IPOL
IMAGE PROCESSING ON LINE

Published in Image Processing On Line on 2015–11–21.
Submitted on 2013–02–01, accepted on 2013–07–19.
ISSN 2105–1232 © 2015 IPOL & the authors CC–BY–NC–SA
This article is available online with supplementary materials,
software, datasets and online demo at
http://dx.doi.org/10.5201/ipol.2015.64

Automatic Detection and Removal of Impulsive Noise in Audio Signals

Laurent Oudre

CMLA, ENS Cachan, France (laurent.oudre@cmla.ens-cachan.fr)

Abstract

This article presents a method for restoring audio signals corrupted by impulsive noise such as clicks, bursts or scratches. The algorithm takes as input a degraded audio signal and automatically detects the locations of the degraded samples and replaces them with more appropriate values. Both steps (detection and interpolation) are based on the assumption that the signal can locally be modeled as a realization of an autoregressive process. Surprisingly, the results obtained on several types of signals (classical, jazz, vocal, etc.) show that a fully automatic method, with a carefully fixed set of parameters, can achieve good performance on a wide range of degraded audio signals.

Source Code

The reviewed source code in C language for this algorithm and an online demo are available from the web page of this article[1]. Compilation and usage instruction are included in the `README.txt` file of the archive.

Keywords: sound processing; denoising; auto-regressive models

1 Introduction

The degradations commonly found in audio signals can be broadly classified into two groups: localized and global [9]. The global degradations affect all the samples of the waveform while the localized ones only affect certain groups of samples and thus cause a discontinuity in the waveform. Among global degradations we can cite background noise, hiss, flutter and certain types of non-linear degradations such as speed variations or distortion. In this article, we focus on localized degradations such as clicks, bursts, outliers, crackles, scratches, etc. These degradations are particularly common in old gramophone recordings [17]. Restoring audio signals corrupted by impulsive noise is a tricky process which can be divided into two steps: *detection* (finding the locations of the degraded samples) and *interpolation* (replacing the degraded samples by more suitable values).

[1] http://dx.doi.org/10.5201/ipol.2015.64

The *detection* step is crucial as the performances of any reconstruction or interpolation method greatly depend on the knowledge of the locations of corrupted samples (unbiased estimation of the parameters, fidelity to the original signal, convergence of iterative algorithms, etc.) [8]. When the degradations only consist of small bursts (with a duration of several samples) with high amplitudes, it is rather easy to locate the corrupted samples (as their amplitudes are particularly high compared to the other samples). In these cases, a simple median filter or the calculation of some derivatives may be sufficient for the detection [12]. Unfortunately, in most cases, the bursts, scratches, clicks, etc., last from less than 20 microseconds to 4 milliseconds, which corresponds to 1 to 200 samples (with a sampling frequency of 44.1 kHz). Furthermore, the number of bursts can rise up to 2000 per second in poor gramophone recordings [8]. Finally, in practice, the amplitudes of the degradation vary greatly not only according to the type of physical defect, but also within the same recorded extract, making it very hard to detect the degradations at first sight. More sophisticated methods exist to deal with more difficult cases, where the locations of the corrupted samples are determined either in the frequency (or transform) domain [3, 14] or in the time domain. In the latter case, a model is often assumed for the signal; the degradations are supposed to correspond to the samples which do not properly fit the model. In most cases, the underlying model is inspired by autoregressive (AR) or autoregressive moving average (ARMA) processes. Indeed, typical audio and speech signals can be efficiently approximated as AR or ARMA processes as those are accurate to model the physical phenomenon of sound production (also known as *source-filter* model). Depending on the methods, the model is either time-invariant (the parameters are constant on a given frame) [17, 18, 6, 7, 4, 5, 1] or time-varying [15, 2] (the parameters slowly vary with time).

The *interpolation* step uses as input the locations of the degraded samples estimated in the detection step. Just like for detection, numerous methods have been developed for interpolation of missing samples in music or speech signals. An overview of these techniques, along with the description of one AR-based method for interpolating missing samples can be found in [16].

This article describes a complete and automatic method for denoising audio signals corrupted by impulsive noise. This method assumes that the signal can be approximated by a *locally* stationary AR process and uses this hypothesis in order to detect the corrupted samples, which are then reconstructed with the method described in [16].

2 Detection Step

2.1 Principle

Consider a finite audio signal $\mathbf{s} \in \mathbb{R}^N$, which is corrupted by a random, additive and impulsive noise $\mathbf{n}$. The observed signal $\mathbf{x} = \{x_t\}_{t=1\dots N}$ can be written as

$$x_t = s_t + n_t. \tag{1}$$

Note that this general additive model is only valid if the degradation is not too severe and if the original signal is still detectable in the observed signal. Should this not be the case, a replacement model would be more appropriate [9]. Also, a more precise approximation of the observed signal might be achieved by adding a third term in Equation (1) corresponding to the broadband noise [14].

Now, since the additive noise is supposed to be impulsive, all the samples are not concerned by the degradation. The noise can be split into two distinct and independent components: a switching process $\mathbf{i}$ with values 0 or 1 and a corrupting noise $\mathbf{v}$ [9] such as

$$n_t = i_t \cdot v_t. \tag{2}$$

The aim of the detector is to estimate as correctly as possible the switching process $\mathbf{i}$. From the noise structure described in Equation (2), we can formulate some remarks:

- We have previously seen that, in practice, the degradations may last up to several hundreds of samples. This means that the values of i_t are not completely independent as the degradations tend to occur in contiguous bursts of corrupted samples.

- The randomness of the values of v_t can greatly influence the performance of the detector. Indeed, if v_t is small, the noise might be undetectable, causing a missed detection and possibly a bad estimation of the burst length. On the contrary, if v_t is large, the estimation of signal parameters is corrupted, which can possibly lead to irrelevant results.

Let us suppose that the original clean signal $\mathbf{s}$ can be approximated as a realization of a locally stationary AR process. Then, there exist an order $p \in \mathbb{N}^*$ and some coefficients $\mathbf{a} = [a_1, \ldots, a_p]^t \in \mathbb{R}^p$, $a_p \neq 0$ such that

$$s_t = -\sum_{k=1}^{p} a_k \, s_{t-k} + e_t, \tag{3}$$

where $\mathbf{e}$ is a zero-mean white noise of variance σ_e^2.

