Postprint of: Niedźwiecki M., Ciołek M., Localization of impulsive disturbances in audio signals using template matching, Digital Signal Processing, Vol. 46 (2015), pp. 253-265, DOI: 10.1016/j.dsp.2015.07.009

Localization of Impulsive Disturbances in Audio Signals Using Template Matching

Maciej Niedźwiecki, *Senior Member, IEEE*, and Marcin Ciołek

***Abstract*— In this paper, a new solution to the problem of elimination of impulsive disturbances from audio signals, based on the matched filtering technique, is proposed. The new approach stems from the observation that a large proportion of noise pulses corrupting audio recordings have highly repetitive shapes that match several typical "patterns". In many cases a representative set of exemplary pulse waveforms can be extracted from the episodes of silence preceding and succeeding the recorded audio material. Based on such a set, a relatively small number of typical noise patterns, called click templates, can be established. To localize noise pulses, the appropriately modified click templates can be correlated with the sequence of one-step-ahead prediction errors yielded by the model-based signal predictor. It is shown that template matching is an efficient and computationally affordable disturbance localization technique – when combined with the classical detection method based on autoregressive modeling, it can improve restoration results. Since click templates can be created for a particular set of recordings, obtained using a particular audio equipment, an important feature of the proposed approach is its source adaptivity. Even though the paper is focused on restoration of archive recordings, the proposed approach is useful in a much wider context, e.g., it can be applied to elimination of impulsive disturbances corrupting telecommunication channels.**

***Index Terms*—outlier detection and elimination, adaptive signal processing, audio restoration, digital archives.**

I. Introduction

ARCHIVED audio recordings are often degraded by impulsive disturbances and wideband noise [1], [2]. Clicks, pops, ticks, crackles and record scratches are caused by aging and/or mishandling of the surface of gramophone records (shellac or vinyl), specks of dust and dirt, faults in the record stamping process (e.g. gas bubbles), and slight imperfections in the record playing surface due to the use of coarse grain filters in the record composition. In the case of magnetic tape recordings, impulsive disturbances can be usually attributed to transmission or equipment artifacts (e.g. electric or magnetic pulses).

Wideband background noise, such as the so-called surface noise of magnetic tapes and phonograph records, is an inherent component of all analog recordings.

This work was supported by the National Science Centre under the agreement UMO-2013/09/B/ST7/ 01582.

The authors are with the Faculty of Electronics, Telecommunications and Computer Science, Department of Automatic Control, Gdańsk University of Technology, Narutowicza 11/12, Gdańsk, Poland (e-mails: maciekn@eti.pg.gda.pl, marcin.ciolek@pg.gda.pl).

Elimination of both types of disturbances from archive audio documents is an important element of saving our cultural heritage. The Polish Radio Archives and the Polish National Library Archives alone contain more than one million archive audio documents with different content (historic speeches, interviews, concerts, studio music recordings etc.), saved on different media, such as piano rolls, phonograph and gramophone records, magnetic tapes etc. The British Library Sound Archive (which is among the largest collections of recorded sound in the world) holds over three million recordings, including over a million of disks and 200,000 tapes. Digitization of these documents is an ongoing process (in Poland carried out, among others, by the Polish National Digital Archives), which will be very soon followed by the next, obvious step – audio restoration. This makes research on audio restoration technology both practically useful and timely.

For the sake of simplicity, in this paper we will deal only with the problem of elimination of impulsive disturbances, i.e., we will assume that the sampled audio signal $y(t)$ has the form

$$y(t) = s(t) + \delta(t) \tag{1}$$

where $t = \ldots, -1, 0, 1, \ldots$ denotes normalized (dimensionless) discrete time, $s(t)$ denotes the undistorted (clean) audio signal, and $\delta(t)$ is the sequence of noise pulses. Later on we will comment how to modify the proposed approach so that it can also work in the presence of additive wideband noise.

Let $d(t)$ be the pulse location function

$$d(t) = \begin{cases} 1 & \text{if} \quad \delta(t) \neq 0 \\ 0 & \text{if} \quad \delta(t) = 0 \end{cases} .$$

The problem of elimination of impulsive disturbances is usually solved in two steps. First, noise pulses are localized. The resulting estimated pulse location function has the form

$$\widehat{d}(t) = \begin{cases} 1 & \text{if the sample is classified} \\ & \text{as an outlier} \\ 0 & \text{otherwise} \end{cases} .$$

Then, at the second stage of processing, all samples regarded as outliers $Y_\delta = \{y(t) : \widehat{d}(t) = 1\}$ are interpolated based on the approved samples $Y_s = \{y(t) : \widehat{d}(t) = 0\}$.

The majority of known approaches to elimination of impulsive disturbances from archive audio signals are based on adaptive prediction – the autoregressive (AR) or autoregressive moving average (ARMA) model of the analyzed signal is continuously updated and used to predict consecutive signal samples [3]–[14]. In the simplest case, further referred to as basic detection scheme, a "detection alarm" is raised, and the predicted sample is scheduled for reconstruction, whenever the absolute value

of the one-step-ahead prediction error becomes too large, namely when it exceeds a prescribed multiple of its estimated standard deviation. The test is then extended to multi-step-ahead prediction errors – detection alarm is terminated when a given number of samples in a row remain sufficiently close to the predicted signal trajectory (or when the length of detection alarm reaches its maximum allowable value). Finally, once the pulse is localized, the corrupted samples are interpolated (using the same signal model which served for detection purposes) based on the uncorrupted neighboring samples. A more sophisticated, Bayesian solution to the problem of noise pulse detection and signal reconstruction (also based on AR modeling) was presented in [6] and [7]. In both cases noise pulses were modeled as additive bursts of noise.

The basic detection scheme was subject to several modifications and extensions.

In [5] and [8] the task of simultaneous signal identification, outlier detection and signal reconstruction was stated as a nonlinear filtering problem and solved using the theory of extended Kalman filter (EKF). The EKF algorithm can be viewed as a combination of two Kalman filters coupled in a nonlinear fashion – the filter designed to track time-varying parameters of the AR signal model, and another one used for the purpose of detection and reconstruction of corrupted samples. As later shown in [9], applying the certainty equivalence projection technique one can partition the EKF algorithm into two weakly coupled subalgorithms responsible for model parameter tracking and signal monitoring/reconstruction, respectively. This has two important practical implications. First, the Kalman filter based parameter tracker can be replaced with a more convenient (easier to tune) exponentially weighted least squares (EWLS) algorithm [15], [16]. Second, the detection/reconstruction algorithm can be put down in the order-recursive form, which results in major computational savings.

Even though yielding satisfactory results when applied to archived music, the AR-model based reconstruction often fails on speech signals, especially those with strong voiced episodes. Since voiced speech sounds are formed by exciting the vocal tract (represented by the AR model) with a periodic train of glottal air pulses, the outlier detector can easily confuse pitch excitation with impulsive noise, which usually results in audible signal distortions. The problem mentioned above can be alleviated if the sparse autoregressive (SAR) model of the audio signal is used instead of the AR model [12]. SAR models capture both short-term correlations (formant structure) and long-term correlations (pitch structure) of the analyzed sound. Owing to this, unlike outlier detectors based on conventional AR models, detectors that incorporate SAR models usually do not confuse pitch-related pulses with noise pulses. This significantly reduces the number of false alarms. Restoration of stereo recordings can be performed by splitting left/right audio tracks and processing them separately. However, improved results can be obtained if both tracks are modeled jointly using the vector autoregressive (VAR) or sparse vector autoregressive (SAR) modeling approach [14]. The benefits of VAR/SVAR modeling can be observed both at the outlier detection stage (more accurate localization of noise pulses) and at the signal interpolation stage (the undistorted material in one track can be used to ”repair” the corrupted fragment in the other track).

When the archive audio signal is analyzed sequentially, forward in time, a sample is regarded as an outlier if it is “inconsistent” with the signal past, which is indicated by excessive values of prediction errors. When signal characteristics change abruptly, e.g. when an entirely new sound starts to build up, all causal detection schemes are prone to generate false detection alarms, calling in question uncorrupted signal samples simply because they do not match the signal past. Since such samples are consistent with the signal “future”, rather than its “past”, the number of false alarms can be significantly reduced if results of forward-time detection are combined with the analogous results of backward-time detection. The latter can be obtained by means of processing audio signal backward in time (provided, of course, that the entire recording is available). In addition to reducing the number and length of false alarms, bidirectional processing allows one to carve detection alarms more carefully (smaller number of overlooked noise pulses, better front/end matching of noise pulses). The set of local, case-dependent fusion rules that can be used to combine forward and backward detection alarms was proposed and experimentally verified in [13].

The common feature of the approaches summarized above is that they all incorporate outlier elimination schemes which do not rely on any information about the size and shape of noise pulses – even if such a prior knowledge is available. To the best of our knowledge, apart from the method described in [4], which focuses on very long disturbances such as record scratches, the only approach proposed so far, which incorporates prior knowledge about noise pulses into pulse detection/elimination procedure, is that described in the recent paper of Ávila and Biscainho [11]. The Bayesian pattern matching procedure proposed there is based on two sequentially sampled models: the AR model of the clean audio signal (with adjustable autoregressive coefficients and adjustable driving noise variance), and an explicit model of the impulsive disturbance (exponentially decaying pulse with adjustable location and shape parameters). The problem of joint detection and estimation of corrupted samples is solved by means of Gibbs sampling – the joint posterior distribution of the clean signal, its AR-model parameters and noise pulse parameters, is searched numerically using a variant of the Metropolis-Hastings algorithm. The resulting numerical iterative procedure is computationally very demanding.

The approach described in this paper is explicit and much simpler. It originates from the observation that in many cases a representative set of impulsive disturbances can be extracted, using simple detection techniques, from the episodes of silence preceding and succeeding the recorded audio material [e.g., separating successive tracks on long-playing records (LPs)]. Based on such a set, a relatively small number of typical noise patterns (click templates) can be established and further used for detection purposes.

