Postprint of: Niedźwiecki M., Ciołek M., Elimination of Impulsive Disturbances From Archive Audio Signals Using Bidirectional Processing, IEEE Transactions on Audio Speech and Language Processing, Vol. 21, Iss. 5 (2013), pp. 1046-1059, DOI: 10.1109/TASL.2013.2244090

Elimination of impulsive disturbances from archive audio signals using bidirectional processing

Maciej Niedźwiecki, *Member, IEEE* and Marcin Ciołek

***Abstract*—In this application-oriented paper we consider the problem of elimination of impulsive disturbances, such as clicks, pops and record scratches, from archive audio recordings. The proposed approach is based on bidirectional processing - noise pulses are localized by combining the results of forward-time and backward-time signal analysis. Based on the results of specially designed empirical tests (rather than on the results of theoretical analysis), incorporating real audio files corrupted by real impulsive disturbances, we work out a set of local, case-dependent fusion rules that can be used to combine forward and backward detection alarms. This allows us to localize noise pulses more accurately and more reliably, yielding noticeable performance improvements, compared to the traditional methods, based on unidirectional processing. The proposed approach is carefully validated using both artificially corrupted audio files and real archive gramophone recordings.**

***Index Terms*—outlier detection and elimination, adaptive signal processing.**

I. Introduction

ARCHIVED audio recordings are often degraded by impulsive disturbances and wideband noise. Clicks, pops and record scratches are caused by aging and/or mishandling of the surface of gramophone records (shellac or vinyl). In the case of magnetic tape recordings, impulsive disturbances can be usually attributed to transmission or equipment artifacts (e.g. electric or magnetic pulses). Broadband noise, such as surface noise of magnetic tapes and phonograph records, is an inherent part of all analog recordings. Elimination of both types of disturbances from archive audio documents is an important element of saving our cultural heritage.

The audio restoration approaches can be divided into frequency-domain methods and time-domain methods [1], [2]. Frequency-domain methods, which are used for broadband noise suppression, include such schemes as adaptive Wiener filtering/smoothing [3]–[4], spectral subtraction [5]–[7] and, more recently, computational auditory scene analysis (CASA) [8]–[11]. In all cases mentioned above, information about time-varying signal/noise characteristics is inferred from short-time spectral analysis of the processed speech or audio. Even though numerous extensions of frequency-domain methods have been proposed over the past 30 years – such as those allowing one to continuously update noise characteristics (which, in the classical variants of Wiener filtering and spectral subtraction, are pre-estimated and fixed) [12], or to take into account perceptual features of human auditory system (signal decomposition using auditory filters, incorporation of masking mechanisms into the process of noise reduction) [13] – their fundamental limitation remains unchanged: they are not capable of removing local degradations caused by impulsive noise. Even the most advanced CASA algorithms, which use harmonicity and temporal continuity cues as a basis for segregation of the acoustic signal into streams corresponding to different sources, can be used only to remove from the corrupted speech signals the long-lasting intrusions such as white noise, "coctail party" noise, or competing speech.

Although removal of broadband noise is not the topic of this paper, it is worth noticing that all frequency-domain methods mentioned above were designed to improve quality of speech signals, where the aesthetic sound evaluation criteria are usually of secondary importance, increased signal-to-noise ratio and/or intelligibility being the main restoration objectives. When applied to archive audio signals (instrumental, vocal) such speech-oriented algorithms may produce distortions and audible artifacts that are hardly acceptable in the field of music restoration, such as over attenuation of high-frequency signal content (typical of Wiener filtering) or "musical noise" (typical of spectral subtraction). For this reason they should be used with caution – for more details see an interesting discussion in [14].

The second approach to audio restoration, which can be used for both broadband and impulsive noise removal, is based on time-domain signal analysis. The methods that fall into this category include the matching filter technique (which incorporates noise templates) [15], and techniques based on parametric (e.g., autoregressive) modeling of of audio signals, such as model-based Bayesian inference methods [16]–[17], and the extended Kalman filtering (EKF) approach [18]–[20]. A remarkable feature of model-based algorithms is their ability to simultaneously detect noise pulses, interpolate the corrupted data values, and attenuate broadband noise.

In this paper, which pursues the time-domain, model-based approach to restoration of audio signals, we focus solely on the problem of elimination of impulsive disturbances.

When attempting to eliminate real impulsive disturbances from real audio signals, one faces several challenges. First, in the classical robust estimation studies, impulsive disturbances, referred to as outliers in the statistical literature, are usually modeled as isolated pulses of unity length, with a certain probability of occurrence and a certain amplitude distribution (with variance much larger than signal variance) [21]. Even though audio signals corrupted by noise pulses generated in this way sound very much like "old recordings," detection of real impulsive disturbances based on such premises

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

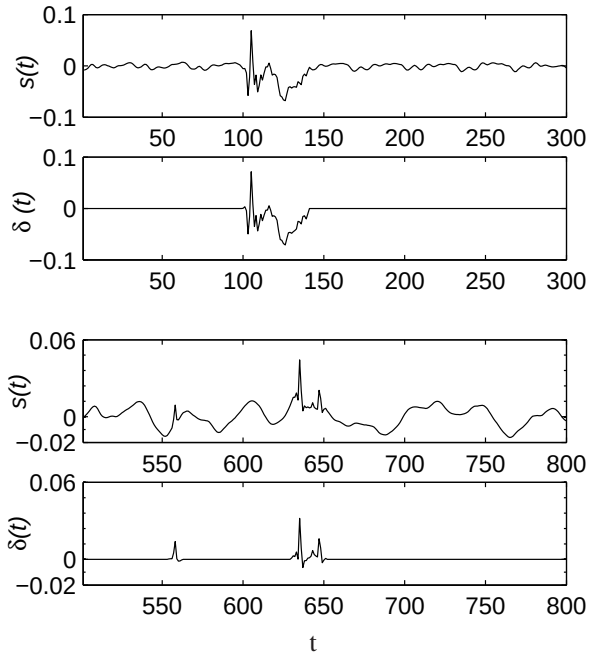

Fig. 1. Two fragments of an archive gramophone recording $s(t)$ and the corresponding noise pulses $\delta(t)$; t denotes discrete time. In both cases the sampling rate was equal to 22.05 kHz.

is usually far from satisfactory. Real disturbances may last for several to several hundred sampling intervals and have different shapes, ranging from simple unimodal to complex multimodal (or even oscillatory) patterns - see Fig. 1. In this paper most of the experiments were performed on clean audio signals corrupted by impulsive disturbances extracted from old gramophone recordings. Such a "disturbance transplantation" technique allows one to check reconstruction algorithms under realistic conditions. At the same time it gives insight into the "anatomy" of the reconstruction process, as the location and shape of each noise pulse is known exactly (allowing one to compute objective performance measures).

The second challenge is performance evaluation. No obvious objective performance measure seems to exist that would allow one to evaluate and compare reconstruction results. Development of objective methods for quality assessment of speech and audio is a long-standing problem. The well-known PESQ (perceptual evaluation of speech quality) tool can be used to assess quality of narrow-band speech degraded by coding distortions, transmission errors, packet loss, time-warping and environmental noise [22]. However, it is reported unsuitable for quantifying quality of noise reduction algorithms [23]. Some more recent results on the objective quality assessment problem are reported in [24], but impulsive disturbances are not among the five classes of audio degradation effects considered there. In case of impulsive disturbances the thing that really matters is whether or not the applied reconstruction procedure produces audible artifacts. Such artifacts may be caused by detection errors (undetected, or only partially detected, noise pulses), interpolation errors (incorrectly interpolated samples), or by both. The most trustful quality measure is that based on subjective listening tests. Objective measures, although useful, should be used with caution. For example, when treated as the performance indicator, the number of undetected noise pulses may be misleading as some of these disturbances may be not audible (audibility strongly depends on the local signal characteristics). Similarly, the number of false detections matters only if the interpolation that follows is of poor quality.