From Equation (1) and Equation (3), we can rewrite x_t only in terms of past values x_{t-k} (since we have no access to s_t)

$$x_t = -\sum_{k=1}^{p} a_k \, s_{t-k} + e_t + n_t, \tag{4}$$

$$x_t = -\sum_{k=1}^{p} a_k \, (x_{t-k} - n_{t-k}) + e_t + n_t, \tag{5}$$

$$d_t = x_t + \sum_{k=1}^{p} a_k \, x_{t-k} = e_t + n_t + \sum_{k=1}^{p} a_k \, n_{t-k}. \tag{6}$$

In Equation (6), the left hand term $d_t = x_t + \sum_{k=1}^{p} a_k \, x_{t-k}$ corresponds to a filtered version of the corrupted data x_t by using the prediction error filter $H(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}$ and shall be called *detection signal* in the following. The right term is composed of the excitation signal e_t, the impulsive noise n_t and the effect of the past p impulsive noise samples.

Note that the transformation provided by the prediction error filter from Equation (1) to Equation (6) has the nice property of transforming the original signal s_t to the excitation signal e_t while keeping the impulsive noise n_t unchanged or increased (providing that the impulsive responses of the previous noise samples do not cancel each other). In practice, for audio or speech signals, the typical amplitude of s_t is up to 10^4 times the one of e_t, making it easier to detect the impulsive noise. From the point of view of the detection problem, the signal-to-noise ratio in Equation (1) is $\frac{E[n_t^2]}{E[s_t^2]}$ while in Equation (6) it becomes $\frac{E[n_t^2]}{E[e_t^2]}$ (contrary to what one may think, the useful signal is here n_t as it is the term necessary for the detection of the noise samples). With typical values for s_t and e_t, the improvement in SNR is up to 40 dB [18].

An easy way to detect the locations of the noisy samples is to threshold the detection signal $|d_t|$. As the values of $|e_t|$ are random and supposedly small, high values of $|d_t|$ correspond to high values of $|n_t|$ and therefore to noisy samples. In particular, if the p previous samples are not corrupted by impulsive noise, the detection signal is exactly the sum of e_t and n_t. The process becomes trickier when some of the p previous samples are indeed degraded and artifacts appear. Also note that the precision in the estimation of the locations of the bursts is limited due to the propagation of the impulsive noise over the next p samples in the detection signal.

2.2 Estimation of the AR Parameters

Another important issue is the estimation of the AR parameters. Indeed, the whole filtering process is based on the *good* estimation of the AR parameters (which only makes sense if the signal can be well modeled with an AR model). If we take into account that the estimated parameters $\hat{a_k}$ are different from the *true* parameters a_k, then the detection signal rewrites

$$\begin{aligned} \hat{d_t} &= x_t + \sum_{k=1}^{p} \hat{a_k}\, x_{t-k} && (7) \\ &= e_t + n_t - \sum_{k=1}^{p} a_k\, s_{t-k} + \sum_{k=1}^{p} \hat{a_k}\, x_{t-k} && (8) \\ &= e_t + n_t + \sum_{k=1}^{p} a_k\, n_{t-k} + \sum_{k=1}^{p} (\hat{a_k} - a_k)\, x_{t-k}. && (9) \end{aligned}$$

We see that an additional term appears in the expression of $\hat{d_t}$, corresponding to the estimation error. Once again, this error depends on the p previous samples but, assuming that the AR parameters are correctly estimated, it should remain lower than the present impulse n_t, allowing to use the criterion d_t for the impulsive noise detection.

Many methods exist for the determination of appropriate AR parameters from a series of audio samples. One of the most efficient and less time-consuming is based on the Yule-Walker equations. These equations, characteristic of AR processes, provide a relationship between the AR parameters and the values of the autocorrelation function. Given an estimation of the autocorrelation function

$$\hat{R}(\tau) = \frac{1}{N} \sum_{k=\tau+1}^{N} x_k\, x_{k-\tau}, \tag{10}$$

the AR parameters are estimated through

$$\begin{bmatrix} \hat{R}(0) & \hat{R}(1) & \cdots & \hat{R}(p-1) \\ \hat{R}(1) & \hat{R}(0) & \cdots & \hat{R}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{R}(p-1) & \hat{R}(p-2) & \cdots & \hat{R}(0) \end{bmatrix} \begin{bmatrix} \hat{a_1} \\ \hat{a_2} \\ \vdots \\ \hat{a_p} \end{bmatrix} = - \begin{bmatrix} \hat{R}(1) \\ \hat{R}(2) \\ \vdots \\ \hat{R}(p) \end{bmatrix}. \tag{11}$$

There exists a fast algorithm (Levinson-Durbin algorithm [13]) dedicated to Toeplitz matrices which allows to solve Equation (11) in $O(p^2)$ operations. This algorithm also provides an estimation of the variance $\hat{\sigma_e}^2$ of the excitation process. More details on the Yule-Walker equations and on the Levinson-Durbin algorithm can be found in [16].

2.3 Direct Thresholding

We have previously seen that, given an estimation $\hat{\mathbf{a}}$ of the AR parameters, the term $d_t = x_t + \sum_{k=1}^{p} \hat{a_k}\, x_{t-k}$ could be used as a decision criterion for detecting corrupted samples. In fact, from Equation (9), we see that when $i_t = 1$ (i.e. impulsive noise is present), and provided that the following conditions are satisfied:

1. the estimation error of the AR parameters $(\hat{a_k} - a_k)$ is small;
2. the excitation signal $|e_t|$ is small compared to $|v_t|$;

3. either the p previous samples are uncorrupted or $\sum_{k=1}^{p} a_k\, n_{t-k}$ is small;

then we can say that $|d_t| \approx |v_t|$. Note that hypotheses 1 & 2 are very likely to be satisfied if the AR approximation fits the original signal.

Before investigating further on these hypotheses, we propose to figure out the general shape of the criterion d_t on a simple example. Let us consider 2000 samples of an audio file (The Beatles) sampled at 44.1kHz. We artificially created a burst of $N_{max} = 50$ samples at the center of the extract. The corrupting noise v_t was simulated by a zero-mean Gaussian white noise with variance 10^{-4}. The AR parameters were estimated with the Levinson-Durbin algorithm as previously described. The value of p was set to $p = 3N_{max} + 2 = 152$ (this empirical expression proved to give acceptable results for bursts whose length is up to $N_{max} = 50$ samples – see [16] for details).

Figure 1 presents the corrupted audio signal x_t along with the corresponding detection signal $|d_t|$ (note that due to its definition, the detection signal is only defined for $t > p$). We can observe that, as far as this audio extract is concerned, hypotheses 1 & 2 seem to be valid, as the values of $|d_t|$ are small when there is no impulsive noise. Similarly, for most of the samples corrupted by impulsive noise, the criterion $|d_t|$ takes high values. There are therefore two main issues which prevent from directly thresholding criterion $|d_t|$ for detection:

- Inside the burst, due to destructive interferences and to the fact that the impulsive noise is random, there are some samples for which the term $n_t + \sum_{k=1}^{p} a_k\, n_{t-k}$ is small, which can possibly lead to missed detections.
- The effects of impulsive noise tend to propagate on the next p samples (this is due to the term $\sum_{k=1}^{p} a_k\, n_{t-k}$), which can possibly lead to false detections.