The contribution of the paper is twofold. First, we propose a new method, based on analysis of the pulse similarity graph, that allows one to create the library of click templates. Second,

we show how typical noise patterns can be detected and localized in the archive recording using the matched filtering technique. We demonstrate that, when combined with the classical AR-model based detection methods, such approach can noticeably improve restoration results.
For clarity reasons, our presentation will be restricted to the sequential AR-model based noise pulse elimination scheme, which is suitable for on-line processing of music signals. Extension of the obtained results to sparse modeling and/or bidirectional processing is straightforward.

Remark: The matched filtering technique was proposed in early publications on elimination of impulsive disturbances [3], [4]. The authors of the above-mentioned papers analyzed an impact that an *idealized* (Kronecker-type) noise pulse has on the output of the AR-model based inverse filter. They suggested that in order to localize such pulses in the input (corrupted audio) signal, one could convolve the sequence of one-step-ahead signal prediction errors, yielded by the inverse filter, with the sequence made up of autoregressive coefficients (put in reverse order), and threshold the obtained results. Quite clearly, this approach does not incorporate any knowledge of typical noise patterns. It can only be used to isolate short unimodal pulses. The technique described in [4] is more along our lines, but it can be used only to isolate very long high-energy disturbances such as record scratches.

II. Creating click templates

While some noise pulses encountered in archive audio recordings have unique (and sometimes rather complicated) shapes, the majority of them form repeatable patterns which can be grouped in a relatively small number of classes represented by click templates. Typical shapes and duration of noise pulses may strongly depend on the recording medium (shellac, vinyl, magnetic tape), the way it was handled in the past (storage conditions, degree of wear), played back (pre-amplifier mode, turntable speed, type of stylus or tape deck), and digitized (sampling rate). Hence, the important feature of the proposed approach is its source adaptivity – click templates can be created for a particular group of recordings (e.g. those coming from a specific LP record) obtained using a particular audio equipment.

A. Extraction of Exemplary Noise Pulses

Exemplary noise pulses can be extracted from the silent parts of archive recordings preceding and/or succeeding the actual soundtracks. Extraction can be performed using any general purpose outlier detection scheme, e.g. by means of adaptive signal thresholding based on the 3-sigma rule.
To create reliable click templates, at least several hundreds of exemplary noise pulses should be collected. While for a single recording there might be not enough material for doing this, collections of many recordings stored on LPs or tapes make the click gathering task relatively easy.

B. Shape Similarity Analysis

The aim of this step is to asses degree of similarity between the extracted click waveforms. Similar waveforms will be grouped, normalized, time-aligned and averaged, forming click templates. As a tool for shape similarity analysis, we will use the quantity known in statistics as Pearson's correlation – the estimate of the correlation (normalized covariance) coefficient between two random variables X and Y. Based on K measurements of X and Y, written as $x(k)$ and $y(k)$, $k = 1, \ldots, K$, the Parson's correlation can be computed using the formula

$$\widehat{\rho}_{XY} = \frac{\sum_{k=1}^{K}[x(k)-\bar{x}][y(k)-\bar{y}]}{\sqrt{\sum_{k=1}^{K}[x(k)-\bar{x}]^2 \sum_{k=1}^{K}[y(k)-\bar{y}]^2}} = \sum_{k=1}^{K} \widetilde{x}(k)\widetilde{y}(k) \tag{2}$$

where

$$\bar{x} = \frac{1}{K}\sum_{k=1}^{K} x(k), \quad \bar{y} = \frac{1}{K}\sum_{k=1}^{K} y(k)$$

denote the estimates of mean values of X and Y, respectively, and

$$\widetilde{x}(k) = \frac{x(k)-\bar{x}}{\sqrt{\sum_{k=1}^{K}[x(k)-\bar{x}]^2}}, \quad \widetilde{y}(k) = \frac{y(k)-\bar{y}}{\sqrt{\sum_{k=1}^{K}[y(k)-\bar{y}]^2}}. \tag{3}$$

denote the normalized measurements. Parson's correlation takes the values in the interval $[-1, 1]$ and is scale-invariant. This makes it a very good tool for shape similarity assessment. Denote by $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_N\}$ the set consisting of N extracted noise pulses where

$$\mathcal{P}_i = \{p_i(1), \ldots, p_i(k_i)\}$$

is the sequence of samples, of length k_i, forming the i-th pulse. Denote by $\widetilde{\mathcal{P}}_i = \{\widetilde{p}_i(1), \ldots, \widetilde{p}_i(k_i)\}$ the sequence of normalized pulse samples, obtained in a way analogous to (3).
When comparing two waveforms, say $\mathcal{P}_i$ and $\mathcal{P}_j$, one should account for their (possibly) different length and lack of alignment. To find the best alignment, we will compute correlation-based similarity scores between $\widetilde{\mathcal{P}}_i$ and the sequence $\widetilde{\mathcal{P}}_j$ shifted by τ samples. Assuming that samples preceding $\widetilde{p}_j(1)$ and succeeding $\widetilde{p}_j(k_j)$ have zero values, the similarity score between $\mathcal{P}_i$ and $\mathcal{P}_j$ for the integer time shift τ can be expressed in the form

$$\rho_{ij}(\tau) = \sum_{\substack{k=1 \\ 1 \le k+\tau \le k_j}}^{k_i} \widetilde{p}_i(k)\widetilde{p}_j(k+\tau) = \sum_{k=\max(1,1-\tau)}^{\min(k_i,k_j-\tau)} \widetilde{p}_i(k)\widetilde{p}_j(k+\tau) \tag{4}$$

where $\tau \in \mathcal{T}_{ij} = [1-k_i, k_j-1]$. Note that the summation range in (4) accounts for differences in the length of the compared sequences. The entire set of correlation coefficients $\rho_{ij}(\tau), \tau \in \mathcal{T}_{ij}$ can be efficiently computed using the FFT-based convolution algorithm.

Denote by

$$\tau_{ij} = \arg\max_{\tau \in \mathcal{T}_{ij}} \rho_{ij}(\tau) \tag{5}$$

the time shift maximizing the similarity score, i.e., the one that guarantees the best alignment of $\widetilde{\mathcal{P}}_i$ and $\widetilde{\mathcal{P}}_j$. To measure the degree of similarity between $\mathcal{P}_i$ and $\mathcal{P}_j$, we will use the following correlation coefficient

$$r_{ij} = \max_{\tau \in \mathcal{T}_{ij}} \rho_{ij}(\tau) = \rho_{ij}(\tau_{ij}) \ . \tag{6}$$

C. Creation of Click Templates

Based on the set of correlation coefficients $\{r_{ij}, i, j = 1, \ldots, N\}$, one can build an undirected similarity graph G showing an internal similarity structure of the analyzed set of noise pulse extracts. This graph has N vertices corresponding to different click waveforms $\mathcal{P}_1, \ldots, \mathcal{P}_N$. If the degree of similarity between $\mathcal{P}_i$ and $\mathcal{P}_j$ is sufficiently high, namely, if $r_{ij} \geq \gamma$, where γ is a threshold close to 1, e.g. $\gamma = 0.95$, the vertices associated with $\mathcal{P}_i$ and $\mathcal{P}_j$ $(i \neq j)$ are connected by an edge. Hence, the adjacency matrix of G has the form

$$\mathbf{L} = [l_{ij}]_{N \times N}, \quad l_{ij} = \begin{cases} 1 & \text{if } r_{ij} \geq \gamma \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases} .$$

Click templates can be obtained by averaging click waveforms corresponding to maximum cliques of G, i.e., its maximum complete subgraphs[1]. The proposed procedure is recursive and can be summarized as follows:

Initialize: $\mathbf{L}_1 \leftarrow \mathbf{L}$, $G_1 \leftarrow G$, $i \leftarrow 1$.

Step 1: Search for the maximum clique Q_i of the graph G_i defined by $\mathbf{L}_i$. If there are several maximum cliques with the same number of vertices n_i, choose the one for which the sum of similarity scores r_{ij} takes the largest value (summation being carried over all edges of Q_i). Alternatively, use a computationally more involved algorithm for finding the weighted maximum clique. If the size of the clique Q_i is sufficiently large, e.g. if $n_i \geq 10$, continue to **Step 2** – otherwise **Stop**.

Step 2: Remove from G_i all vertices and edges of Q_i, forming a new graph G_{i+1} with adjacency matrix $\mathbf{L}_{i+1}$ ($\mathbf{L}_{i+1}$ can be obtained by zeroing the corresponding rows and columns of $\mathbf{L}_i$). Set $i \leftarrow i + 1$ and return to **Step 1**.

Once all cliques of sufficient size are found, their "centers" are localized. Denote by $\mathcal{S}_i = \{\widetilde{\mathcal{P}}_j, j \in J_i\}$ the set of normalized pulse waveforms associated with the clique Q_i (J_i is the set indicating which vertices of G belong to Q_i). The central element of $\mathcal{S}_i$, denoted by $\widetilde{\mathcal{P}}_{j_i}$, is the one for which the sum of outgoing edge weights (similarity scores) is maximized

$$j_i = \arg\max_{j \in J_i} \sum_{\substack{l \in J_i \\ l \neq j}} r_{jl} \ .$$

Such element can be interpreted as the one that is "most similar" to the remaining elements of $\mathcal{S}_i$.

[1]Every two vertices of a complete (sub)graph must be connected by an edge. The maximum subgraph is the one with the largest number of vertices.

All waveforms grouped in $\mathcal{S}_i$ are extended with zeros on both sides, aligned with respect to the central waveform $\widetilde{\mathcal{P}}_{j_i}$, and averaged. Note that the optimal alignment shifts $\tau_{j_i l}, l \in J_i$ were already computed at the pre-processing stage, cf. (5). Since averaging shows tendency to create long tails (small but non-zero values preceding and succeeding the main pulse activity), and since such tails have a marginal impact on the subsequent shape similarity analysis, click templates are obtained by trimming the averaged waveforms, namely by removing from their beginning and end all samples with absolute values smaller than 5% of the peak value.

Remark 1: When the length of a noise pulse is too short, shape matching becomes an ill-posed problem. For example, when a pulse of length 1 (Kronecker-type) is compared with *any* signal fragment of length 1 (isolated sample), the similarity score takes always its maximum value equal to 1, losing its discriminative value. For this reason the minimum length of click templates was restricted to 4 – all shorter templates were eliminated. Note, however, that short noise pulses can be easily handled by classical outlier detectors – see Section 4A for further details.