Finally, the third challenge, particularly relevant in the case of audio processing, is due to signal nonstationarity. Most of the existing detection and interpolation procedures are based on the hypothesis of stationarity of the analyzed data. For this reason they work satisfactorily when signal characteristics change slowly over time (i.e., when the signal is "locally stationary"), but may fail in the presence of their abrupt changes. "Unpredictable events", such as emergence of new sounds, may easily fool the classical outlier detection algorithms. One of our key observations is that, unlike impulsive disturbances, such forward-unpredictable events are often backward-predictable. This allows one to distinguish more precisely between natural sounds and noise pulses.

The main contribution of this paper is the demonstration that impulsive disturbances can be eliminated more efficiently if the results of forward-time signal analysis are combined with the analogous results of its backward-time analysis. Such a bidirectional processing allows one to localize noise pulses more accurately and more reliably, yielding noticeable performance improvements compared to unidirectional processing.

To the best of our knowledge, the idea of bidirectional processing was previously exploited only once, in the paper of Canazza *et al.* [20]. The method proposed there is based on combining restoration results obtained independently by means of forward-time and backward-time signal processing. The output signal is evaluated as a linear (convex) combination of its forward and backward components. The weighting coefficients are computed based on the local forward/backward prediction error statistics. Our approach is different. Based on the results of tests, performed on real audio signals, corrupted by real impulsive disturbances, we work out a set of local, case-dependent fusion rules that are further used to combine forward and backward detection alarms. Such an analysis is carried out *prior* to signal interpolation and allows one to "carve" detection alarms more carefully. This results in better disturbance coverage statistics (smaller number of shorter false alarms, smaller number of overlooked noise pulses, better front/back matching of noise pulses) and, in effect, in better restored sound quality.

II. Unidirectional processing

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. No statistical model of the disturbance (quantifying the frequency of

occurrence, length or shape of noise pulses) is assumed to be available. By $d(t)$ we will denote 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 can be decomposed into two subproblems:

1) Localization of noise pulses

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

2) Interpolation of samples regarded as outliers $Y_\delta = \{y(t) : \widehat{d}(t) = 1\}$ based on the approved samples $Y_s = \{y(t) : \widehat{d}(t) = 0\}$.

Most of the existing impulsive disturbance elimination techniques are based on autoregressive (AR) or sparse autoregressive (SAR) signal modeling, and model-based adaptive prediction: an on-line identification of the AR/SAR model of the audio signal is carried out and its results are used to predict new samples from the old ones. If the magnitude of the prediction error is too large (e.g. if it exceeds three standard deviations of its nominal value), the sample is classified as an outlier and scheduled for interpolation.

A. Approach Based on Classical AR Modeling

In this approach the sampled audio signal $s(t)$ is represented by the following AR model of order r

$$s(t) = \sum_{i=1}^{r} a_i s(t-i) + n(t) \tag{2}$$

where $a_1, \ldots, a_r$ are the so-called autoregressive coefficients and $n(t)$ denotes white driving noise. Model coefficients are continuously updated using a parameter tracking algorithm – such as exponentially weighted least squares (EWLS), least mean squares (LMS) or Kalman filter (KF) based [25], [26] – which yields $\widehat{a}_1(t), \ldots, \widehat{a}_r(t)$. Denote by $\boldsymbol{\theta} = [a_1, \ldots, a_r]^{\mathrm{T}}$ the vector of autoregressive coefficients and by $\boldsymbol{\varphi}(t) = [y(t-1), \ldots, y(t-r)]^{\mathrm{T}}$ – the regression vector, made up of r past signal values. The EWLS algorithm, known of its good tracking capabilities, can be summarized as follows

$$\begin{aligned} \widehat{\boldsymbol{\theta}}(t) &= \widehat{\boldsymbol{\theta}}(t-1) + \mathbf{k}(t)\varepsilon(t|t-1) \\ \varepsilon(t|t-1) &= y(t) - \boldsymbol{\varphi}^{\mathrm{T}}(t)\widehat{\boldsymbol{\theta}}(t-1) \\ \mathbf{k}(t) &= \frac{\boldsymbol{\Sigma}(t-1)\boldsymbol{\varphi}(t)}{\lambda + \boldsymbol{\varphi}^{\mathrm{T}}(t)\boldsymbol{\Sigma}(t-1)\boldsymbol{\varphi}(t)} \\ \boldsymbol{\Sigma}(t) &= \frac{1}{\lambda}\left[\mathbf{I} - \mathbf{k}(t)\boldsymbol{\varphi}^{\mathrm{T}}(t)\right]\boldsymbol{\Sigma}(t-1) \end{aligned} \tag{3}$$

where $\widehat{\boldsymbol{\theta}}(t) = [\widehat{a}_1(t), \ldots, \widehat{a}_r(t)]^{\mathrm{T}}$ is the vector of parameter estimates and λ, $0 < \lambda < 1$, denotes the so-called forgetting constant, determining estimation memory of the tracking algorithm. Recursive estimation of autoregressive coefficients is stopped each time a new noise pulse is detected. It is resumed when the process of reconstruction of the corrupted fragment is finished.

Detection alarm starts at the instant $t+1$: $\widehat{d}(t+1) = 1$, when the magnitude of the AR model-based one-step-ahead prediction error exceeds μ times its estimated standard deviation (typically $\mu \in [3, 5]$)[1]

$$\begin{aligned} &\widehat{d}(t+1) = 1 \quad \text{if}: \\ &|\varepsilon(t+1|t)| = |y(t+1) - \widehat{y}(t+1|t)| > \mu\widehat{\sigma}_{\varepsilon(t+1|t)} \end{aligned} \tag{4}$$

where

$$\widehat{y}(t+1|t) = \sum_{i=1}^{r} \widehat{a}_i(t) y(t+1-i), \quad \widehat{\sigma}_{\varepsilon(t+1|t)} = \widehat{\sigma}_n(t)$$

and $\widehat{\sigma}_n^2(t)$ denotes the local estimate of the driving noise variance, obtained by means of averaging the recently observed squared one-step-ahead prediction errors (after excluding outliers).

$$\widehat{\sigma}_n^2(t) = \begin{cases} \gamma\widehat{\sigma}_n^2(t-1) + (1-\gamma)\varepsilon^2(t|t-1) & \text{if} \quad \widehat{d}(t) = 0 \\ \widehat{\sigma}_n^2(t-1) & \text{if} \quad \widehat{d}(t) = 1 \end{cases} .$$

The coefficient $\gamma, 0 < \gamma < 1$, denotes another forgetting constant which determines the estimation memory of the averaging algorithm.