Let us prove these assumptions on our example thanks to some quality metrics. We define the precision P and the recall R as

$$P = \frac{\text{number of detected samples which are indeed corrupted}}{\text{number of detected samples}}, \tag{12}$$

$$R = \frac{\text{number of detected samples which are indeed corrupted}}{\text{number of corrupted samples}}. \tag{13}$$

If the detection process is flawless (i.e. if it is possible to detect the corrupted samples by thresholding the criterion $|d_t|$), then there should exist a threshold such as $P = 1$ and $R = 1$. If the threshold is too low then too many samples are detected and P decreases; if the threshold is too high, no sample is detected and R decreases.

We assume here that the *detection threshold* λ can be written as

$$\lambda_K = K\hat{\sigma}_e, \tag{14}$$

where $\hat{\sigma}_e$ is the estimated value of the excitation standard-deviation. This formulation takes into account that criterion d_t depends on both e_t and n_t and that if the excitation signal e_t can take high values, then the threshold should be adapted in consequence.

For our example, we have tried all detection thresholds λ_K with K between $K = 0$ and $K = 12$ with a step of 10^{-4} and each time, we have calculated the associated precision and recall. The plot we obtain is displayed in Figure 2. The upper-left corner of the figure corresponds to high thresholds while the lower-right corner corresponds to low thresholds. We clearly see in this figure that, as far as our example is concerned, it is impossible to find a threshold which allows to perfectly detect all corrupted samples. Moreover, the best compromise (i.e. the closest point to the optimal coordinates $(R, P) = (1; 1)$) gives a recall of 66% and a precision of 86.84% which is somehow unsatisfactory.

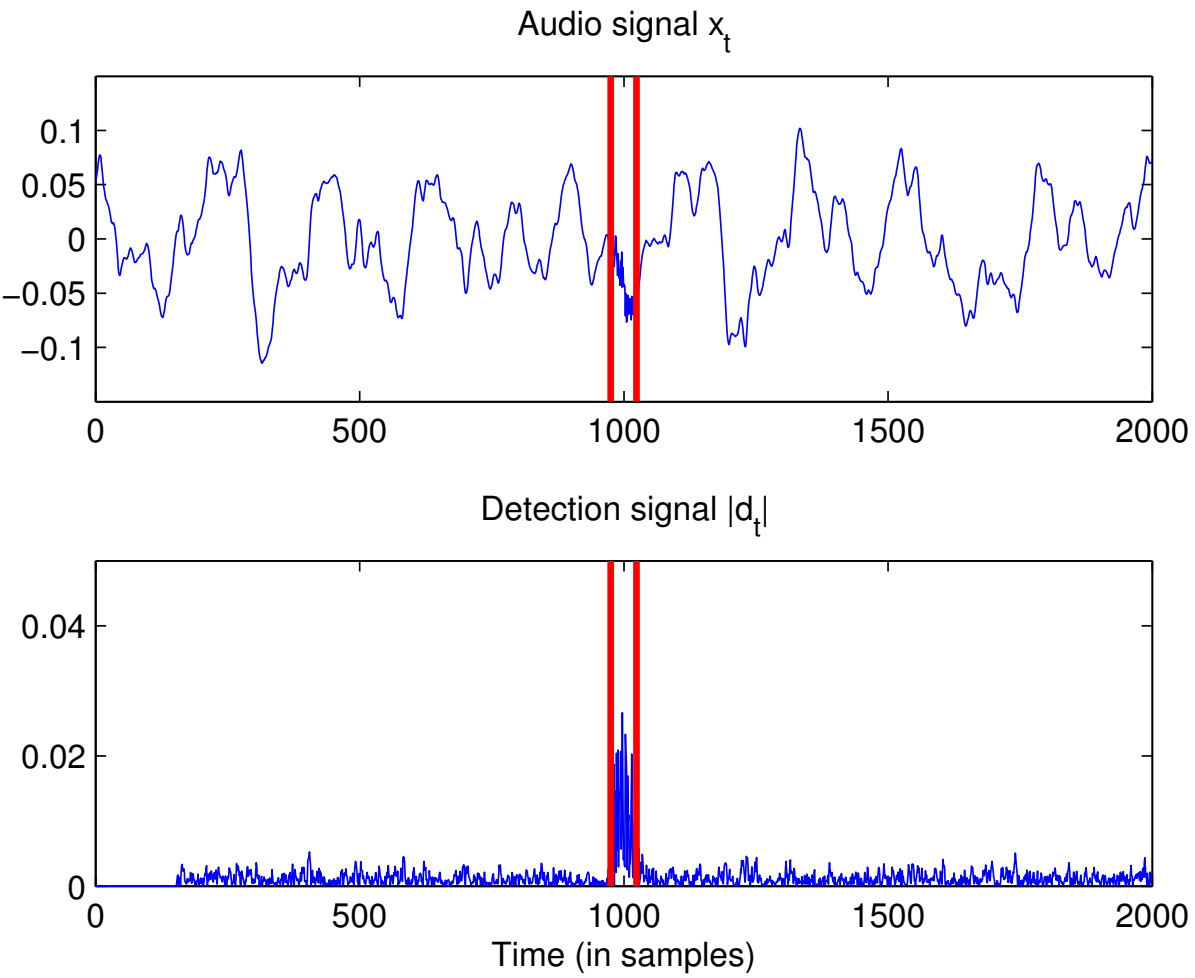

Figure 1: Example of detection signal for an audio signal of pop music artificially corrupted by impulsive Gaussian noise.

2.4 Improvements and Variants

All the variants described in this section build on this basic threshold-based detection but aim to add some pre- or post-processing steps within the calculation of the criterion so as it really detects bursts of contiguous samples. If impulsive noise corrupts bursts and not individual samples, the task of noise detection is in fact reduced to the identification of the number of bursts present on the frame, along with their first and last samples.

2.4.1 Single Burst

When only one burst is present, an easy solution could then consist in finding the first and last samples for which the criterion is greater than the threshold and decide that all the samples in between belong to the burst. When conducting this simple change in the experiment described in Section 2.3 and testing all detection threshold values λ_K for K comprised between $K = 0$ and $K = 12$, the best compromise now provides a recall of 94% and a precision of 100%: in fact the burst boundaries are correctly estimated except for the three first samples of the burst which remain undetected.