Remark 2: All maximum cliques can be efficiently searched using the algorithm described in [17] and based on the well-known Bron-Kerbosch maximal clique finding algorithm [18] (the MATLAB code `maximalcliques.m` is available from the Mathworks repository `mathworks.com/matlabcentral`).
When the search is restricted to just one (any) maximum clique, much faster algorithms are available – see e.g. [19] (the C++ code can be found on the first author's web page `sicmm.org/~konc/maxclique`).

Remark 3: The method of matched filtering can work without creating click templates. In such a case, each time a noise pulse is detected, one should correlate the sequence of prediction errors with every waveform contained in the click database. Since click databases may consist of hundreds of exemplary click waveforms, such a brute-force approach would be computationally prohibitive. Additionally, since many click waveforms match, with high degree of accuracy, just a few typical patterns, using click templates seems to be a pretty natural and elegant solution.

D. Source Adaptivity

Both the length and the shape of click waveforms may strongly depend on the source of audio material. Fig. 1 shows the first 14 click templates obtained by means of processing 500 pulse waveforms extracted from an old gramophone record. Fig. 2 shows an analogous set of templates obtained for an archive magnetic tape recording corrupted with electrical interference noise. In both cases sampling rate was equal to 48 kHz. Note that while most of the typical gramophone clicks are unimodal or bimodal, the electrical clicks usually form characteristic oscillatory patterns. The advantage of the proposed approach is its ability to incorporate such source-specific knowledge into the process of detection of impulsive disturbances.

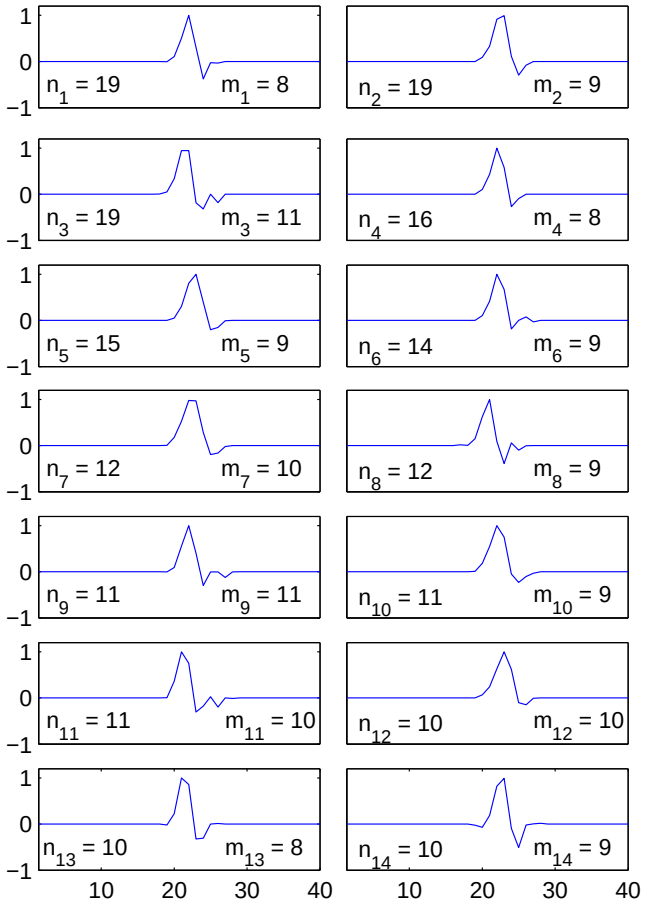

Fig. 1. A collection of 14 click templates obtained for an old gramophone recording. Information about the size of the corresponding clique (n_i) and the length of the click template (m_i) is displayed below each plot. To preserve the original look of noise pulses all waveforms were amplitude-normalized but not debiased.

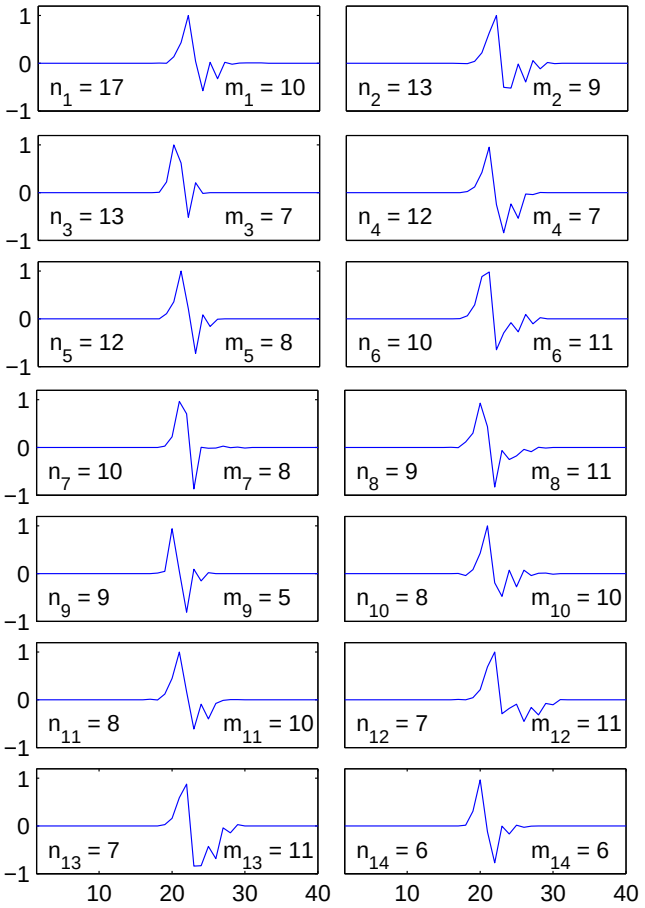

Fig. 2. A collection of 14 click templates obtained for a magnetic tape recording corrupted with electrical interference noise. Information about the size of the corresponding clique (n_i) and the length of the click template (m_i) is displayed below each plot. To preserve the original look of noise pulses all waveforms were amplitude-normalized but not debiased.

III. Detection of Typical Noise Patterns

Our procedure for localization of click templates in audio signals will be based on the technique known in telecommunications as matched filtering. Classical matched filtering is used to detect known symbols transmitted over a noisy channel, i.e., buried in additive white measurement noise [20]. This can be achieved by correlating symbol templates with the received signal and thresholding the obtained results. When the measurement noise is not white, the matched filtering technique can be still used, provided that the analyzed signal is whitened prior to template matching. We will use this approach to localize typical noise patterns.

A. Starting the Matching Procedure

We will assume that the noiseless audio signal $s(t)$ obeys the following AR model of order r

$$s(t) = \sum_{j=1}^{r} a_j s(t-j) + n(t) \tag{7}$$

where $a_j, j = 1, \ldots, r$, denote autoregressive coefficients and $n(t)$ denotes zero-mean white driving noise with variance σ_n^2.

To simplify the presentation, we will assume that parameters of the AR model are known at the moment of carrying out the detection procedure. The adaptive detection procedure can be obtained by replacing the known model parameters with their most recent estimates, provided by the signal identification algorithm – see Section 5.

The minimum-variance one-step-ahead prediction of $s(t)$ is given by

$$\widehat{s}(t+1|t) = \sum_{j=1}^{r} a_j s(t-j+1).$$

When the past r measurements are outlier-free, it also holds that

$$\widehat{y}(t+1|t) = \sum_{j=1}^{r} a_j y(t-j+1).$$

Detection procedure is started each time the outlier alarm is raised, i.e., when the magnitude of the one-step-ahead prediction error $\varepsilon(t+1|t) = y(t+1) - \widehat{y}(t+1|t)$ exceeds μ times its standard deviation $\sigma_\varepsilon(t+1|t) = \sigma_n$

$$|\varepsilon(t+1|t)| > \mu\sigma_\varepsilon(t+1|t) \tag{8}$$

where μ is the detection threshold multiplier determined experimentally (usually the best results are obtained for $\mu \in [3, 5]$; $\mu = 3$ corresponds to the well-known "3-sigma" rule used for detection of outliers in Gaussian signals).

B. Whitening

When $s(t)$ obeys the AR model, whitening can be achieved by passing $y(t)$ through the corresponding inverse filter $A(q^{-1}) = 1 - \sum_{j=1}^{r} a_j q^{-j}$, where q^{-1} denotes the backward shift operator. According to (1) and (7), it holds that

$$\varepsilon(t|t-1) = A(q^{-1})y(t) = n(t) + \delta^f(t) \tag{9}$$

where $\delta^f(t) = A(q^{-1})\delta(t)$. This means that the problem of detection of typical noise patterns in the signal $y(t)$ can be reformulated as a problem of detection of suitably modified patterns (original patterns "shaped" by the inverse filter) in the prediction error signal $\varepsilon(t|t-1)$. Since the term $n(t)$, appearing in (9), denotes white noise, the second formulation is consistent with the classical matched filtering problem statement.

Prior to applying the matched filtering procedure to the sequence of prediction errors, one should create a new set of templates, further referred to as secondary templates, which reflect the changes introduced to impulsive noise patterns by the whitening filter.

Denote by

$$\mathcal{C}_i = \{\widetilde{c}_i(1), \ldots, \widetilde{c}_i(m_i)\}, \quad 1 \le i \le L$$

the i-th primary template (the averaged and normalized click waveform) and by $M = \max_{1 \le i \le L} m_i$ – the length of the longest template.

Let

$$\bar{c}_i(k) = \begin{cases} 0 & k \le 0 \\ \widetilde{c}_i(k) & 1 \le k \le m_i \\ 0 & k > m_i \end{cases} \tag{10}$$

and

$$c_i^f(k) = \bar{c}_i(k) - \sum_{j=1}^{r} a_j \bar{c}_i(k-j), \quad k = 1, \ldots, k_i \tag{11}$$

where $k_i = m_i + r$.

The set of secondary templates

$$\mathcal{C}_i^f = \{\widetilde{c}_i^f(1), \ldots, \widetilde{c}_i^f(k_i)\}, \quad 1 \le i \le L$$

can be obtained by means of normalizing the sequences $\{c_i^f(1), \ldots, c_i^f(k_i)\}$ generated using (11). The length of the longest secondary template will be denoted by K: $K = \max_{1 \le i \le L} k_i = M + r$.