The detection process is continued for multi-step-ahead predictions, i.e., the absolute values of the k-step-ahead prediction errors $\varepsilon(t+k|t) = y(t+k) - \widehat{y}(t+k|t)$, $k = 2, 3, \ldots$, are checked against the corresponding thresholds $\mu\widehat{\sigma}_{\varepsilon(t+k|t)}$. The alarm ends at the instant $t + k_0 + 1$: $\widehat{d}(t + k_0 + 1) = 0$, if r consecutive prediction errors are sufficiently small, namely

$$|\varepsilon(t + k_0 + i|t)| \leq \mu\widehat{\sigma}_{\varepsilon(t+k_0+i|t)}, \quad i = 1, \ldots, r \tag{5}$$

or if the length of the detection alarm k_0 reaches the prescribed value $k_{\max}$. To avoid "accidental acceptancies" of corrupted samples localized in the middle of long-lasting artifacts (such as the one depicted in Fig. 1), it is set $\widehat{d}(t+1) = \ldots = \widehat{d}(t+k_0) = 1$ – even if for some value(s) of k, $1 < k < k_0$, the prediction error remains below the corresponding threshold. Detection alarms determined in this way always form solid blocks of "ones" preceded and succeeded by at least r "zeros". The quantity $\widehat{y}(t+k|t)$ can be obtained as a concatenation of k one-step-ahead predictions, namely

$$\widehat{y}(t+k|t) = \sum_{i=1}^{r} \widehat{a}_i(t)\widehat{y}(t+k-i|t) \tag{6}$$

where $\widehat{y}(t+j|t) = y(t+j|t)$ for $j \leq 0$. The variance of the multi-step prediction errors can be evaluated recursively using the following algorithm proposed by Stoica [27]

$$\begin{aligned} \widehat{\sigma}^2_{\varepsilon(t+k|t)} &= \widehat{\sigma}^2_{\varepsilon(t+k-1|t)} + \widehat{\sigma}_n^2(t) f_{k-1}^2(t) \\ f_{k-1}(t) &= g_{k-1}^0(t) \\ g_k^i(t) &= g_{k-1}^{i+1}(t) + \widehat{a}_{i+1}(t) f_{k-1}(t) \\ & i = 0, \ldots, r-1 \\ & k = 2, \ldots, k_{\max} \end{aligned} \tag{7}$$

[1]The value $\mu = 3$ corresponds to the well-known "three sigma" rule used to detect outliers in Gaussian signals. Since audio signals are generally non-Gaussian, very often better results are obtained for $\mu > 3$.

with initial conditions: $\widehat{\sigma}^2_{\varepsilon(t+1|t)} = \widehat{\sigma}^2_n(t)$, $f_0(t) = 1$ and $g^i_1(t) = \widehat{a}_{i+1}(t), \quad i = 0, \ldots, r-1$.

When the detection process is finished, the sequence of irrevocably distorted samples $\{s(t+1), \ldots, s(t+k_0)\}$ is interpolated using the available signal model (2). The projection-based interpolation is based on r samples preceding the missing block, and r samples succeeding the block – see Section III. In [19] all quantities needed to carry out the detection/interpolation process are evaluated by the extended Kalman filter (EKF).

B. Approach Based on Sparse AR Modeling

The procedure described above, based on AR modeling, often fails on speech signals, especially those with strong voiced episodes. The reason is not difficult to find. Since voiced speech sounds are formed by means of exciting the vocal tract (represented by the AR model) with a periodic train of glottal air pulses, the outlier detector is prone to confuse pitch excitation with noise pulses. Interestingly, the same effect can be observed for audio signals with strong vocal components, and for purely instrumental music with contribution from some wind instruments, such as trumpet, saxophone or clarinet [28].

The problem mentioned above can be overcome using sparse autoregressive modeling [29], [30]. The SAR model of an audio signal can be defined in the form

$$s(t) = \sum_{i=1}^{r} a_i s(t-i) + \sum_{j=\tau+1}^{\tau+q} a_j s(t-j) + n(t) \qquad (8)$$

where the quantities τ ($\tau \gg r$) and q are chosen in such a way that $\tau + 1 \leq T \leq \tau + q$, where T denotes the fundamental period of the signal, e.g. in the case of speech signals the period of pitch excitation (if present). Even though formally of order $p = \tau + q$, such a model is sparse as it contains only $r + q \ll p$ nonzero coefficients.

Sparse AR models capture both short-term correlations [taken care of by the first component on the right-hand side of (8)] and long-term correlations [taken care of by the second component on the right hand side of (8)] of the analyzed time series.

To better understand advantages of sparse modeling, consider a signal governed by (2) in the case where $\{n(t)\}$ is a periodic train of pulses of arbitrary shape (rather than white noise). Denote by T the period of such an external excitation. Since, under steady state conditions, the signal $s(t)$ is also periodic with period T, it obeys the following sparse model

$$s(t) = a_T s(t-T), \quad a_T = 1.$$

Note that such a model yields zero prediction errors at *all* time instants t, including the moments of periodic input activity. This explains good predictive capabilities of SAR models in the presence of mixed excitation (stochastic + periodic) – provided, of course, that the period T is carefully chosen. Fig. 2 shows errors yielded by AR-based and SAR-based adaptive predictors applied to fragments of two audio signals: voiced speech and trumpet music. In both cases the results obtained for the AR model of order 10 reveal the presence of a periodic excitation, which can be easily confused with noise pulses. When the same model is extended with just one long-term component ($q = 1$) and when T is set to the estimated pitch period, signal predictions become much more accurate and prediction errors – free of periodic excitation-related components.

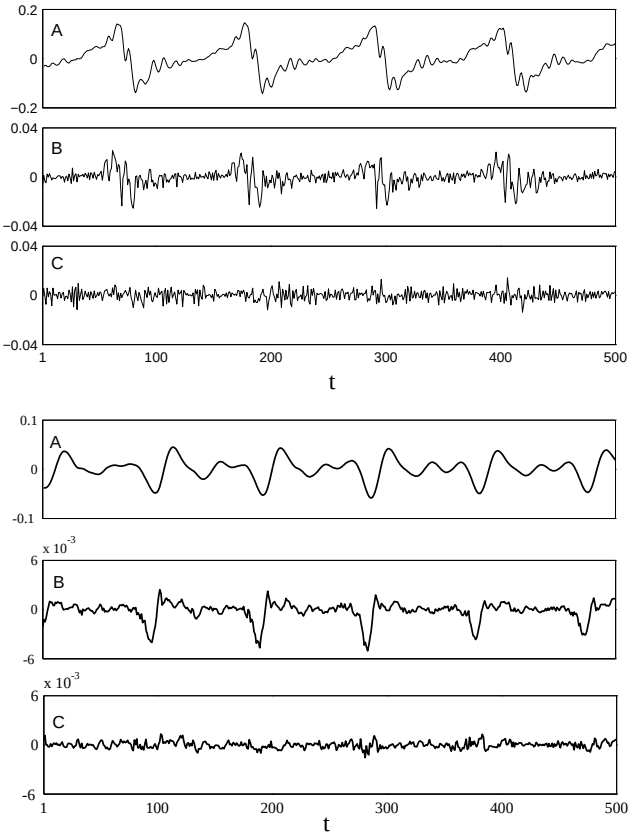

Fig. 2. Results obtained for two audio signals: voiced speech (three upper plots) and trumpet music (three lower plots). The corresponding plots show: the original audio signals (A), prediction errors yielded by the continuously updated AR model (B), and prediction errors yielded by the continuously updated SAR model (C). The order of the short-term part of both models is the same and equal to $r = 10$.

The main problem with the model (8) is that no identification algorithms seem to exist that can guarantee its stability. Additionally, since the order of the model $p = \tau + q$ is large (usually exceeding 100, even for moderate sampling rates), stability tests that could eliminate unstable models are hardly practical. Since unstable models may lead to detection and interpolation errors (such as self-oscillatory interpolation artifacts), model stability is an important practical issue.

The stability problem, mentioned above, can be easily solved if the SAR model is seeked in the following factorized form, widely used for predictive coding of speech, e.g. in CELP coders [31], [32]

$$s(t) = \sum_{i=1}^{r} \alpha_i s(t-i) + x(t) \qquad (9)$$

$$x(t) = \beta x(t-T) + n(t). \qquad (10)$$

In the speech coding context, equation (9) describes the so-

called formant filter, characterized by formant coefficients $\alpha_1, \ldots, \alpha_r$, and equation (10) describes pitch filter, characterized by the pitch coefficient β. The formant filter and the pitch filter form a cascade. Stability of the factorized model is guaranteed if both filters (formant and pitch) are stable, which can be easily achieved using appropriate estimation tools and simple stability enforcement mechanisms [31].