Unfortunately, despite the good results obtained in this example, some questions remain open: how can we deal with real-time scenarios where several bursts are present and where the degradations

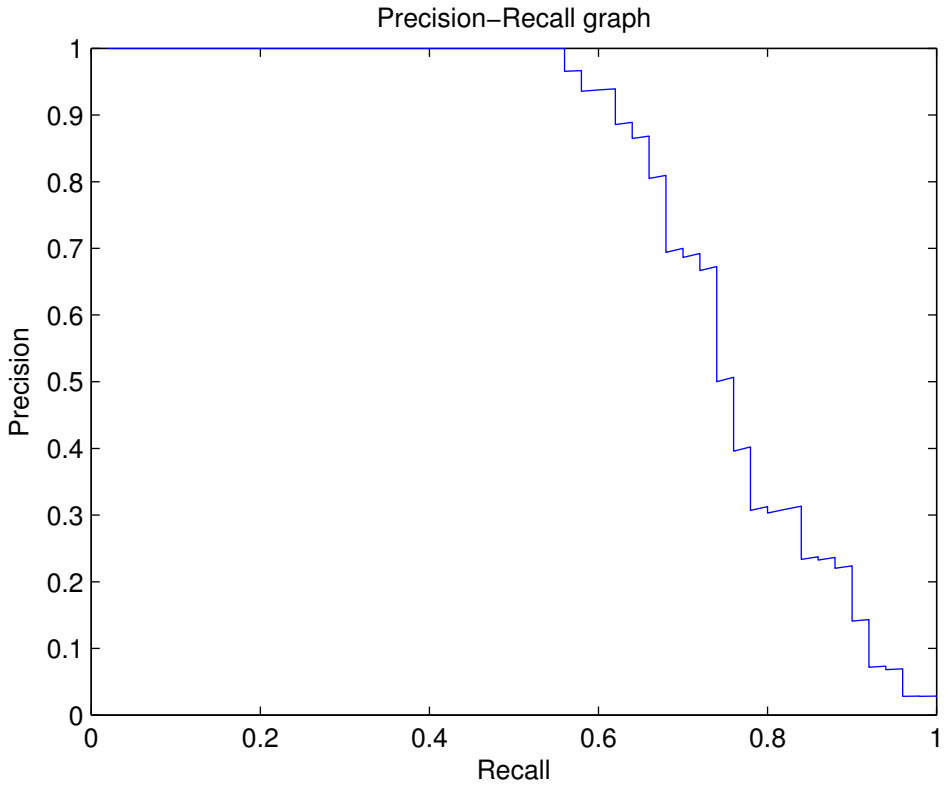

Figure 2: Precision-recall graph obtained by thresholding criterion $|d_t|$.

are not easily detected? How can we determine a threshold which could work for different types of degradations?

2.4.2 Multiple Bursts

Figure 3 shows the detection signal obtained on a real-life example of classical music corrupted by clicks. Since the maximum length of a burst N_{max} is not known, we have considered the same value for p than for the previous experiment ($p = 152$). The criterion d_t seems to be able to identify one large click (sample 1000), one moderate click (sample 900) and several small clicks which are hard to locate (samples 530, 560 and 1750).

The method explained in Section 2.4.1 can no longer be applied in this case, as multiple bursts are apparently present on the frame. Yet, this method can be adapted to multiple bursts by defining a *fusion parameter* b which corresponds to the maximum number of consecutive samples within a burst whose values are lower than the threshold λ_K [1].

In practice, the detection process is now composed of two steps:

1. Find the set of samples $T \subset \{1, \ldots, N\}$ such as $\forall t \in T, |d_t| > \lambda_K$.
2. If $(t, t') \in T^2$ and $|t - t'| \leq b$, then add all the samples comprised between t and t' into T.

In [1], the authors propose to use the value of $b = 20$ (which is equivalent to 0.45 ms for a sampling frequency of 44.1 kHz). The results obtained on our real-life example for $b = 1$ (basic thresholding) and $b = 20$ and different values for K (2, 2.5 and 3) are displayed on Figure 4. We clearly see that the number of detected bursts varies much with threshold λ_K but also with b. Intuitively, the lower λ_K, the more samples satisfy the condition $|d_t| > \lambda_K$ and the larger the number of detected bursts. The parameter b acts in the opposite direction and tends to merge distinct bursts if they are close to