A typical set of secondary templates, obtained by inverse filtering and normalizing primary templates shown in Fig. 1, is depicted in Fig. 3.

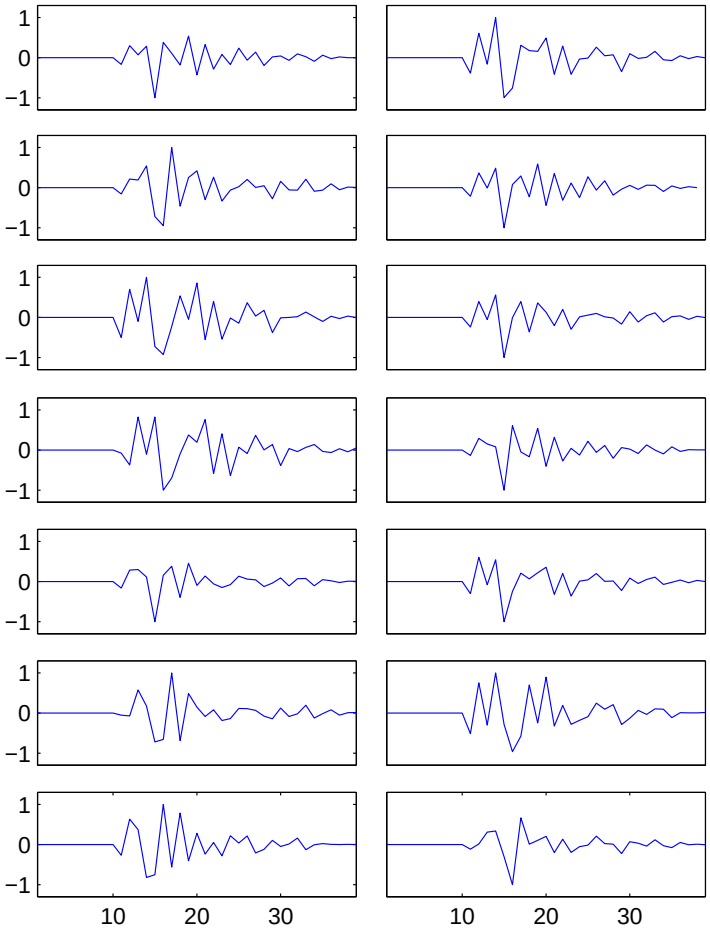

Fig. 3. A typical collection of 14 secondary click templates.

C. Pattern Detection

In accordance with our earlier findings, when (8) holds true, i.e., when detection alarm is raised at the instant $t+1$, the sequence of l "past" (where l denotes a small integer number) and K "future" one-step-ahead prediction errors

$$\{\varepsilon(t-l+1|t-l), \ldots, \varepsilon(t+K|t+K-l)\}$$

will be checked for the presence of appropriately time-shifted secondary templates $\mathcal{C}_i^f$. This can be achieved by computing the corresponding similarity scores

$$g_i(\tau) = \sum_{k=1}^{k_i} \widetilde{c}_i^f(k)\widetilde{\varepsilon}(t_0+\tau+k|t_0+\tau+k-1) \tag{12}$$

where $t_0 = t-l$, $\tau \in [0, l]$ is an integer number which denotes the alignment shift, and

$$\{\widetilde{\varepsilon}(t_0+\tau+1|t_0+\tau), \ldots, \widetilde{\varepsilon}(t_0+\tau+k_i|t_0+\tau+k_i-1)\}$$

is the sequence of normalized prediction errors [note that normalization, governed by (3), must be performed independently for each value of τ]. Time alignment is necessary to account for uncertainties embedded in triggering the detection alarm, i.e., determining the moment at which the matching process should start. When the shape similarity test is run only for $l = \tau = 0$, the results deteriorate.

Denote by τ_i the optimal alignment shift for the template $\mathcal{C}_i^f$

$$\tau_i = \arg \max_{\tau \in [0,l]} |g_i(\tau)|.$$

Note that, in order to make results insensitive to polarity of the detected noise pulses, verification is based on checking the absolute value of the similarity score.
The best-matching template $\mathcal{C}_{i_0}^f$ is the one that maximizes the optimized similarity score

$$i_0 = \arg \max_{1 \le i \le L} |g_i(\tau_i)|.$$

D. Verification

After finding the best-matching template, some sort of verification is needed to confirm that the best match is "sufficiently good". To assure that this is the case, it will be required that

$$|g_{i_0}(\tau_{i_0})| \ge \gamma_0 \tag{13}$$

where $\gamma_0 < \gamma$ is the similarity threshold (to account for the presence of noise, γ_0 is set to a smaller value than γ).
When the condition (13) is met, the detected noise pulse will be regarded as typical, matching the template $\mathcal{C}_{i_0}$, otherwise it will be classified as atypical and handled differently.

IV. Localization and interpolation of corrupted signal samples

A. Localization of Noise Pulses

The detection scheme described in Section 3 should be regarded as an *extension* of the classical AR-model based detection approach, rather than its replacement. Whenever noise pulse has a typical shape, matching one of the click templates, its localization can be usually done more precisely using matched filtering than using the classical general purpose scheme.
1) Atypical Patterns: When the noise pulse detected at the instant t does not match any of the templates, the classical prediction-based approach is used, i.e., the test is extended to multi-step-ahead prediction errors $\varepsilon(t+k|t) = y(t+k) - \widehat{y}(t+k|t)$, $k > 1$, where

$$\widehat{y}(t+k|t) = \sum_{j=1}^{r} a_j \widehat{y}(t+k-j|t), \quad k > 1$$

and $\widehat{y}(t+k|t) = y(t+k)$ for $k \le 0$. The detection alarm, started at the instant $t+1$, is terminated at the instant $t+n+1$ if r consecutive prediction errors are sufficiently small

$$|\varepsilon(t+n+j|t)| \le \mu \sigma_\varepsilon(t+n+j|t), \quad j = 1, \ldots, r$$

or if the length n of the detection alarm reaches its maximum allowable value denoted by $n_{\max}$. Hence, the corresponding detection alarm forms a solid block of "ones"

$$\widehat{d}(k) = 1 \ \text{ for } \ k \in \widehat{D}_t = [\widehat{t}_{\mathrm{B}}, \widehat{t}_{\mathrm{E}}]$$
$$\widehat{t}_{\mathrm{B}} = t+1, \ \widehat{t}_{\mathrm{E}} = t+n \ .$$

The multi-step-ahead prediction error variances $\sigma_\varepsilon^2(t+k|t)$, $k > 1$, can be evaluated using the recursive algorithm proposed by Stoica [22]. See [13] for a more detailed description of the entire detection procedure.
In order to achieve further performance improvements, the simple detection scheme described above, based on the open-loop multiple-step-ahead signal prediction, can be replaced with a more sophisticated scheme based on decision-feedback prediction. In this case prediction errors $\varepsilon(t+j|t+j-1)$ and the corresponding standard deviations $\sigma_\varepsilon(t+j|t+j-1)$ are evaluated on-line by the Kalman filtering algorithm which takes into account its earlier accept/reject decisions, i.e., decisions taken at the instants $t+1, \ldots, t+j-1$. The stopping rule is the same as in the open-loop approach. For a more detailed description of this technique see [9].
2) Typical Patterns: When a particular noise template $\mathcal{C}_{i_0}$ is detected, the classical outlier detection procedure is not pursued and the detection alarm has the form

$$\widehat{d}(k) = 1 \ \text{ for } \ k \in \widehat{D}_t = [\widehat{t}_{\mathrm{B}}, \widehat{t}_{\mathrm{E}}]$$
$$\widehat{t}_{\mathrm{B}} = t_0 + \widehat{\tau}_{i_0} + 1, \ \widehat{t}_{\mathrm{E}} = t_0 + \widehat{\tau}_{i_0} + m_{i_0} \ .$$

B. Interpolation of Distorted Signal Samples

In the classical approach, applied to atypical noise pulses, the corrupted signal samples $y(\widehat{t}_{\mathrm{B}}), \ldots, y(\widehat{t}_{\mathrm{E}})$ are interpolated based on r samples preceding and r samples succeeding the reconstructed fragment – the details can be found e.g. in [23] and [13].
When the detected noise pulse matches one of the click templates, one can choose between two reconstruction options: interpolation (as in the classical approach) or compensation. In the second case the clean signal is recovered by subtracting from the corrupted fragment the appropriately modified (scaled and bias-corrected) click template.
Suppose that the noise pulse shaped by the inverse filter coincides with the time-shifted, scaled and bias-corrected secondary template $\mathcal{C}_i^f$, namely

$$\delta^f(t_0 + \tau_i + k) = \alpha_i \widetilde{c}_i^f(k) + \beta_i, \quad k = 1, \ldots, k_i \tag{14}$$

where α_i and β_i denote the scale and bias correction coefficients. Let

$$\mathbf{w}_i^f = \begin{bmatrix} w_i^f(1) \\ \vdots \\ w_i^f(k_i) \end{bmatrix}, \quad \mathbf{e}_i = \begin{bmatrix} e_i(1) \\ \vdots \\ e_i(k_i) \end{bmatrix}, \quad \boldsymbol{\eta}_i = \begin{bmatrix} \eta_i(1) \\ \vdots \\ \eta_i(k_i) \end{bmatrix}$$
$$\mathbf{r}_i = \begin{bmatrix} \widetilde{c}_i^f(1) \\ \vdots \\ \widetilde{c}_i^f(k_i) \end{bmatrix}, \quad \mathbf{h}_i = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

where

$$w_i^f(k) = \delta^f(t_0 + \tau_i + k)$$
$$e_i(k) = \varepsilon(t_0 + \tau_i + k | t_0 + \tau_i + k - 1)$$
$$\eta_i(k) = n(t_0 + \tau_i + k)$$
$$k = 1, \ldots, k_i.$$