The factorized model (9) - (10) can be easily converted into to the generic sparse form (8) by setting $\tau = T-1, q = r+1$ and

$$a_i = \alpha_i, \quad i = 1, \ldots, r$$
$$a_T = \beta, \quad a_{T+i} = -\beta\alpha_i, \quad i = 1, \ldots, r$$

The SAR-based detection of impulsive disturbances can be carried out in the analogous way as the AR-based detection; the projection-based interpolation that follows is based on p samples preceding the missing block, and r samples succeeding the block – for more details see [30].

III. Bidirectional processing

A. Need for Bidirectional Processing

When processing has to be performed on-line, detection of noise pulses must rely on the signal past. The resulting causal detection algorithms, such as the ones described in the previous section, localize and schedule for interpolation fragments that are "unpredictable", i.e., inconsistent with the signal past. Most of impulsive disturbances fall into this category. Unfortunately, outlier detectors based on forward consistency checks have also some obvious limitations – whenever characteristics of the proposed audio signals change abruptly, e.g. at the beginning of new sounds, they generate false detection alarms. Since many of the questioned fragments are consistent with the signal future, rather than its past, the number of false alarms can be reduced if detection is based on backward consistency checks, which is possible when the analyzed signal is prerecorded and processed (in the off-line mode) backward in time. Listening tests show that the results of anticausal, reverse-time processing (both detection and interpolation) are better than those produced by causal procedures. The most likely explanation of this fact is that natural sounds have some asymetric features, namely their rise times are usually much shorter than their decay times. Hence, when adapting to time-varying signal characteristics, the backward-time signal predictor has an easier task than its forward-time counterpart.

Even though backward-time processing yields generally better results than forward-time processing, a closer inspection shows that the best performance can be achieved if the results of forward-time and backward-time detection/interpolation are combined appropriately. The corresponding fusion rules will be proposed and evaluated in Section VI.

From this point on, we will assume that two detection signals are available: $\widehat{d}_{\mathrm{f}}(t)$ and $\widehat{d}_{\mathrm{b}}(t)$, obtained by means of forward-time and backward-time processing, respectively. Similarly, by $\widehat{\boldsymbol{\theta}}_{\mathrm{f}}(t)/\widehat{\boldsymbol{\theta}}_{\mathrm{b}}(t)$, $\varepsilon_{\mathrm{f}}(t)/\varepsilon_{\mathrm{b}}(t)$ and $\widehat{\sigma}^2_{\varepsilon_{\mathrm{f}}}(t)/\widehat{\sigma}^2_{\varepsilon_{\mathrm{b}}}(t)$ we will denote parameter estimates, one-step-ahead prediction errors and innovation variance estimates, respectively, yielded by the forward-time/backward-time identification algorithms. The backward-time algorithm is identical with the forward-time one but it processes time-reversed data (to guarantee compatibility with the results of forward-time analysis, all signals produced by the backward-time algorithm are time-reversed again).

B. Mathematical Foundations

Denote by $\widetilde{s}(t) = s(N-t+1)$, $t = 1, \ldots, N$, where N is the number of available data samples, the time-reversed version of the signal $s(t)$. Note that any stationary AR signal governed by (2) has also the following reverse time representation:

$$\widetilde{s}(t) = \sum_{i=1}^{r} a_i \widetilde{s}(t-i) + \widetilde{n}(t) \tag{11}$$

where $\{\widetilde{n}(t)\}$ denotes white noise which is different from $\{n(t)\}$, but has the same variance: $\mathrm{var}[\widetilde{n}(t)] = \mathrm{var}[n(t)] = \sigma_n^2$. The proof is straightforward – since an autocorrelation function of a stationary proces is symmetric, it holds that

$$R_{\widetilde{s}}(\tau) = \mathrm{E}[\widetilde{s}(t)\widetilde{s}(t-\tau)] = \mathrm{E}[s(N-t+1)s(N-t+\tau+1)]$$
$$= R_s(-\tau) = R_s(\tau), \quad \forall t$$

which means that signals $s(t)$ and $\widetilde{s}(t)$ have the same autocorrelation function and hence they obey the same Yule-Walker equations. Note that the model (11) can be equivalently written down in the form

$$s(t) = \sum_{i=1}^{r} a_i s(t+i) + \widetilde{n}(t) \tag{12}$$

which relates the current signal value to its "future" values. The same argument applies to SAR models (8) which are nothing but high-order AR models with few nonzero coefficients. Reverse-time representation of an AR process should not be confused with its backward Markovian representation, the concept exploited in the theory of Kalman smoothing [33]. Backward representation is based on the state-space model of an AR process

$$\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t-1) + \mathbf{b}n(t)$$
$$s(t) = \mathbf{b}^{\mathrm{T}}\mathbf{x}(t) \tag{13}$$

where $\mathbf{x}(t) = [s(t), \ldots, s(t-r+1)]^{\mathrm{T}}$ denotes the state vector and

$$\mathbf{A} = \begin{bmatrix} a_1 & \ldots & a_{r-1} & a_r \\ 1 & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & \ldots & 1 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The backward Markovian equivalent of (13) takes the form

$$\mathbf{x}(t) = \mathbf{A}_*\mathbf{x}(t+1) + \mathbf{b}_* n_*(t)$$
$$s(t) = \mathbf{b}^{\mathrm{T}}\mathbf{x}(t) \tag{14}$$

where $\mathbf{A}_* = \mathbf{A}^{-1}$, $\mathbf{b}_* = -\mathbf{A}^{-1}\mathbf{b}$ and $n_*(t) = -n(t+1)$. The backward AR representation (14) differs from the reverse-time representation (12). The Markovian model (14) is less suitable for our purposes due to the fact that it loses sparsity when expressed in the input-output form similar to (8).

C. Bidirectional Interpolation

Consider a fragment of the signal scheduled for interpolation that starts at the instant t_1 and ends at the instant t_2, covering $l = t_2 - t_1 + 1$ samples. Let $\mathcal{T}_0 = [t_1, t_2]$. Given that the vector of AR coefficients $\boldsymbol{\theta}$ is known and that at least r samples preceding and r samples succeeding the interpolated block are available, the optimal, in the mean-squared sense, estimates of the missing fragment $\{s(t), t \in \mathcal{T}_0\}$ can be obtained from [34]

$$\begin{aligned}&\{\widehat{s}(t_1), \ldots, \widehat{s}(t_2)\}\\ &\quad = \arg \min_{s(t_1),\ldots,s(t_2)} \sum_{t=t_1}^{t_2+r} \left[s(t) - \sum_{i=1}^{r} a_i s(t-i) \right]^2\\ &\quad = \arg \min_{\boldsymbol{\psi}_m} \boldsymbol{\psi}^{\mathrm{T}} \mathbf{B}\mathbf{B}^{\mathrm{T}} \boldsymbol{\psi} \end{aligned} \tag{15}$$

where $\boldsymbol{\psi} = [s(t_1 - r), \ldots, s(t_2 + r)]^{\mathrm{T}}$ is the vector of all samples involved in (15), $\boldsymbol{\psi}_m = [s(t_1), \ldots, s(t_2)]^{\mathrm{T}}$ is the vector of missing samples, and $\mathbf{B}$ denotes the $(r+l) \times (2r+l)$ matrix made up of autocorrelation coefficients

$$\mathbf{B} = \begin{bmatrix} a_r & a_{r-1} & \ldots & -1 & 0 & 0 & \ldots & 0 \\ 0 & a_r & \ldots & a_1 & -1 & 0 & \ldots & 0 \\ \vdots & & & & & & & \vdots \\ 0 & 0 & \ldots & 0 & a_r & \ldots & a_1 & -1 \end{bmatrix}.$$

Denote by $\mathcal{K} = \{1, \ldots, r, r+l+1, \ldots, 2r+l\}$ the set of indices characterizing location of $2r$ known samples within the analyzed audio fragment $\{s(t_1 - r), \ldots, s(t_2 + r)\}$ of length $2r + l$. Similarly, denote by $\mathcal{U} = \{r+1, \ldots, r+l\}$ the set indicating positions of l unknown samples. By $\mathbf{B}_m = \mathbf{B}_{|\mathcal{K}>}$ we will denote the $(r+l) \times l$ matrix obtained after removing from $\mathbf{B}$ columns indicated by the set $\mathcal{K}$. Similarly, $\mathbf{B}_o = \mathbf{B}_{|\mathcal{U}>}$ will denote the $(r+l) \times 2r$ matrix obtained from $\mathbf{B}$ after removing its columns indicated by the set $\mathcal{U}$.