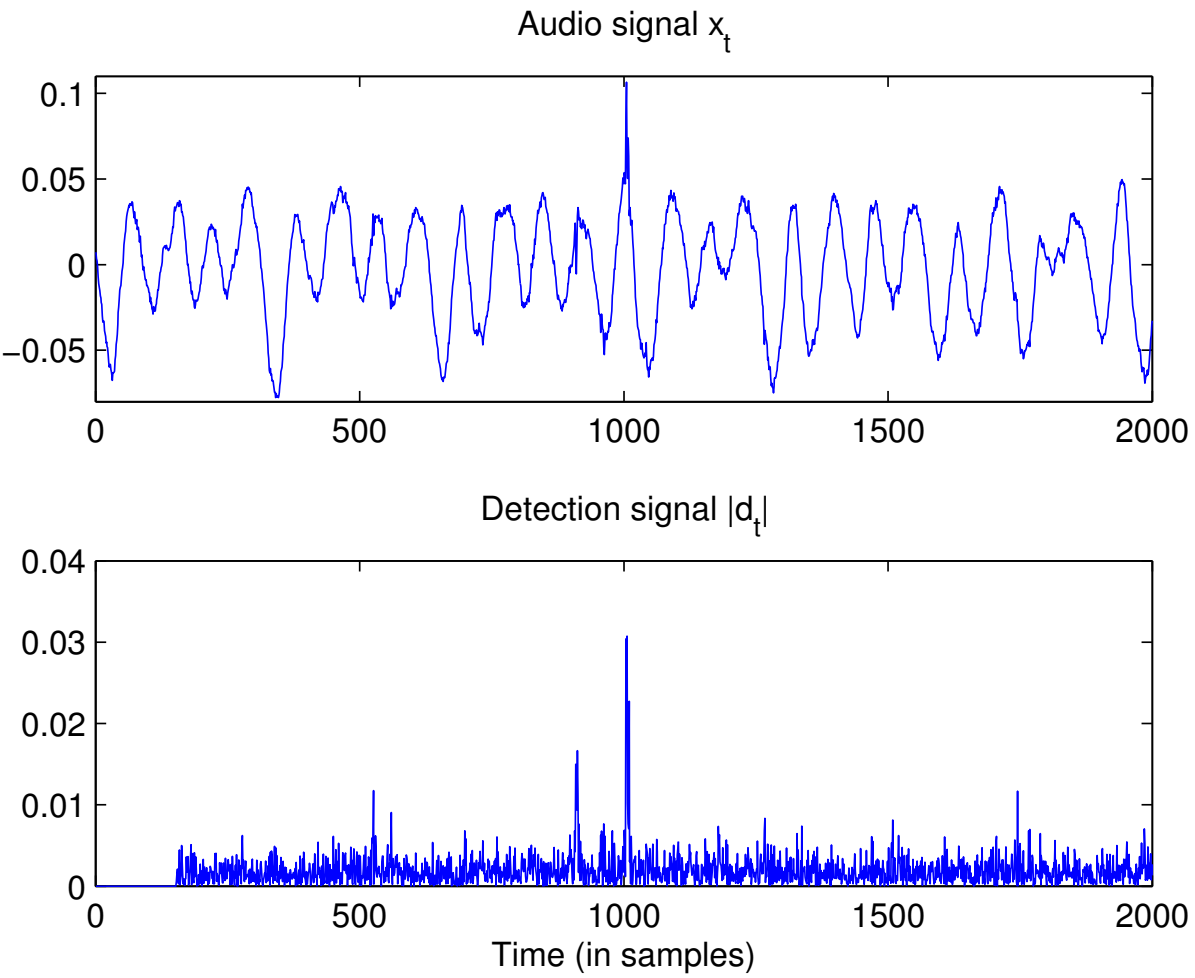

Figure 3: Example of detection signal for real-life audio signal of classical music corrupted by clicks.

each other. One may hope that by adjusting these two thresholds, it should be possible to estimate the accurate number of bursts.

3 Interpolation Step

This article concentrates on the detection of impulsive noise samples; for the interpolation we rely on the method described in [16]. It is an implementation of the algorithm introduced in [11]. This classic method offers the good property of being based on AR models as well, which allows us to factorize the calculation of the AR parameters and use them for both the detection and interpolation phase. In a nutshell, the method selects the values for the missing samples that minimize the sum of the absolute prediction error by the AR model. We refer the reader to [16] for more details. The parameters linked to this method are p and N_w, i.e., the order of the model and the window length; these parameters are set to the same values used for the detection phase.

4 Summary of the Denoising Method

Now, if we seek to apply our method on real-life examples, we must take into account that a 30 seconds excerpt with a sampling frequency of 44.1 kHz contains $N = 1323000$ samples, which is way

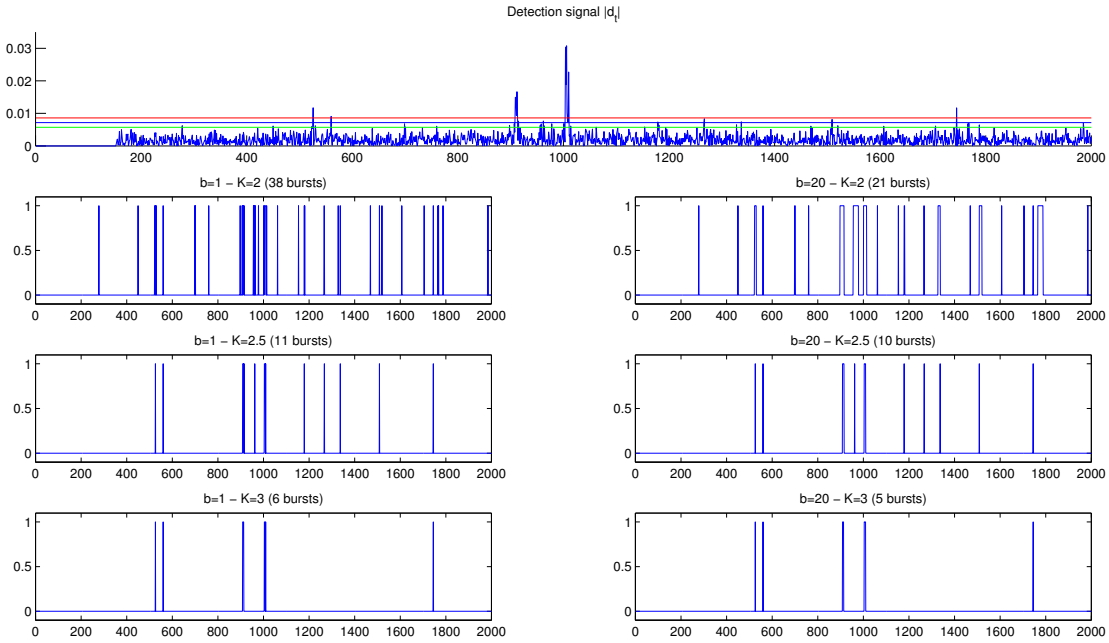

Figure 4: Detected bursts for different values of λ_K and b: the vectors take value 1 when impulsive noise is detected and 0 elsewhere.

too large for our method. Indeed, it is unrealistic to model several seconds of music with one single AR model since the stationarity assumed by the model is only valid on a local scale. We need to divide the audio signal into frames before processing it with the detection+interpolation algorithms.

As far as the frame processing is concerned, the signal is divided into overlapping frames of length N_w with a hop size of N_h samples. In practice, we choose $N_h = N_w/4$, corresponding to a 75% overlap. In order to reconstruct the information provided by the different overlapping frames which correspond to the same period, we use the overlap-add (OLA) procedure [10]. Note that the estimation of the AR parameters and the calculation of d_t requires that the first p samples of the frame are left outside of the detection+interpolation process and that the interpolation method assumes that no degradation is present in the last p samples (if it is not the case, the degradations will be processed in the next frame thanks to the overlapping effect).

The whole process is summarized in Algorithm 1.

The whole process is likely to provide better results when iterated several times. Indeed, in the first iteration the estimation of the AR parameters (which are used in both the detection and interpolation steps) is actually performed on the degraded signal and is therefore necessarily biased and may cause some distortions or mis-detections. The output of the first iteration, which hopefully is *cleaner* that the input degraded signal, should provide better estimates and thus better results. Of course, the number of iterations N_{iter} should remain limited as at some point either the signal has efficiently been de-noised and extra iterations are useless or the model does not fit the signal and extra iterations would artificially force the signal to fit the model causing distortions.

5 Results

The method described in the previous section will be now tested on several examples.