Using the vector notation introduced above, one can rewrite (14) in the form

$$\mathbf{w}_i^f = \alpha_i \mathbf{r}_i + \beta_i \mathbf{h}_i. \tag{15}$$

Note also that, according to (9),

$$\mathbf{e}_i = \boldsymbol{\eta}_i + \mathbf{w}_i^f. \tag{16}$$

Based on (15) and (16), the least squares estimates of the coefficients α_i and β_i can be obtained by minimizing the quadratic cost function

$$J_i(\alpha_i, \beta_i) = \| \mathbf{e}_i - \alpha_i \mathbf{r}_i - \beta_i \mathbf{h}_i \|^2 .$$

Let

$$\{\widehat{\alpha}_i, \widehat{\beta}_i\} = \arg\min_{\alpha_i, \beta_i} J(\alpha_i, \beta_i) .$$

It can be easily shown that

$$\widehat{\alpha}_i = \frac{f_1 f_5 - f_3 f_4}{f_1 f_2 - f_3^2} , \quad \widehat{\beta}_i = \frac{f_2 f_4 - f_3 f_5}{f_1 f_2 - f_3^2}$$

where the coefficients $f_1, \ldots, f_5$ are given by

$$\begin{aligned}
f_1 &= \| \mathbf{h}_i \|^2 = k_i \\
f_2 &= \| \mathbf{r}_i \|^2 = \sum_{k=1}^{k_i} [\widetilde{c}_i^f(k)]^2 \\
f_3 &= \mathbf{h}_i^{\mathrm{T}} \mathbf{r}_i = \sum_{k=1}^{k_i} \widetilde{c}_i^f(k) \\
f_4 &= \mathbf{h}_i^{\mathrm{T}} \mathbf{e}_i = \sum_{k=1}^{k_i} e_i(k) \\
f_5 &= \mathbf{r}_i^{\mathrm{T}} \mathbf{e}_i = \sum_{k=1}^{k_i} \widetilde{c}_i^f(k) e_i(k).
\end{aligned}$$

Since template coefficients are normalized, it holds that $f_2 = 1$ and $f_3 = 0$, leading to

$$\widehat{\alpha}_i = f_5 , \quad \widehat{\beta}_i = \frac{f_4}{k_i}.$$

Using the above estimates, one arrives at the following estimate of the secondary pulse waveform

$$\widehat{w}_i^f(k) = \widehat{\delta}^f(t_0 + \tau_i + k) = \widehat{\alpha}_i \widetilde{c}_i^f(k) + \widehat{\beta}_i, \quad k = 1, \ldots, k_i.$$

Finally, to obtain the estimates of the primary pulse waveform

$$\widehat{w}_i(k) = \widehat{\delta}(t_0 + \tau_i + k), \quad k = 1, \ldots, m_i$$

one should undo the changes introduced by the inverse filter. The corresponding recursive formula takes the form

$$\widehat{w}_i(k) = \widehat{w}_i^f(k) + \sum_{j=1}^{m_i} \widehat{w}_i(k-j), \quad k = 1, \ldots, m_i$$

where $\widehat{w}_i(k) = 0$ for $k \leq 0$.

The procedure summarized above should be performed *only* for the best-matching template $\mathcal{C}_{i_0}^f$. The compensation that follows takes the form

$$\widehat{s}(k) = y(k) - \widehat{\delta}_{i_0}(k), \quad k \in \widehat{D}_t \tag{17}$$

Our experiments have shown that interpolation yields better results than compensation. For this reason the compensation approach is not recommended – see Section 7 for more details.

V. Adaptive Detection

So far we have assumed that parameters of the AR signal model are constant and known. The adaptive detection, matching and interpolation procedures can be obtained by replacing the autoregressive coefficients $a_1, \ldots, a_r$ and the driving noise variance σ_n^2 with their estimates [$\widehat{a}_1(t), \ldots, \widehat{a}_r(t)$ and $\widehat{\sigma}_n^2(t)$, respectively] yielded by the finite-memory signal identification/tracking algorithm, such as the well-known EWLS algorithm. Let

$$\boldsymbol{\theta} = \begin{bmatrix} a_1 \\ \vdots \\ a_r \end{bmatrix}, \quad \boldsymbol{\varphi}(t) = \begin{bmatrix} y(t-1) \\ \vdots \\ y(t-r) \end{bmatrix}.$$

The EWLS estimator minimizes the exponentially weighted sum of squared modeling errors

$$\widehat{\boldsymbol{\theta}}(t) = \arg\min_{\boldsymbol{\theta}} \sum_{k=1}^{t} \lambda^{t-k} [y(k) - \boldsymbol{\varphi}^{\mathrm{T}}(k)\boldsymbol{\theta}]^2$$

where λ, $0 < \lambda < 1$, denotes the so-called forgetting constant. The value of λ should be chosen so as to trade off the bias and variance components of the mean-squared parameter tracking error [15], [16].

The recursive algorithm for computation of $\widehat{\boldsymbol{\theta}}(t)$ has a well-known form

$$\begin{aligned}
\varepsilon(t|t-1) &= y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\widehat{\boldsymbol{\theta}}(t-1) \\
\widehat{\boldsymbol{\theta}}(t) &= \widehat{\boldsymbol{\theta}}(t-1) + \mathbf{k}(t)\varepsilon(t|t-1) \\
\mathbf{k}(t) &= \frac{\mathbf{P}(t-1)\boldsymbol{\varphi}(t)}{\lambda + \boldsymbol{\varphi}^{\mathrm{T}}(t)\mathbf{P}(t-1)\boldsymbol{\varphi}(t)} \\
\mathbf{P}(t) &= \frac{1}{\lambda}[\mathbf{I} - \mathbf{k}(t)\boldsymbol{\varphi}^{\mathrm{T}}(t)]\mathbf{P}(t-1)
\end{aligned} \tag{18}$$

The exponentially weighted estimate of the driving noise variance can be obtained using the following recursive formula

$$\widehat{\sigma}_n^2(t) = \lambda_0 \widehat{\sigma}_n^2(t-1) + (1-\lambda_0)\varepsilon^2(t|t-1) \tag{19}$$

where λ_0, $0 < \lambda_0 < 1$, is a forgetting constant, usually smaller than λ.

The order of the autoregression r can be fixed or chosen adaptively using the generalized Akaike's criterion [21].

The algorithms (18) and (19) are run as long as signal measurements are regarded as outlier-free. When detection alarm is raised, e.g. if $\widehat{d}(t+1) = 1$, parameter estimation is temporarily stopped. Estimation is resumed at the instant $t + n + r$, where n denotes the length of detection alarm.

Remark: When adaptive detection procedure is carried out, the true values of the AR coefficients appearing in (11) are replaced with their most recent estimates $\widehat{a}_1(t), \ldots, \widehat{a}_r(t)$. This means that, unlike primary click templates, secondary templates are model parameter dependent, i.e., they must be re-established each time a new noise pulse is detected.

VI. Alarm Extension Technique

Some modifications are recommended to address the problem of "soft" pulse edges. As argued in [13], typical geometry of local damages of the recording medium (e.g. groove damages) results in pulses that usually start and end in a "gentle" way – abrupt changes, which constitute the main "body" of a click, are preceded and succeeded by much smaller but systematic changes resulting in smooth pre-click and post-click signal distortions. This effect becomes more pronounced when the sampling rate grows. In the presence of soft edges, detection alarms are seldom triggered at the very beginning of noise pulses, which may result in small but audible distortions of the reconstructed audio material. The problem described above can be alleviated by decreasing the detection multiplier μ, i.e., by making the detector more sensitive to unpredictable signal changes. This, however, may dramatically increase the number and length of detection alarms, causing the overall degradation of the results. A practical solution, proposed in [13] and recommended also here, is to shift back the front edge of detection alarm (once triggered) by a small, fixed number of samples Δ_1. This means that if detection alarm is raised at the instant $t+1$, the template matching procedure is initialized at the instant $t-\Delta_1+1$ instead of $t+1$. $\Delta_1 = 4$ is usually a good choice for 44.1 and 48 kHz recordings.

For the same reason (remember that click templates are created by trimming the average pulse waveforms), once detection alarm based on template matching is determined, it is beneficiary to widen it prior to interpolation by moving back its front edge, and moving forward its back edge by a small, fixed number of samples Δ_2: $\widehat{t}_{\mathrm{B}} \leftarrow (\widehat{t}_{\mathrm{B}} - \Delta_2)$, $\widehat{t}_{\mathrm{E}} \leftarrow (\widehat{t}_{\mathrm{E}} + \Delta_2)$. For 44.1 kHz and 48 kHz recordings $\Delta_2 = 1$ is our experimentally verified choice.

VII. Experimental results

The paper is illustrated with the results of objective tests, carried out on clean audio signals corrupted with real impulsive disturbances, and subjective (listening) tests performed on real archive gramophone recordings. All audio files and the results of their processing are available at `http://eti.pg.edu.pl/katedra-systemow-automatyki/badanie`.

A. Artificially Corrupted Audio Files

Our repository of clicks was made up of 1000 click waveforms extracted from an old gramophone record (under 48 kHz sampling) and randomly divided into two sets $\mathcal{P}_A$ and $\mathcal{P}_B$ (each containing $N = 500$ clicks), which were further used for training and validation purposes, respectively. Clicks were extracted using the bidirectional processing algorithm described in [13], which is very precise in determining both the beginning and end points of each noise pulse.

Based on the training set $\mathcal{P}_A$, 14 click templates, shown in Fig. 1, were established ($\gamma = 0.95$). Only a few seconds were needed to complete this task on a standard PC. The information about the size of the corresponding clique (n_i) and the length of click template (m_i) is displayed beneath each plot depicted in Fig. 1. Note that $\sum_{i=1}^{14} n_i = 130$, which means that 26% of all extracted noise waveforms were utilized in the process of formation of click templates.

Our audio test base consisted of 60 clean recordings contaminated with click waveforms randomly drawn from the set $\mathcal{P}_B$. Clean audio recordings contained from 23 to 29 seconds of classical music sampled at the rate of 48 kHz: 29 fragments of jazz music (15 vocal, 14 instrumental) and 31 fragments of classical music (23 instrumental, 3 choir, 5 opera). The audio material was chosen so as to cover different temporal and spectral features of audio signals. Clicks were picked at random from the set $\mathcal{P}_B$ and added every 300 signal samples. Such a regular spacing between consecutive noise pulses was a deliberate choice as regularly occurring signal distortions/imperfections are more audible than those appearing in irregular time constellations.