According to [35], the optimal estimate (15), which can be interpreted as the orthogonal projection of the vector of unknown samples $\boldsymbol{\psi}_m$ on the space spanned by known samples, is given by the following formula

$$\widehat{\boldsymbol{\psi}}_m = -\left(\mathbf{B}_m^{\mathrm{T}} \mathbf{B}_m\right)^{-1} \mathbf{B}_m^{\mathrm{T}} \mathbf{B}_o \boldsymbol{\psi}_o. \tag{16}$$

where $\widehat{\boldsymbol{\psi}}_m = [\widehat{s}(t_1), \ldots, \widehat{s}(t_2)]^{\mathrm{T}}$, and $\boldsymbol{\psi}_o = [s(t_1 - r), \ldots, s(t_1-1), s(t_2+1), \ldots, s(t_2+r)]^{\mathrm{T}}$ denotes the vector of known samples preceding and succeeding the block of missing samples $\{s(t), t \in \mathcal{T}_0\} = \{s(t_1), \ldots, s(t_2)\}$.

Since it holds that $s(t) = y(t)$ for $t \in [t_1 - r, t_1 - 1] \cup [t_2 + 1, t_2 + r]$, the interpolation formula given above can be symbolically written down in the form

$$\{\widehat{s}(t), t \in \mathcal{T}_0\} = h\left[\boldsymbol{\varphi}_{\mathrm{f}}(t_1), \boldsymbol{\varphi}_{\mathrm{b}}(t_2), \boldsymbol{\theta}\right] \tag{17}$$

where $\boldsymbol{\varphi}_{\mathrm{f}}(t) = [y(t-1), \ldots, y(t-r)]^{\mathrm{T}}$ and $\boldsymbol{\varphi}_{\mathrm{b}}(t) = [y(t+1), \ldots, y(t+r)]^{\mathrm{T}}$ denote the forward regression vector, and the backward regression vector, respectively [note that the first part of the vector $\boldsymbol{\psi}_o$ coincides with $\boldsymbol{\varphi}_{\mathrm{f}}(t_1)$, and its second part is made up of the elements of the vector $\boldsymbol{\varphi}_{\mathrm{b}}(t_2)$].

When the coefficients of the AR signal model are not known, they can be replaced with their estimates. Three approaches to adaptive interpolation were considered:

a) Forward reconstruction

Interpolation is based on the forward-time AR model, i.e., the vector $\boldsymbol{\theta}$ is replaced with its estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{f}}(t_1 - 1)$, yielded by the forward-time algorithm:

$$\{\widehat{s}_{\mathrm{f}}(t), t \in \mathcal{T}_0\} = h\left[\boldsymbol{\varphi}_{\mathrm{f}}(t_1), \boldsymbol{\varphi}_{\mathrm{b}}(t_2), \widehat{\boldsymbol{\theta}}_{\mathrm{f}}(t_1 - 1)\right]. \tag{18}$$

b) Backward reconstruction

Interpolation looks similarly as in the previous case, except that it incorporates parameter estimates yielded by the backward-time algorithm:

$$\{\widehat{s}_{\mathrm{b}}(t), t \in \mathcal{T}_0\} = h\left[\boldsymbol{\varphi}_{\mathrm{b}}(t_2), \boldsymbol{\varphi}_{\mathrm{f}}(t_1), \widehat{\boldsymbol{\theta}}_{\mathrm{b}}(t_2 + 1)\right]. \tag{19}$$

c) Mixed reconstruction

Following Canazza, De Poli and Mian [20], interpolation can be obtained as a convex combination of the results yielded by the forward-time and backward-time algorithms:

$$\widehat{s}_{\mathrm{fb}}(t) = w_{\mathrm{f}} \widehat{s}_{\mathrm{f}}(t) + w_{\mathrm{b}} \widehat{s}_{\mathrm{b}}(t), \quad t \in \mathcal{T}_0 \tag{20}$$

where

$$w_{\mathrm{f}} = \frac{\widehat{\sigma}^2_{\varepsilon_{\mathrm{b}}}(t_2 + 1)}{\widehat{\sigma}^2_{\varepsilon_{\mathrm{f}}}(t_1 - 1) + \widehat{\sigma}^2_{\varepsilon_{\mathrm{b}}}(t_2 + 1)}$$

$$w_{\mathrm{b}} = \frac{\widehat{\sigma}^2_{\varepsilon_{\mathrm{f}}}(t_1 - 1)}{\widehat{\sigma}^2_{\varepsilon_{\mathrm{f}}}(t_1 - 1) + \widehat{\sigma}^2_{\varepsilon_{\mathrm{b}}}(t_2 + 1)}$$

are the weights that depend on the local predictive performance of both algorithms. Note that $w_{\mathrm{f}} + w_{\mathrm{b}} = 1$.

Interpolation based on the SAR model can be carried out in an analogous way as described for the AR model.

D. Bidirectional Detection of Noise Pulses

1) Preliminary Considerations: In this section we will work out the rules allowing one to combine decisions $\widehat{d}_{\mathrm{f}}(t)$ and $\widehat{d}_{\mathrm{b}}(t)$ yielded by the forward and backward SAR-based outlier detectors, respectively. Each binary detection signal can be regarded as a sequence of detection alarms, further denoted by $D_{\mathrm{f}}(i)$ and $D_{\mathrm{b}}(i)$:

$$\widehat{d}_{\mathrm{f}}(t) = \begin{cases} 1 & \text{if} \quad t \in \cup_{i=1}^{n_{\mathrm{f}}} D_{\mathrm{f}}(i) \\ 0 & \text{otherwise} \end{cases}$$

$$\widehat{d}_{\mathrm{b}}(t) = \begin{cases} 1 & \text{if} \quad t \in \cup_{i=1}^{n_{\mathrm{b}}} D_{\mathrm{b}}(i) \\ 0 & \text{otherwise} \end{cases}$$

where

$$D_{\mathrm{f}}(i) = [\underline{t_{\mathrm{f}}}(i), \overline{t_{\mathrm{f}}}(i)], \quad D_{\mathrm{b}}(i) = [\underline{t_{\mathrm{b}}}(i), \overline{t_{\mathrm{b}}}(i)].$$

The quantities $\underline{t_{\mathrm{f}}}(i)$, $\underline{t_{\mathrm{b}}}(i)$ and $\overline{t_{\mathrm{f}}}(i)$, $\overline{t_{\mathrm{b}}}(i)$, such that

$$\underline{t_{\mathrm{f}}}(i) \ \leq \ \overline{t_{\mathrm{f}}}(i), \quad \underline{t_{\mathrm{b}}}(i) \leq \overline{t_{\mathrm{b}}}(i),$$

denote the beginning and the end of the ith forward/backward detection alarm, respectively.