Our detection+interpolation method depends on five parameters:

Algorithm 1: Detection and removal of impulsive noise

Input: Degraded signal $\mathbf{x}$, Parameters K, b, p, N_w, N_{iter}
Output: Restored signal, set of samples detected as noise T

1 **Do** N_{iter} **times**
2 Pad the signal with zeros: N_w zeros are added before and after the signal samples
3 Add $\left(\lceil \frac{N+N_w}{N_h} \rceil N_h - N_w\right) - N$ zeros at the end of the signal so as to round up the number of frames
4 Divide the signal into overlapping frames of length N_w with 75% overlap
5 Estimate the AR parameters $\hat{\mathbf{a}}$ and $\hat{\sigma_e}$ on each frame through the Levinson-Durbin algorithm [16]
6 Calculate the criterion d_t on each frame for $t = p+1, \cdots, N-p$; use thresholds λ_K and b for detecting the locations of the corrupted samples T
7 Reconstruct the corrupted samples on each frame with the interpolation method described in [16] and the AR parameters calculated on Step 4
8 Multiply each frame with a Hamming window of size N_w
9 Add iteratively all the frames with a 75% overlap so as to reconstruct the signal
10 Remove the N_w first and the $\left(\lceil \frac{N+N_w}{N_h} \rceil N_h\right) - N$ last samples

- The detection threshold $\lambda_K = K\hat{\sigma_e}$
- The fusion parameter b
- The order of the AR model p
- The window length N_w
- The number of iterations N_{iter}

The first two parameters are linked to the detection step, while the last three are common for the detection and interpolation steps. In this article, we propose to use the default values for p and N_w, as described in [16], i.e. $p = 3N_{max} + 2$ and $N_w = 8p$, as this choice allows us to assume that the interpolation process is optimized. In practice, the maximum length for a burst is around $N_{max} = 100$ samples, so we set $p = 302$ and $N_w = 2416$.

5.1 Influence of λ_K and b

Intuitively:

- If λ_K is high, the system is designed to only detect strong bursts and clicks with high amplitudes. If λ_K is low, all clicks will be detected but it is likely that so will background noise.
- The parameter b indirectly controls the length of the detected bursts. Note that this parameter also greatly influences the interpolation process: indeed, the results are not the same when reconstructing two small bursts or one large burst (even if only 1 sample separates the two bursts). Also, b should be large enough to create contiguous bursts of reasonable length but unfortunately large bursts are also harder to reconstruct.

We have tested these hypotheses on a 30 seconds degraded extract of classical music by Mussorgsky[2]. This real-life extract is corrupted by clicks and crackles but also by some background

[2]This file was used by Godsill for the evaluation of his reconstruction methods http://www-sigproc.eng.cam.ac.uk/~sjg/springer/index.html

noise (hiss). The extract sounds like an old gramophone recording and exhibits a range of click-type defects from huge "pops" to small crackles.

We have tried several values for λ_K ($K = 1$, 1.5, 2, 2.5 and 3) and b (all integer values between 1 and 30). We present in Figure 5 the number of samples detected by our method as noise (in % of the total number of samples) and the average length of the detected bursts (in samples) in this case. We confirm in this example that the combination of λ_K and b allows us to obtain various configurations for our system that detect various numbers and shapes of bursts. Indeed, in this excerpt, the number of noise samples varies from 1% to 82% while the average length varies from 1 to 230 samples (0.02 ms to 5.2 ms). Even before hearing the reconstructed signals, we can see that some of the configurations are unrealistic as in most of the cases it is assumed that impulsive degradations only corrupt up to 10% of the samples [8].

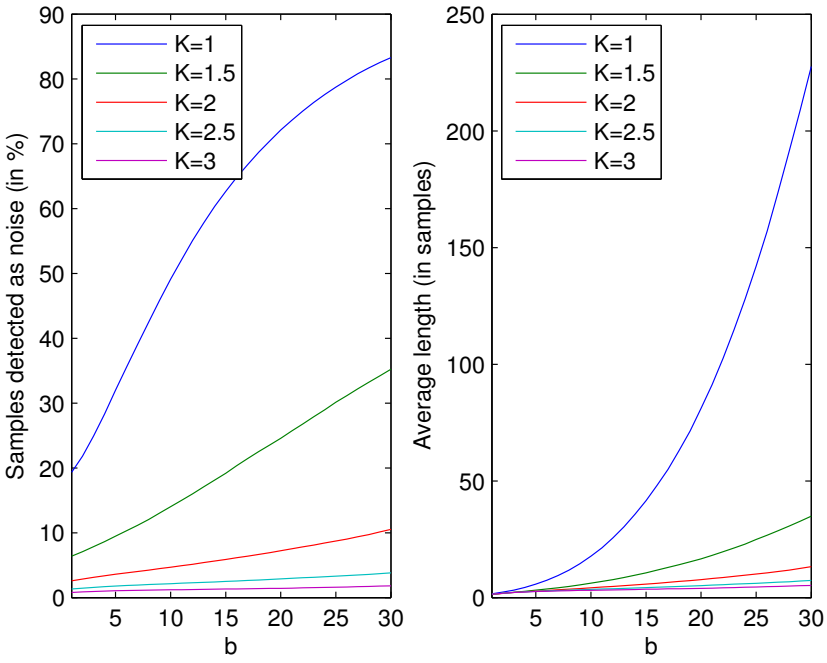

Figure 5: Number of samples detected as noise (in % of the total number of samples) and average length of the detected bursts for different values of K and b on a 30 seconds extract of classical music by Mussorgsky.

The evaluation of all these configurations is purely subjective and depends on the original uncorrupted signal (which is unfortunately not available). In particular, if the original signal is distorted then it is hard to distinguish the artifacts due to the reconstruction process (which tend to occur for large bursts) and the original distortions. Yet, some basic hearing tests can bring insightful results. For example:

- For $K = 1$ and $b = 1$, almost all the clicks are still present and the signal is slightly distorted (musical noise).
- For $K = 3$ and $b = 1$, it seems like no processing has been done.
- For $K = 1$ and $b = 30$, all of the clicks are removed but the signal is highly distorted (musical noise) and muffled.
- For $K = 3$ and $b = 30$, only some of the clicks are removed but the signal seems less distorted than in the previous configurations.

From these results, we can conclude that the parameter b is essential to our system since the interpolation process is only efficient when the bursts last for at least several samples. The threshold K allows us to control the intensity of the denoising process: nevertheless, we must remember that the method presented in this article only aims to remove impulsive noise and is not designed to deal with broadband noise.

Our tests on three other extracts (two jazz pieces by King Oliver and Louis Armstrong, one classical piece by Mozart, also used by Godsill) showed that a good compromise (clicks removed/no distortion) could be obtained with K around 2 and b around 20. For these 4 audio signals, the system tuned with these parameters detects between 5 and 7 % of the samples as noise, which seems realistic.