Performance evaluation was made for 4 approaches: the open-loop prediction based approach (A), the open-loop prediction based approach combined with template matching (A^*), the decision-feedback prediction based approach (B), and the decision-feedback prediction based approach combined with template matching (B^*).

All compared detection/reconstruction algorithms incorporated AR models of order $r = 20$. Signal identification was carried out using the exponentially weighted algorithms (18) and (19), equipped with the forgetting factors $\lambda = 0.995$ and $\lambda_0 = 0.991$, respectively. The detection multiplier was set to $\mu = 4.5$, the validation threshold was set to $\gamma_0 = 0.8$, and the alarm extension parameters – to $\Delta_1 = 4$ and $\Delta_2 = 1$. The number of pre-alarm prediction errors involved in template matching was set to $l = 4$.

Prior to comparing detection efficiency of different approaches, the adaptive AR-model based interpolation algorithm, supported with information about the exact location of inserted clicks [$\widehat{d}(t) \equiv d(t)$] was run on each of 60 test recordings and the results were evaluated via listening tests. The purpose of this "ground truth" experiment was to check how much signal interpolation alone (carried out in the presence of perfect detection of inserted noise pulses) affects the final reconstruction results. Since in all cases listening tests reported no audible difference between the original audio material and the reconstructed one, it was clear that all audible distortions (if any) observed later, when adaptive interpolation was combined with adaptive detection, must have been caused by detection errors, such as missing detections, inaccurate detections and false detections.

To evaluate performance of different detection/reconstruction algorithms, we used the perceptual evaluation of audio quality (PEAQ) tool [24], [25]. PEAQ scores take negative values that range from -4 (very annoying distortions) to 0 (imperceptible distortions). The PEAQ standard uses a number of psychoacoustical evaluation techniques which are combined to give a measure of the quality difference between the original audio signal and its processed version. Even though it was introduced as an objective method to measure the quality of perceptual coders, without any reference to audio restoration, we have found it useful for our purposes as it gives scores that are well correlated with the results of time consuming listening tests. Some caution is still required in the interpretation of

PEAQ scores. While in telecommunication applications signal distortions are more or less evenly spread over time, in our current context they affect only isolated fragments of the audio material. As a result, much higher (i.e., much closer to 0) values of the PEAQ score, which is a "per sample" distortion measure, are needed to guarantee high quality of the restored audio. We have found out experimentally that, in the case of elimination of impulsive disturbances, the PEAQ threshold above which signal distortions can be regarded as imperceptible is roughly equal to -0.1.

Similarly, according to our experience, the differences between two approaches that reach or exceed the level of 0.1 in terms of the associated PEAQ scores, i.e., $|\mathrm{PEAQ}_1 - \mathrm{PEAQ}_2| \geq 0.1$, are usually audible.

In addition to PEAQ-based evaluation, two other objective measures of fit were used to quantify the obtained results – the degree of overlapping between the true and estimated pulse location functions, and the pulse energy coverage statistic. Both measures, defined below, are indirect as they quantify accurateness of the detection process.

Consider two detection alarms: the estimated one $\widehat{D}_t = [\widehat{t}_\mathrm{B}, \widehat{t}_\mathrm{E}]$ and its "ideal" version $D_t = [t_\mathrm{B}, t_\mathrm{E}]$, reflecting the true location of the detected noise pulse. Assuming that both alarms at least partially overlap ($\widehat{D}_t \cap D_t \neq \oslash$), their similarity measure can be defined as

$$s_t = \frac{\min(t_\mathrm{E}, \widehat{t}_\mathrm{E}) - \max(t_\mathrm{B}, \widehat{t}_\mathrm{B}) + 1}{\max(t_\mathrm{E}, \widehat{t}_\mathrm{E}) - \min(t_\mathrm{B}, \widehat{t}_\mathrm{B}) + 1} \quad [\%] \, .$$

The coefficient s_t takes its maximum value, equal to 100%, when two alarms coincide, i.e., $\widehat{t}_\mathrm{B} = t_\mathrm{B}$ and $\widehat{t}_\mathrm{E} = t_\mathrm{E}$. In all other cases it takes values smaller than 100% – the more so, the larger the discrepancies in the location and size of the compared binary alarm pulses. The degree of overlapping statistic s is defined as the average value of s_t computed for all detection alarms.

The pulse energy coverage statistic, proposed in [13], measures the percentage of the overall energy of noise pulses captured by the detector

$$c = \frac{\sum_{t \in T_{\widehat{\mathrm{d}}}} \delta^2(t)}{\sum_{t \in T_\mathrm{d}} \delta^2(t)} \quad [\%]$$

where $T_{\widehat{\mathrm{d}}} = \{t : \widehat{d}(t) = 1 \ \wedge \ d(t) = 1\}$, $T_\mathrm{d} = \{t : d(t) = 1\}$. Tab. I summarizes performance statistics for the compared approaches. Additionally, in the part that compares PEAQ scores, the ground truth (GT) results are shown [obtained when the signal is reconstructed under the perfect knowledge of pulse location: $\widehat{d}(t) \equiv d(t)$], as well as the results obtained for the corrupted signals prior to reconstruction (REF).

The algorithms compared in Tab. I are listed in the order of increasing PEAQ scores. The worst PEAQ scores were obtained for the algorithm A, and the best scores – for the algorithm B*. Note that template matching improves performance of the methods it is combined with. The improvement is significant in the case of the algorithm A, and much smaller but consistent and audible in the case of the algorithm B. Typical detection/interpolation results, obtained using the recommended approach B*, are shown in Fig. 4.

TABLE I

DIRECT (PEAQ) AND INDIRECT OBJECTIVE PERFORMANCE MEASURES FOR THE COMPARED DETECTION SCHEMES: THE OPEN-LOOP PREDICTION BASED APPROACH (A), THE OPEN-LOOP PREDICTION BASED APPROACH COMBINED WITH TEMPLATE MATCHING FILTERING (A*), THE DECISION-FEEDBACK PREDICTION BASED APPROACH (B), AND THE DECISION-FEEDBACK PREDICTION BASED APPROACH COMBINED WITH TEMPLATE MATCHING (B*). ADDITIONALLY, THE FIRST TABLE SHOWS RESULTS OBTAINED FOR THE CORRUPTED AUDIO FILES (REF), AND GROUND TRUTH RESULTS (GT), OBTAINED WHEN THE SIGNAL IS RECONSTRUCTED UNDER THE PERFECT KNOWLEDGE OF PULSE LOCATION. AV_{10} AND AV_{60} DENOTE AVERAGE RESULTS OBTAINED FOR THE SET OF 10 DISTINGUISHED FILES AND FOR ALL 60 FILES, RESPECTIVELY.

INTERPRETATION OF PEAQ SCORES: 0 = IMPERCEPTIBLE (SIGNAL DISTORTIONS), -1 = PERCEPTIBLE BUT NOT ANNOYING, -2 = SLIGHTLY ANNOYING, -3 = ANNOYING, -4 = VERY ANNOYING.

PEAQ

Audio file	GT	REF	A	A*	B	B*
1	-0.03	-3.68	-3.22	-1.10	-0.2	-0.16
2	-0.06	-3.69	-3.34	-1.05	-0.16	-0.14
3	-0.05	-3.70	-3.7	-2.97	-0.63	-0.47
4	-0.01	-3.59	-3.43	-1.39	-0.32	-0.28
5	-0.09	-3.77	-3.69	-2.60	-0.36	-0.20
6	-0.06	-3.84	-3.65	-3.48	-1.48	-1.05
7	-0.02	-3.82	-3.45	-2.14	-0.80	-0.49
8	-0.03	-3.55	-3.14	-2.76	-1.78	-0.92
9	-0.02	-3.13	-3.11	-0.95	-0.46	-0.37
10	-0.01	-3.01	-2.86	-1.30	-0.92	-0.68
AV_{10}	-0.04	-3.58	-3.36	-1.97	-0.71	-0.47
AV_{60}	-0.05	-3.75	-3.53	-2.69	-0.87	-0.62

Degree of overlapping [%]

Audio file	A	A*	B	B*
1	31.86	52.55	51.46	56.59
2	32.08	53.76	54.33	58.34
3	38.96	53.12	57.81	56.57
4	33.59	51.17	50.98	54.54
5	47.41	54.80	65.13	58.67
6	29.79	42.62	36.26	47.50
7	30.21	41.95	38.45	46.21
8	23.27	31.27	29.82	37.23
9	27.85	43.68	44.72	49.84
10	22.46	29.60	28.16	31.91
AV_{10}	31.75	45.45	45.71	49.74
AV_{60}	34.67	48.61	43.06	52.60

Pulse energy coverage [%]

Audio file	A	A*	B	B*
1	88.07	97.58	93.55	98.82
2	88.84	98.19	95.07	99.52
3	92.14	98.56	96.43	99.48
4	86.09	97.15	91.81	98.34
5	95.22	99.34	98.53	99.93
6	95.08	98.82	99.91	99.84
7	93.76	97.81	98.61	99.26
8	92.32	96.43	97.69	98.88
9	87.78	97.20	98.73	98.88
10	86.45	92.98	96.86	96.96
AV_{10}	90.58	97.40	96.72	98.99
AV_{60}	94.49	98.06	98.57	99.05

The indirect performance measures provide interesting insights into the compared detection schemes. Note that incorporation of template matching into the open-loop and decision-feedback prediction based approaches increases both the pulse energy coverage and the degree of overlapping statistics. This proves that the proposed scheme guarantees better placement of

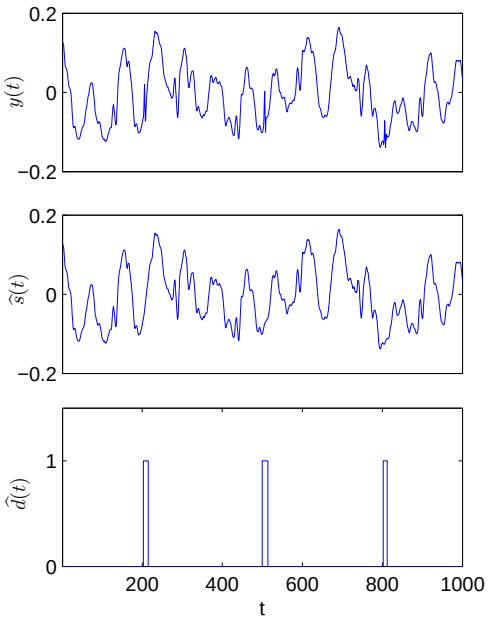

Fig. 4. Typical detection/interpolation results obtained using the recommended approach. Clicks were added to a clean signal at instants 200, 500 and 800.

detection alarms, i.e., better front/end matching of noise pulses (while the energy coverage can be easily increased by simply widening detection alarms, the simultaneous increase in the energy coverage *and* degree of overlapping can be achieved only by better fitting the true pulse location).