As already remarked in Section II-A, it holds that

$$\underline{t_{\mathrm{f}}}(i+1) - \overline{t_{\mathrm{f}}}(i) > r, \quad \underline{t_{\mathrm{b}}}(i+1) - \overline{t_{\mathrm{b}}}(i) > r \ , \tag{21}$$

i.e., the consecutive detection alarms are separated by at least r no-alarm decisions. This is the minimum distance allowing one

to decompose the problem of interpolation of $n_{\rm f}/n_{\rm b}$ blocks of missing samples into $n_{\rm f}/n_{\rm b}$ local interpolation tasks analyzed in the previous subsection. Note, however, that the analogous separation between the forward and backward detection alarms is not guaranteed, which means that when analyzed jointly, such alarms may form complicated patterns. For this reason, formation of the joint detection signal $\widehat{d}_{\rm fb}(t)$, based on the results of both forward-time and backward-time analysis, is a nontrivial task.

The simplest approach to combining results of forward-time and backward-time detection is the one based on global decision rules, such as the intersection rule ($\cap$)

$$\widehat{d}_{\rm fb}(t) = \begin{cases} 1 & \text{if } \ \widehat{d}_{\rm f}(t) = 1 \text{ and } \widehat{d}_{\rm b}(t) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

or the union rule ($\cup$)

$$\widehat{d}_{\rm fb}(t) = \begin{cases} 1 & \text{if } \ \widehat{d}_{\rm f}(t) = 1 \text{ or } \widehat{d}_{\rm b}(t) = 1 \\ 0 & \text{otherwise} \end{cases} . \tag{23}$$

In the first case detection alarm is raised only when the sample is questioned by both detectors, and in the second case – when it is questioned by at least one of the detectors. Preliminary tests have shown that neither of these rules works satisfactorily in practice. The intersection rule is too conservative – it tends to overlook many small noise pulses and produces underfitted (too short) detection alarms. The union rule is too liberal – it yields many overfitted (too long) detection alarms which, after interpolation, result in audible signal distortions.

To avoid problems mentioned above, different configurations of forward and backward detection alarms, further referred to as detection patterns, were divided into several classes and subclasses. Each class was analyzed separately in order to determine the best way of combining detection alarms. The final detection decision is a result of application of a certain number of local, case-dependent decision rules, called atomic fusion rules, rather than using a single global rule applicable to all cases.

2) Preprocessing: Unlike artificially generated noise pulses, real impulsive disturbances corrupting audio signals are rarely confined to isolated samples. Moreover, most of them have "soft" edges (the more so, the higher sampling rate) which stems from the typical geometry of local damages of the recording medium (e.g. groove damages). The straightforward consequence of this fact is that detection alarms are seldom triggered at the very beginning of noise pulses. This may lead to small but audible distortions of the reconstructed audio material. Although detection delays can be reduced, or even eliminated, by lowering the detection multiplier μ, i.e., by making the outlier detector more sensitive to "unpredictable" signal changes, the improvement comes at a price: low detection thresholds may dramatically increase the number and length of detection alarms, causing the overall degradation of the results. An alternative solution, which works pretty well in practice, is based on shifting back the beginning of each detection alarm (once determined) by a small fixed number of samples further denoted by ϵ. The resulting modified detection alarms have the form

$$D^*_{\rm f}(i) = \left[\ \underline{t^*_{\rm f}}(i), \overline{t_{\rm f}}(i)\ \right], \quad D^*_{\rm b}(i) = \left[\ \underline{t_{\rm b}}(i), \overline{t^*_{\rm b}}(i)\ \right]$$

where[2]

$$\underline{t^*_{\rm f}}(i) = \underline{t_{\rm f}}(i) - \epsilon, \quad \overline{t^*_{\rm b}}(i) = \overline{t_{\rm b}}(i) + \epsilon.$$

The corresponding modified forward and backward detection signals will be denoted by $\widehat{d}^*_{\rm f}(t)$ and $\widehat{d}^*_{\rm b}(t)$, respectively. Under 22.05 kHz and 44.1 kHz sampling the best results were obtained for $\epsilon = 2$, wich means that the front edge of each detection alarm is shifted back by 2 samples.

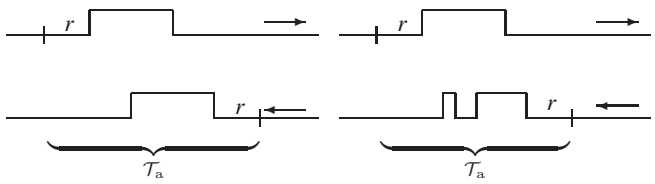

Fig. 3. Examples of elementary (left figure) and complex (right figure) detection patterns. Brackets show analysis frames $\mathcal{T}_{\rm a}$.

3) Atomic Fusion Rules: Following the interpolation guidelines we will sort out detection alarms in consecutive analysis frames $\mathcal{T}_{\rm a}(k) = [\ \underline{t_{\rm a}}(k), \overline{t_{\rm a}}(k)\ ]$, $k = 1, \ldots, n_{\rm fb}$ defined as the *minimum-length* intervals that start and end with r no-alarm decisions

$$\widehat{d}^*_{\rm f}(t) = \widehat{d}^*_{\rm b}(t) = 0 \quad \text{for} \quad t \in [\ \underline{t_{\rm a}}(k), \underline{t_{\rm a}}(k) + r - 1\ ]$$
$$\widehat{d}^*_{\rm f}(t) = \widehat{d}^*_{\rm b}(t) = 0 \quad \text{for} \quad t \in [\ \overline{t_{\rm a}}(k) - r + 1, \overline{t_{\rm a}}(k)\ ]$$

and contain at least one forward or backward detection alarm:

$$\widehat{d}^*_{\rm f}(\underline{t_{\rm a}}(k) + r) = 1 \quad \text{and/or} \quad \widehat{d}^*_{\rm b}(\underline{t_{\rm a}}(k) + r) = 1$$
$$\widehat{d}^*_{\rm f}(\overline{t_{\rm a}}(k) - r) = 1 \quad \text{and/or} \quad \widehat{d}^*_{\rm b}(\overline{t_{\rm a}}(k) - r) = 1$$

– see Fig. 3.

Situations where the analysis frame covers at most one forward detection alarm and at most one backward detection alarm will be referred to as elementary detection patterns; the remaining ones will be termed complex patterns – see Fig. 3.

Note that the adjacent analysis frames can partially overlap (they may share up to r samples at their beginning and/or end).

Detection patterns can be divided into several classes and subclasses.

A-patterns: Elementary patterns that belong to class A are made up of one forward detection alarm, say $D^*_{\rm f}(i)$, and one backward alarm, say $D^*_{\rm b}(j)$. Both alarms overlap, i.e.

$$D^*_{\rm f}(i) \cap D^*_{\rm b}(j) \neq \emptyset.$$

This class can be divided into 5 mutually exclusive subclasses – see Fig. 4

[2]Not to destroy the alarm separability condition (21), this modification is not introduced if the distance from the preceding detection alarm is smaller than $r + \epsilon$, i.e., when $\underline{t_{\rm f}}(i) - \overline{t_{\rm f}}(i-1) - \epsilon < r$ (for forward-time alarms) and $\underline{t_{\rm b}}(i+1) - \overline{t_{\rm b}}(i) - \epsilon < r$ (for backward-time alarms). In cases like this, a shorter extension is applied, namely the one that does not violate the separability condition.