5.2 Influence of N_{iter}

On our four extracts, we realized that by using $K = 2$, $b = 20$ and two iterations, all the clicks were visibly removed without any obvious distortion. With three iterations the results were similar but after four or five iterations musical noise appears. Figure 6 presents a restoration example on one frame of classical music by Mussorgsky. We see in this example that after two iterations, most of the clicks are removed: the remaining degradations are either clicks of low amplitude or background noise.

We see in Figure 7 the details of the computation that can be found in the results page of the demo. In the first iteration of the algorithm, 7.23% of the samples were detected as noise, while on the second iteration, 7.54% were detected as noise. In all, 12.46% of the samples have been processed (note that the overlap between the first and second iteration is rather small: 2.31%).

6 Conclusion

The algorithm presented in this article is a simple yet efficient technique to restore audio signals corrupted by impulsive noise. The limited range of parameters and the fact that the detection and interpolation phases share common parameters make it possible to quickly process signals in an automatic way. The results show that our method is able to deal with different types of signals (classical, jazz, vocal) and different types of degradations.

Acknowledgment

Work partly founded by the European Research Council (advanced grant Twelve Labours 246961), and the Office of Naval research (ONR grant N00014-14-1-0023).

References

[1] E. Alvarez, R. Mendez, and G. Langwagen, *Detection of clicks using sinusoidal modeling for the confirmation of the clicks*, in Proceedings of the International Conference on Digital Audio Effects (DAFx), Naples, Italy, 2004, pp. 27–32.

[2] S. Canazza, G. De Poli, and G.A. Mian, *Restoration of audio documents by means of extended Kalman filter*, IEEE Transactions on Acoustics, Speech and Signal Processing, 18 (2010), pp. 1107–1115. http://dx.doi.org/10.1109/TASL.2009.2030005.

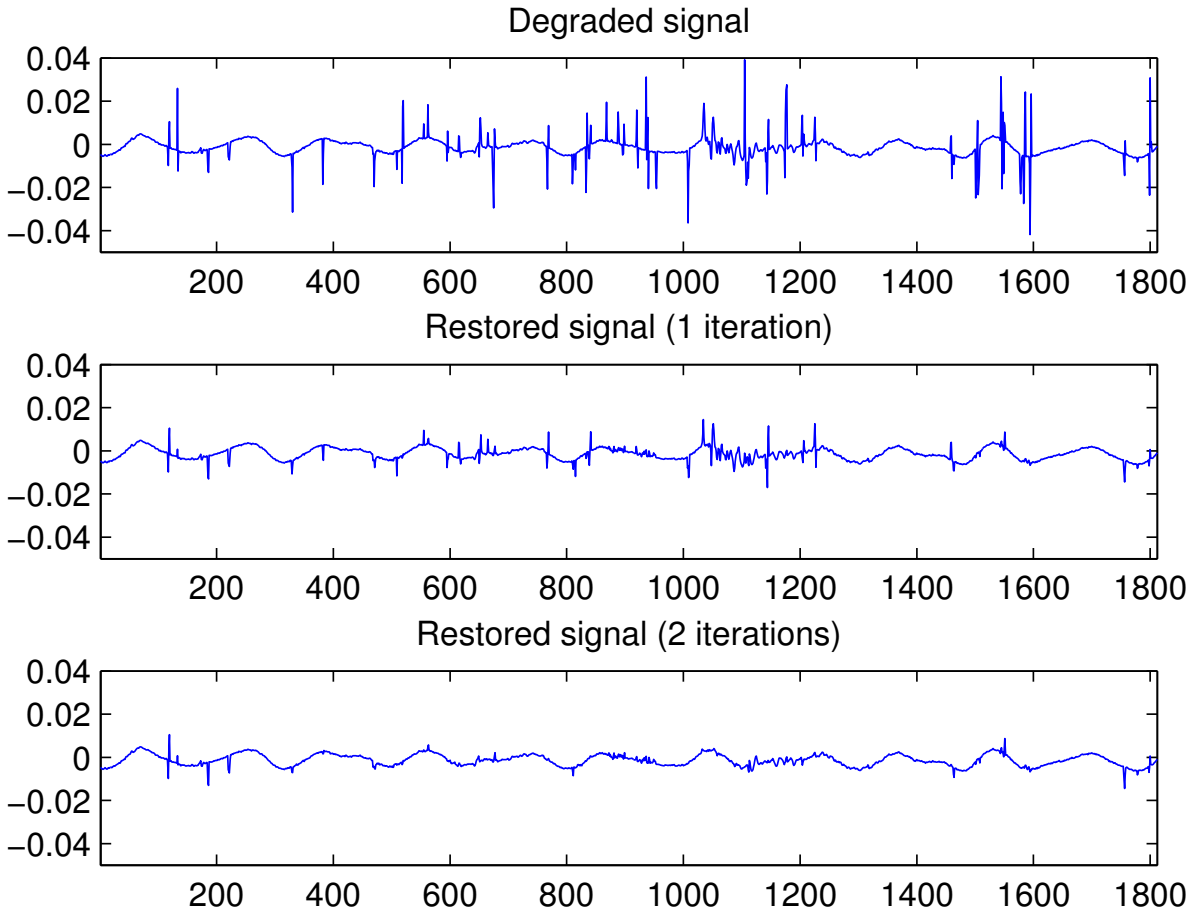

Figure 6: Restoration examples obtained on a degraded extract of classical music by Mussorgsky. The parameters used are $K = 2$, $b = 20$, $p = 302$, $N_w = 2416$ and the first and last p samples are not displayed as they are not part of the interpolation process.

[3] A. Czyzewski, *Some methods for detection and interpolation of impulsive distortions in old audio recordings*, in Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 1995, pp. 139–142. http://dx.doi.org/10.1109/ASPAA.1995.482976.

[4] P.A.A. Esquef, L.W.P. Biscainho, P.S.R. Diniz, and F.P. Freeland, *A double-threshold-based approach to impulsive noise detection in audio signals*, in Proceedings of the European Signal Processing Conference (EUSIPCO), Tempere, Finland, 2000, pp. 2014–2044.

[5] P. Esquef, M. Karjalainen, and V. Valimaki, *Detection of clicks in audio signals using warped linear prediction*, in Proceedings of the International Conference on Digital Signal Processing (DSP), 2002, pp. 1085–1088.

[6] S.J. Godsill and P.J.W. Rayner, *A Bayesian approach to the detection and correction of error bursts in audio signals*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, 1992, pp. 261–264. http://dx.doi.org/10.1109/ICASSP.1992.226070.