Tab. II summarizes click detection statistics obtained for the recommended scheme B^* (decision-feedback prediction based approach combined with template matching). The number of clicks added to each test file was equal to $N_0 = 3034$. The table shows the number of detected pulses N_d and the number of detected pulses that matched one of the click templates N_m. Additionally, it shows the number of correct (true) detections (N_dc, N_mc), the number of incorrect (false positive) detections (N_di, N_mi), and two measures of success rate (N_mc / N_0, N_mc / N_dc).

According to Tab. II, for the set of distinguished 10 files the percentage of correctly detected typical noise pulses N_mc / N_0 ranges from 81% to 93%. Since the matching procedure is initialized only when the detection alarm is raised, this statistic incorporates "losses" introduced by the classical outlier detector. The more meaningful success rate, defined as N_mc / N_dc, ranges from 89% to 97%.

Tab. III shows the average detection statistics and the average PEAQ scores obtained, using the recommended method B^*, for a different number of click templates L. Averaging was performed over 60 audio files. As expected, the PEAQ improvement rate quickly decays with the number of incorporated templates. In the case considered $L = 6$ seems to be the best performance-complexity tradeoff. Importantly, since the PEAQ scores almost monotonically decrease with growing L, there is no performance penalty for overstatement of the number of templates (actually, a small improvement can be achieved when N is increased from 6 to 14).

TABLE III
COMPARISON OF THE AVERAGE DETECTION STATISTICS AND AVERAGE PEAQ SCORES OBTAINED, USING THE RECOMMENDED METHOD B^*, FOR A DIFFERENT NUMBER OF INCORPORATED CLICK TEMPLATES L. N_m - THE NUMBER OF DETECTED PULSES THAT MATCHED ONE OF THE CLICK TEMPLATES, N_mc - THE NUMBER OF CORRECT (TRUE) DETECTIONS, N_mc - THE NUMBER OF INCORRECT (FALSE POSITIVE) DETECTIONS. THE TOTAL NUMBER OF CLICKS ADDED WAS EQUAL TO 3034.

L	N_m	N_mc	N_mi	PEAQ
0	0	0	0	-0.87
1	915	914	1	-0.78
2	1818	1672	146	-0.69
3	2060	1916	144	-0.69
4	2143	2000	143	-0.66
5	2354	2210	145	-0.67
6	2427	2283	144	-0.65
7	2455	2308	147	-0.66
8	2481	2335	146	-0.65
9	2521	2376	145	-0.65
10	2556	2398	157	-0.66
11	2720	2561	158	-0.63
12	2735	2576	158	-0.62
13	2799	2624	174	-0.61
14	2820	2634	186	-0.62

As mentioned in Section 4B, once a typical noise pulse has been localized, the corrupted fragment can be "repaired" in two ways – using signal interpolation, or by means of subtracting the appropriately scaled and bias-compensated click template from the corrupted signal (compensation technique). Tab. IV shows comparison of PEAQ scores for two variants of processing mentioned above. According to these results, interpolation yields better results than compensation – this finding was later confirmed by listening tests. The main problem with compensation is that it often produces some low-energy but audible artifacts, which degrade the overall quality of the reconstructed audio material – see Fig. 5. These artifacts occur simply because the shape of the actual noise pulse resembles but usually slightly differs from the shape of the best-matching click template.

TABLE IV
COMPARISON OF PEAQ SCORES FOR TWO VARIANTS OF PROCESSING: TEMPLATE MATCHING FOLLOWED BY PULSE SUBTRACTION (B^-) AND TEMPLATE MATCHING FOLLOWED BY SIGNAL INTERPOLATION (B^*). AV_{10} AND AV_{60} DENOTE AVERAGE RESULTS OBTAINED FOR THE PRESENTED SET OF 10 FILES AND FOR ALL 60 FILES, RESPECTIVELY.

PEAQ

Audio file	B^-	B^*
1	-3.47	-0.16
2	-3.56	-0.14
3	-3.68	-0.47
4	-3.51	-0.28
5	-3.70	-0.20
6	-3.83	-1.05
7	-3.78	-0.49
8	-3.65	-0.92
9	-2.98	-0.37
10	-3.00	-0.68
AV_{10}	-3.52	-0.47
AV_{60}	-3.70	-0.62

TABLE II

CLICK DETECTION STATISTICS OBTAINED FOR THE RECOMMENDED SCHEME B*: THE NUMBER OF DETECTED PULSES (N_{d}), AND THE NUMBER OF DETECTED PULSES THAT MATCH ONE OF THE TEMPLATES (N_{m}). THE TOTAL NUMBER OF CLICKS ADDED WAS EQUAL TO $N_0 = 3034$. ADDITIONALLY, THE TABLE SHOWS THE NUMBER OF CORRECT (TRUE) DETECTIONS (N_{dc}, N_{mc}), THE NUMBER OF INCORRECT (FALSE POSITIVE) DETECTIONS (N_{di}, N_{mi}), AND TWO SUCCESS RATES (N_{mc}/N_0, $N_{\mathrm{mc}}/N_{\mathrm{dc}}$). THE AVERAGE DETECTION STATISTICS (AV_{10}, AV_{60}) WERE ROUNDED TO THE NEAREST INTEGERS.

Audio file	N_{d}	N_{dc}	N_{di}	N_{m}	N_{mc}	N_{mi}	N_{mc}/N_0	$N_{\mathrm{mc}}/N_{\mathrm{dc}}$
1	2877	2839	38	2575	2551	24	0.84	0.90
2	3011	2965	46	2816	2779	37	0.91	0.94
3	3037	2933	104	2702	2625	77	0.86	0.89
4	2779	2754	25	2465	2451	14	0.81	0.89
5	3068	3027	41	2776	2749	27	0.90	0.91
6	3732	3030	702	3291	2736	555	0.90	0.90
7	3632	2942	690	3260	2733	527	0.90	0.93
8	4361	2887	1474	3832	2654	1178	0.87	0.92
9	3349	2923	426	3174	2831	343	0.93	0.97
10	4553	2727	1826	3972	2513	1459	0.83	0.92
AV_{10}	3440	2903	537	3086	2662	424	0.88	0.92
AV_{60}	3212	2921	290	2820	2634	186	0.87	0.90

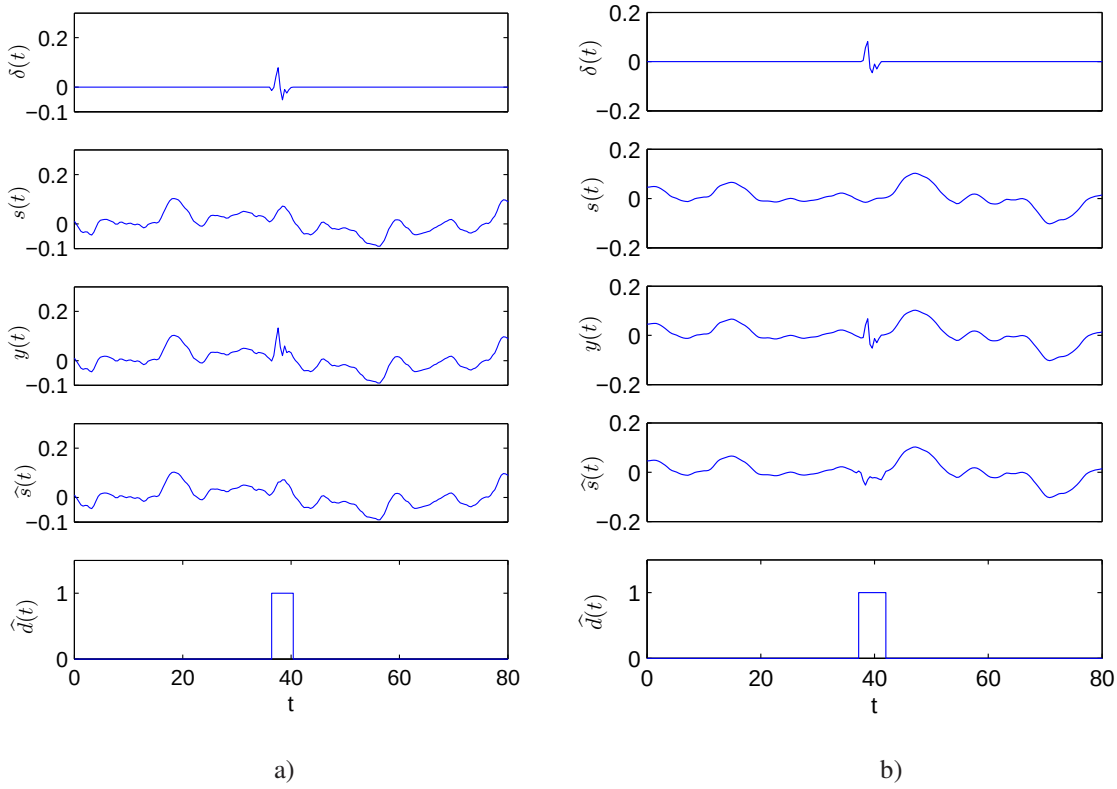

Fig. 5. Examples of successful (a) and unsuccessful (b) compensation-based elimination of noise pulses. In each group the corresponding plots show (from top to bottom): noise pulse, clean audio signal, corrupted audio signal, reconstructed audio signal, and estimated location of noise pulse.