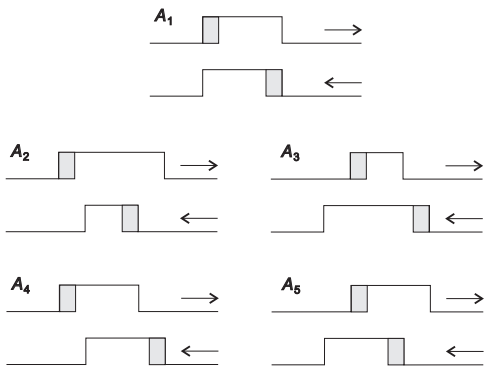

Fig. 4. Subclasses of A-class detection patterns. The plots show the results of forward detection ($\rightarrow$) and backward detection ($\leftarrow$). Shaded areas denote extensions added at the preprocessing stage.

A_1: the forward and backward alarms coincide

$$D^*_{\rm f}(i) = D^*_{\rm b}(j)$$

A_2: the backward alarm is a subset of the forward alarm

$$D^*_{\rm b}(i) \subset D^*_{\rm f}(j) \quad \text{and} \quad D^*_{\rm f}(i) \neq D^*_{\rm b}(j)$$

A_3: the forward alarm is a subset of the backward alarm

$$D^*_{\rm f}(i) \subset D^*_{\rm b}(j) \quad \text{and} \quad D^*_{\rm f}(i) \neq D^*_{\rm b}(j)$$

A_4: the forward alarm starts/ends before the backward alarm starts/ends

$$\underline{t^*_{\rm f}}(i) < \underline{t_{\rm b}}(j), \quad \overline{t_{\rm f}}(i) < \overline{t^*_{\rm b}}(j)$$

A_5: the backward alarm starts/ends before the forward alarm starts/ends

$$\underline{t_{\rm b}}(j) < \underline{t^*_{\rm f}}(i), \quad \overline{t^*_{\rm b}}(j) < \overline{t_{\rm f}}(i)$$

In each of the cases listed above, three rules of combining forward and backward detection alarms were examined – the union rule ($\cup$):

$$D_{\rm fb}(k) = D^*_{\rm f}(i) \cup D^*_{\rm b}(j)$$

the intersection rule ($\cap$):

$$D_{\rm fb}(k) = D^*_{\rm f}(i) \cap D^*_{\rm b}(j)$$

and the "front edge - front edge" rule (FF):

$$D_{\rm fb}(k) = \left[\, \underline{t^*_{\rm f}}(i), \overline{t^*_{\rm b}}(j) \,\right].$$

In the latter case the aggregated detection alarm starts at the front edge of the forward alarm and ends at the front edge of the backward alarm (which, after time reversal, becomes its back edge). The FF rule is practically motivated – it is known that the moment of triggering the detection alarm is usually determined more precisely than the moment of its termination. This is because the variance of the multi-step prediction error grows with the prediction horizon, making the corresponding outlier detector increasingly tolerant to untypical signal features.

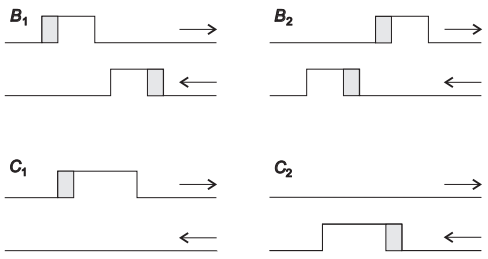

Fig. 5. Subclasses of B-class and C-class detection patterns. The plots show the results of forward detection ($\rightarrow$) and backward detection ($\leftarrow$). Shaded areas denote extensions added at the preprocessing stage.

B-patterns: Elementary detection patterns that belong to class B are made up by pairs of non overlapping detection alarms

$$D^*_{\rm f}(i) \cap D^*_{\rm b}(j) = \emptyset$$

that are separated by less than r samples (otherwise they would fall into separate analysis intervals – see class C below). This class was divided into 2 subclasses, depending on which alarm comes first – see Fig. 5

B_1: the forward alarm precedes the backward alarm

$$\overline{t_{\rm f}}(i) < \underline{t_{\rm b}}(j)$$

B_2: the backward alarm precedes the forward alarm

$$\overline{t^*_{\rm b}}(j) < \underline{t^*_{\rm f}}(i).$$

Two fusion rules were examined for this class of patterns: the "compactified union" rule ($\sqcup$):

$$D_{\rm fb}(k) = \left[\, \min\{\underline{t^*_{\rm f}}(i), \underline{t_{\rm b}}(j)\}, \max\{\overline{t_{\rm f}}(i), \overline{t^*_{\rm b}}(j)\} \,\right]$$

and the intersection (no alarm) rule

$$D_{\rm fb}(k) = \emptyset.$$

C-patterns: Elementary detection patterns that belong to class C consist of single detection alarms: either forward ones (C_1) or backward ones (C_2) – see Fig. 5. Initially only two fusion rules were considered in this case: the union rule (raise alarm) and the intersection alarm (do not raise alarm). A closer inspection of C-patterns showed that, in the majority of cases, the noise pulses (if present) occur in the close vicinity of the front edge of the corresponding detection alarms. Based on this observation, the following "front edge" rule (F) was added:

$$D_{\rm fb}(k) = \left[\, \underline{t^*_{\rm f}}(i), \overline{t^*_{\rm f}}(i) \,\right], \quad D_{\rm fb}(k) = \left[\, \underline{t^*_{\rm b}}(j), \overline{t^*_{\rm b}}(j) \,\right]$$

where

$$\overline{t^*_{\rm f}}(i) = \underline{t_{\rm f}}(i) + \epsilon, \quad \underline{t^*_{\rm b}}(j) = \overline{t_{\rm b}}(j) - \epsilon.$$

According to the F rule, the back/front edges of C-class detection alarms are placed ϵ samples away from their original front/back edges.[3] This means that the front/back edge sample is "sandwitched" between ϵ preceding samples (added at the preprocessing stage) and ϵ succeeding samples. Therefore, unless the alarm separability condition enforces limitations, the length of the resulting alarm is always equal to $2\epsilon + 1$.

[3] Should such positioning of back/front edges violate the alarm separability condition, a smaller shift is applied.

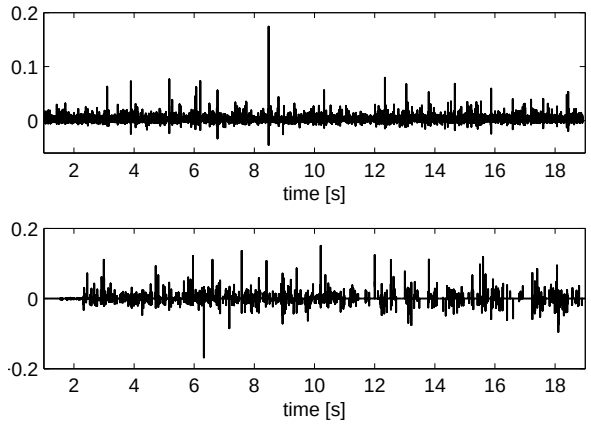

Fig. 6. Impulsive disturbances (extracted from archive gramophone recordings) used for learning (upper plot) and validation (lower plot) of detection fusion rules.

D-patterns: Class D is made up of all complex detection patterns, i.e., those which incorporate more than 2 forward/backward detection alarms that cannot be subdivided into elementary patterns – an example of such a pattern is shown in Fig. 3. For complex patterns three fusion rules, described earlier, were considered and experimentally evaluated: the "compactified union" rule, the intersection rule, and the "front edge - front edge" rule.

IV. Experimental results

A. Training Data

Our training data base was made up of 10 recordings of classical music (Bach, Mozart, Vivaldi, Smetana), chosen so as to cover different temporal and spectral features of audio signals. Each test recording was obtained under the sampling rate of $f_{\mathrm{s}} = 22.05$ kHz and contained from 23 to 29 seconds of the audio material.

