 2010 to 2011 Dr. Gerkmann was a postdoctoral researcher at the Sound and Image Processing Lab at the Royal Institute of Technology (KTH), Stockholm, Sweden. From 2011 to 2015 he was a professor for Speech Signal Processing at the Universität Oldenburg, Oldenburg, Germany. During 2015 to 2016 he was a Principal Scientist for Audio & Acoustics at Technicolor Research & Innovation in Hanover, Germany. Since 2016 he is a professor for Signal Processing at the University of Hamburg, Germany. His research interests are on digital signal processing algorithms for speech and audio, communication devices, hearing instruments, audio-visual media, and human-machine interfaces. Timo Gerkmann is a Senior Member of the IEEE.

•

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.