Finally, Tab. V, which shows processing times of different algorithms, gives some idea of their relative computational complexity. Since none of the algorithms was optimized in any way, the corresponding indications should be regarded as approximate. Note that the algorithms A and B are comparable in terms of computational burden. Incorporation of template matching (algorithms A* and B*) doubles the corresponding processing times.

TABLE V

AVERAGE PROCESSING TIMES (EXPRESSED IN SECONDS) OF THE COMPARED DETECTION/RECONSTRUCTION ALGORITHMS.

Algorithm	A	A*	B	B*
Av_{60}	6.04	6.02	9.02	7.03

B. Archive Audio Files

Our last test was performed on 8 real archive gramophone recordings, sampled at 48 kHz and containing from 24 to 31

seconds of audio material. These test recordings came from two different sources: a compilation album of the opera's greatest arias, sung by Mario del Monaco (4 recordings, 58 templates), and a compilation album of the Mississippi Delta Blues music, performed by different artists (4 recordings, 24 templates). Two approaches were compared: the decision-feedback prediction based approach (B) and the decision-feedback prediction based approach combined with matched filtering (B*). Since, in the case considered, the reference (clean) audio files were not available, the evaluation had to rely on listening tests. During each test, each of 20 test persons was asked to grade the compared recordings (B, B* and the original audio file) on the quality scale between 0 and 100. The scale was divided into five equal intervals ([0,20], [21,40], etc.) with the following description: bad, poor, fair, good and excellent. Listeners were adviced to ignore the wideband surface noise (present in all recordings) when grading the quality of declicking. The order of experiments and the order of the test files within each experiment were randomized. All auditions were made using the same audio set equipped with high-quality headphones designed for critical audio monitoring. The compared recordings, or their selected fragments, could be played back as many times as needed to reach the final conclusion.

Note that the popular MUSHRA test (Multi Stimulus test with Hidden Reference and Anchor), used to evaluate the quality of audio coders [26], is not applicable in the case considered. MUSHRA should not be used when the test sounds have a near-transparent quality or when the reference sounds have low quality. In our case both problems occur.

The conclusions of listening tests are similar to those reached – for artificially corrupted audio files – using the PEAQ tool: for all archive recordings the results yielded by the proposed method B* were preferred by the majority of listeners. The click detection statistics obtained for the algorithm B* is shown in Tab. VII.

TABLE VII
CLICK DETECTION STATISTICS OBTAINED FOR THE ALGORITHM B* APPLIED TO ARCHIVE AUDIO FILES: THE NUMBER OF DETECTED PULSES (N_{d}), THE NUMBER OF DETECTED PULSES THAT MATCHED ONE OF THE CLICK TEMPLATES (N_{m}), AND PROPORTION OF PULSES REGARDED AS TYPICAL ONES ($N_{\mathrm{m}}/N_{\mathrm{d}}$).

Archive recording	N_{d}	N_{m}	$N_{\mathrm{m}}/N_{\mathrm{d}}$
blues 1	1151	831	0.72
blues 2	966	751	0.78
blues 3	2931	1948	0.66
blues 4	1002	453	0.45
aria 1	888	811	0.91
aria 2	842	721	0.86
aria 3	1737	1493	0.86
aria 4	2680	2344	0.87

VIII. CONCLUSION

The click localization approach proposed in this paper is based on the observation that the majority of noise pulses corrupting archive audio files have highly repetitive shapes that match a relatively small number of typical noise patterns, called click templates. Click templates can be created based on the set of exemplary noise pulses extracted from silent parts of archive audio recordings. To localize typical noise pulses, the pre-processed click templates can be correlated with the sequence of one-step-ahead prediction errors yielded by the autoregressive model based signal predictor. We have shown that when incorporated into the classical general purpose outlier detection scheme, such a selective disturbance localization technique can improve quality of the reconstructed audio material. The paper is illustrated with the results of objective tests, carried out on clean audio signals corrupted with real impulsive disturbances, and subjective (listening) tests performed on real archive gramophone recordings.

REFERENCES

[1] S.V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, Wiley, 1996.

[2] J.S. Godsill, and P.J.W. Rayner, *Digital Audio Restoration*, Springer-Verlag, 1998.

[3] S.V. Vaseghi and P.J.W. Rayner, "Detection and suppression of impulsive noise in speech communication systems," *IEE Proceedings*, vol. 137, no. 1, pp. 38–46, 1990.

[4] S.V. Vaseghi and R. Frayling-Cork, "Restoration of old gramophone recordings," *J. Audio Eng. Soc.*, vol. 40, no. 10, pp. 791–801, 1992.

[5] M. Niedźwiecki and K. Cisowski, "Adaptive scheme for elimination of broadband noise and impulsive disturbances from audio signals," *Proc. 14th Colloque GRETSI*, Juan-les-Pins, France, pp. 519–522, 1993.

[6] S.J. Godsill and P.J.W. Rayner, "A Bayesian approach to the restoration of degraded audio signals," *IEEE Trans. Speech, Audio Process.*, vol. 3, no. 4, pp. 267–278, 1995.

[7] S.J. Godsill and P.J.W. Rayner, "Statistical reconstruction and analysis of autoregressive signals in impulsive noise using the Gibbs sampler," *IEEE Trans. Speech, Audio Process.*, vol. 6, no. 4, pp. 352–372, 1998.

[8] M. Niedźwiecki and K. Cisowski, "Adaptive scheme for elimination of broadband noise and impulsive disturbances from AR and ARMA signals," *IEEE Trans. Signal Process.*, vol. 44, no. 3, pp. 528–537, 1996.

[9] M. Niedźwiecki, "Identification of time-varying processes in the presence of measurement noise and outliers," *Proc. 11th IFAC Symposium on System Identification*, Fukuoka, Japan, pp. 1765-1770, 1997.

[10] S. Canazza, G. De Poli, and G.A. Mian, "Restoration of audio documents by means of extended Kalman filter," *IEEE Trans. Audio, Speech Language Process.*, vol. 18, no. 6, pp. 1107-1115, 2010.

[11] F.R. Ávila and L.W.P. Biscainho, "Bayesian restoration of audio signals degraded by impulsive noise modeled as individual pulses," *IEEE Trans. Audio, Speech Language Process.*, vol. 20, no. 9, pp. 2470-2481, 2012.

[12] M. Niedźwiecki and M. Ciołek, "Elimination of clicks from archive speech signals using sparse autoregressive modeling," *Proc. 20th European Signal Processing Conference*, Bucharest, Romania, pp. 2615-2619, 2012.

[13] M. Niedźwiecki and M. Ciołek, "Elimination of impulsive disturbances from archive audio signals using bidirectional processing," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 1046-1059, 2013.

[14] M. Niedźwiecki and M. Ciołek, "Elimination of impulsive disturbances from stereo audio recordings," *Proc. 22nd European Signal Processing Conference*, Lisbon, Portugal, pp. 1 - 5, 2014.

[15] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, 1979.

[16] M. Niedźwiecki, *Identification of Time-varying Processes*, Wiley, 2001.

[17] F. Cazals and C. Karande, "A note on the problem of reporting maximal cliques," *Theor. Comput. Sci.*, vol. 407, no. 1-3, pp. 564-568, 2008.

[18] C. Bron and J. Kerbosch, "Algorithm 457: Finding all cliques of an undirected graph," *Comm. ACM*, vol. 16, no. 9, pp. 575-577, 1973.

[19] J. Konc and D. Janezic, "An improved branch and bound algorithm for the maximum clique problem," *MATCH Commun. Math. Comput. Chem.*, vol. 58, pp. 569-590.

[20] S. Haykin, *Communication Systems*, Wiley, 2001.

[21] M. Niedźwiecki, "On the localized estimators and generalized Akaike's criteria," *IEEE Trans. Automat. Contr.*, vol. 29, pp. 970-983, 1981.

[22] P. Stoica, "Multistep prediction of autoregressive signals," *Electron. Lett.*, vol. 29, no. 6, pp. 554–555, 1993.

[23] M. Niedźwiecki,"Statistical reconstruction of multivariate time series," *IEEE Trans. Signal Process.*, vol. 41, no. 1, pp. 451–457, 1993.

TABLE VI

COMPARISON OF THE RESULTS YIELDED BY 2 DECLICKING ALGORITHMS BASED ON DECISION-FEEDBACK PREDICTION SCHEME (B) AND DECISION-FEEDBACK PREDICTION SCHEME COMBINED WITH MATCHED FILTERING (B*). ALL TESTS WERE PERFORMED ON FRAGMENTS OF REAL ARCHIVE GRAMOPHONE RECORDINGS BY A GROUP OF 20 TEST PERSONS. THE NUMBERS IN EACH ROW SHOW THE AVERAGE GRADES OF THE EVALUATED RECORDINGS ON THE QUALITY SCALE BETWEEN 0 (VERY BAD) AND 100 (EXCELLENT). NUMBERS IN THE BRACKETS SHOW HOW MANY TIMES THE EVALUATED ALGORITHM EARNED AN EQUAL OR BETTER SCORE THAN ITS COMPETITORS (SINCE SOME OF THE SCORES WERE EQUAL, THE NUMBERS DO NOT SUM UP TO 20). THE BEST SCORES ARE SHOWN IN BOLDFACE.

Archive	Blues				Aria			
recording	1	2	3	4	1	2	3	4
original	29.3 [0]	32.4 [0]	26.0 [0]	26.7 [0]	41.5 [0]	40.2 [0]	37.3 [0]	37.0 [0]
B	65.6 [10]	66.0 [10]	60.1 [10]	66.2 [4]	78.2 [9]	75.2 [4]	76.5 [10]	70.0 [7]
B*	**65.7 [13]**	**68.6 [15]**	**61.3 [12]**	**73.0 [17]**	**78.3 [12]**	**79.9 [17]**	**80.2 [14]**	**77.4 [17]**

[24] ITU-R Recommendation BS.1387, "Method for Objective Measurements of Perceived Audio Quality," 1998.

[25] P. Kabal, "An Examination and Interpretation of ITU-R Recommendation BS.1387: Perceptual Evaluation of Audio Quality," Department of Electrical & Computer Engineering, McGill University, Canada, 2003.

[26] ITU-R Recommendation BS.1534-1, "Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," 2003.