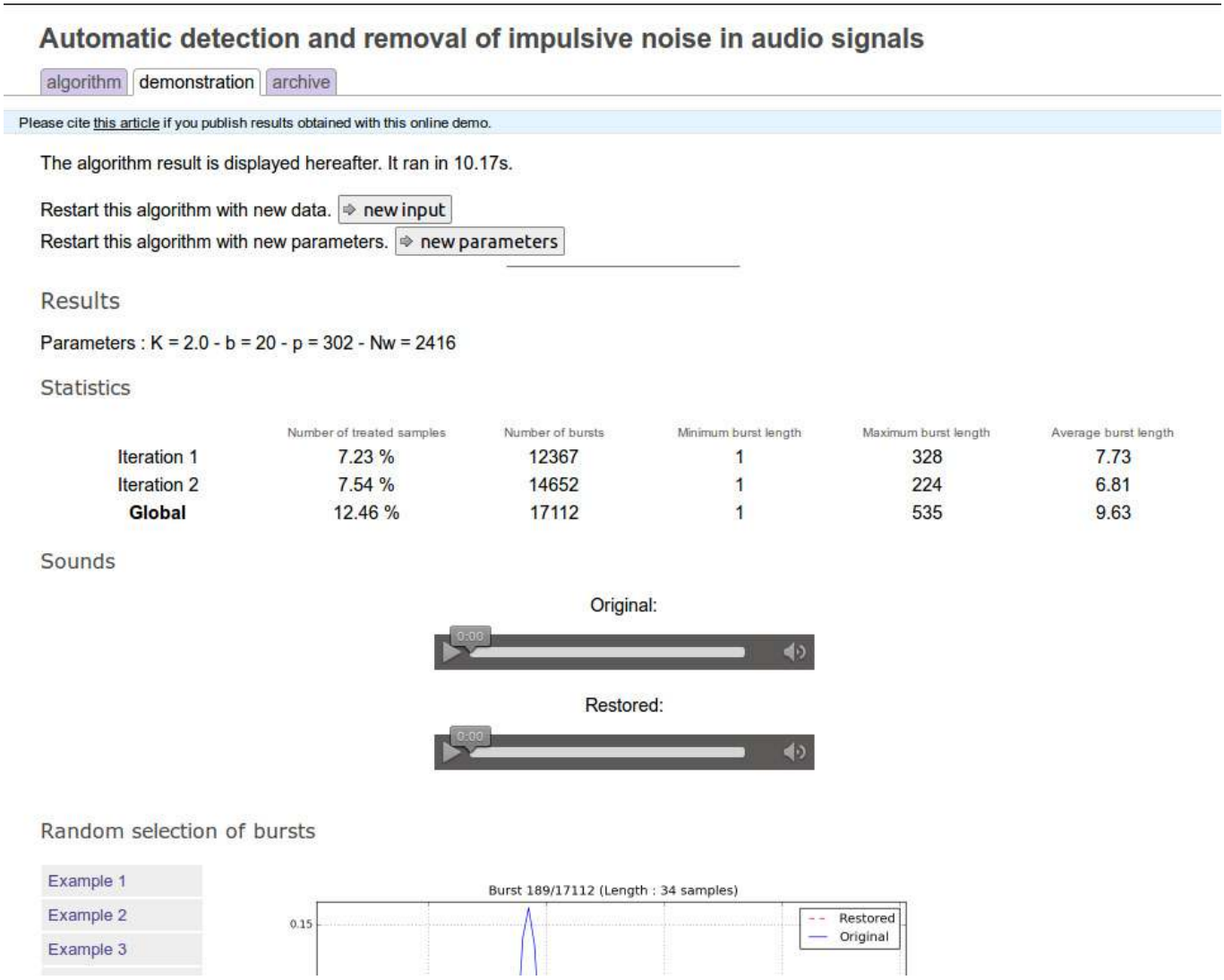

Figure 7: Output of the demo for the restoration of classical music by Mussorgsky. The parameters used are $K = 2$, $b = 20$, $p = 302$, $N_w = 2416$.

[7] ——, *A Bayesian approach to the restoration of degraded audio signals*, IEEE Transactions on Speech and Audio Processing, 3 (1995), pp. 267–278. http://dx.doi.org/10.1109/89.397091.

[8] ——, *Digital Audio Restoration - A Statistical Model-Based Approach*, Springer-Verlag London, 1998, ch. 5 - Removal of clicks, pp. 99–134.

[9] S.J. Godsill, P.J.W. Rayner, and O. Cappé, *Digital audio restoration*, vol. 437 of The International Series in Engineering and Computer Science, Springer, 2002, pp. 133–194. http://dx.doi.org/10.1007/0-306-47042-X_4.

[10] D. Griffin and J. Lim, *Signal estimation from modified short-time Fourier transform*, IEEE Transactions on Acoustics, Speech and Signal Processing, 32 (1984), pp. 236–243. http://dx.doi.org/10.1109/TASSP.1984.1164317.

[11] A. Janssen, R. Veldhuis, and L. Vries, *Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes*, IEEE Transactions on Acoustics, Speech and Signal Processing, 34 (1986), pp. 317–330. http://dx.doi.org/10.1109/TASSP.1986.1164824.

[12] I. Kauppinen, *Methods for detecting impulsive noise in speech and audio signals*, in Proceedings of the International Conference on Digital Signal Processing (DSP), 2002, pp. 967–970. http://dx.doi.org/10.1109/ICDSP.2002.1028251.

[13] N. Levinson, *The Wiener RMS (root mean square) error criterion in filter design and prediction*, Journal of Mathematics and Physics, 25 (1947), pp. 261–278.

[14] J. Murphy and S. Godsill, *Joint Bayesian removal of impulse and background noise*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011, pp. 261–264. http://dx.doi.org/10.1109/ICASSP.2011.5946390.

[15] M. Niedzwiecki and K. Cisowski, *Adaptive scheme for elimination of broadband noise and impulsive disturbances from AR and ARMA signals*, IEEE Transactions on Signal Processing, 44 (1996), pp. 528–537. http://dx.doi.org/10.1109/78.489026.

[16] L. Oudre, *Interpolation of missing samples in sound signals based on autoregressive modeling*, submitted to Image Processing On Line (IPOL).

[17] S.V. Vaseghi and P.J.W. Rayner, *A new application of adaptive filters for restoration of archived gramophone recordings*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New York, New York, USA, 1988, pp. 2548–2551. http://dx.doi.org/10.1109/ICASSP.1988.197163.

[18] SV. Vaseghi and PJW. Rayner, *Detection and suppression of impulsive noise in speech communication systems*, in IEEE Proceedings on Communications, Speech and Vision, vol. 137, 1990, pp. 38–46.