Impulsive disturbances were "extracted" from the archive gramophone recording – the F. Schubert song (lied) "An die Musik" (opus 88, No. 4). This heavily corrupted, harmonically simple recording, with a strong bass line, allowed us to isolate a large variety of impulsive disturbances ranging from small pops to large scratches. The song was first declicked using a commercial audio restoration package (CEDAR). Then the difference between the original signal and its declicked version was computed to find localization and shape of noise pulses. Finally, a visual inspection of the two signals mentioned above was performed to eliminate obvious errors due to false detections or poor-quality interpolations – the corresponding fake noise pulses were removed. In this way we created a 19 seconds long recording containing 2674 isolated noise pulses covering 13428 samples – see Fig. 6. The same procedure was applied to extract another sequence of noise pulses (606 pulses covering 4099 samples), also shown in Fig. 6, that was later used for validation purposes.

As test signals for selection of fusion rules we used clean audio signals corrupted by the extracted disturbances. Prior to adding noise pulses, all audio signals were scaled so as to make their energy content in the corrupted part identical with that of the source of the disturbance signal.

B. Performance Evaluation Tools

Several, both objective and subjective, measures of fit were used for the purpose of evaluation of different detection fusion approaches.

The first three measures quantify the accurateness of the detection process. The degree of overfitting is defined in the form

$$\mathrm{o} = \frac{n_{\mathrm{o}}}{n}\ [\%]$$

where n_{o} denotes the number of elementary false positive decisions, i.e., the number of time instants for which it holds that $\widehat{d}_{\mathrm{fb}}(t) = 1$ while $d(t) = 0$, and n denotes the number of time instants for which it holds that $d(t) = 1$ (the accumulated length of all noise pulses).

Similarly, the degree of underfitting is given by

$$\mathrm{u} = \frac{n_{\mathrm{u}}}{n}\ [\%]$$

where n_{u} is the number of elementary false negative decisions, i.e., the number of time instants for which it holds that $\widehat{d}_{\mathrm{fb}}(t) = 0$ while $d(t) = 1$. Note that the first statistic includes, among others, false detection alarms, and the second statistic includes overlooked noise pulses.

Finally, the coverage statistic measures the percentage of the overall energy of noise pulses $\delta(t)$ captured by the detector

$$\mathrm{c} = \frac{\sum_{t\in\mathcal{T}_{\mathrm{c}}} \delta^2(t)}{\sum_{t\in\mathcal{T}} \delta^2(t)}\ [\%]$$

where $\mathcal{T}_{\mathrm{c}} = \{t : \widehat{d}_{\mathrm{fb}}(t) = 1 \text{ and } d(t) = 1\}$ and $\mathcal{T} = \{t : d(t) = 1\}$.

The next two measures try to assess the quality of the reconstructed audio material. The sum of squared differences between the reconstructed signal and the clean (uncorrupted) signal reflects, to some extent, the quality of the perceived sound but can be easily dominated by the results of handling (or, in fact, mishandling) a small number of large pulses. The second statistic – the number of "local victories" (further denoted by 'v') – is free of this drawback. It shows the number of cases (corresponding to subsequent analysis frames) where the particular method of processing (detection + interpolation) yields the best results, in the mean squared sense, compared to the other methods.

Even though each of the objective measures of fit, described above, yields scores that are to some extent correlated with the subjectively perceived quality of reconstruction, listening tests turned out to be unavoidable.

The blind multiple choice ordering test was used, during which the test person was asked to indicate the best recording in each of the analyzed groups of recordings. To avoid confusion, in cases where the quality of two or more recordings in a group was comparable, more than one recording could be chosen as the "best" one.

The perceptual mean opinion score (MOS) test, frequently used to evaluate the effects of removal of wideband noise [20], was deliberately skipped, as it produced inconsistent results

when used to grade the effects of elimination of high-intensity impulsive noise.

C. Program Settings

The outlier detection algorithm was based on the factorized SAR model of the audio signal. The coefficients of the formant filter (9), of order $r = 6$, were updated using the method of least squares with exponential data windowing (LSEW), different from the more frequently used method of exponentially weighted least squares. Unlike the EWLS algorithm (3), the LSEW algorithm guarantees stability of the formant filter – see [30] for more details. The forgetting constant of the LSEW algorithm was set to $\lambda_0 = 0.992$. The forgetting constant of the residual noise variance estimator was also set to $\gamma = 0.992$. The parameters T and β of the pitch filter (10) were estimated using the method described in [30]. The fundamental period T was searched in the interval $[T_{\min}, T_{\max}]$, where $T_{\min} = 20$ and $T_{\max} = 600$.

The maximum length of detection alarms was set to $k_{\max} = 125$, and the alarm extension parameter – to $\epsilon = 2$. Finally, the detection multiplier (one of the most important "tuning knobs" of the detection algorithm) was set to $\mu = 3.5$.

D. Comparison of Interpolation Schemes

Since detection of impulsive disturbances is followed by interpolation of irrevocably distorted fragments, the quality of the reconstructed audio signal depends on performance achieved at both processing stages. To check how much interpolation influences the final effect, and to choose the best interpolation formula, we examined the interpolation results obtained in the case where localization of noise pulses was known exactly, i.e., $\widehat{d}(t) \equiv d(t)$ (perfect detection).

Table I shows comparison of the averaged sums of squared SAR-model-based interpolation errors obtained for the three interpolation methods described in Section III-C: forward-time interpolation (18), backward-time interpolation (19), and mixed interpolation (20). Since, for each of the test recordings, the best results were obtained for mixed interpolation, we incorporated this method in all experiments reported below.

In addition to the quantitative analysis, summarized in Table I, listening tests were performed. In all 10 cases the effects of applying mixed interpolation were hardly audible. This means that most, if not all, audible artifacts occuring when interpolation is combined with adaptive (i.e., nonideal) detection of noise pulses are caused by detection errors such as missing detections, inaccurate detections and false detections.

E. Selection of Atomic Fusion Rules

Table II summarizes experimental results obtained for all classes of detection patterns and all 10 test recordings (clean audio signals contaminated with the sequence of noise pulses extracted from an archive recording). The performance measures ('o', 'u', 'c', MSE, 'v') shown in the table were averaged over all detection patterns of a given type found in all 10 test recordings. To enable listening tests focused on a particular class of detection patterns, test recordings were prepared in a special way. For example, to compare 3 atomic fusion rules associated with the A_2 pattern ($\cup, \cap$, FF), 3 variants of each test recording were created, confined to A_2 interventions only – all other analysis frames were filled with the undistorted audio material. Since such listening tests are very time-consuming – for each of 10 recordings 24 variants of processing, gathered in 10 groups, had to be evaluated – we relied on the opinion of three experts in the field of sound restoration (experienced sound engineers).

Note that, due to the multiple choice option, the "number one" subjective ranking scores usually do not sum up to 10 (the number of recordings). For example, the scores assigned by the first expert (E1) to the rules $\cup$, $\cap$ and FF for A_2 detection patterns are equal to 8, 0 and 9, respectively. This means that at least 7 recordings (out of 10) obtained by means of applying the union rule were regarded by him as comparable with those obtained by applying the "front edge - front edge" rule.

The experts' choice was indicated in the last column of Table II and, in a more synthetic form, in Table III.

One of the important practical questions we tried to answer was whether the selection of fusion rules based on the opinions of experts can be replaced with automatic selection based on comparison of objective quality measures, such as maximization of 'c' or 'v' scores, or minimization of the MSE scores. This would allow one to perform a more systematic search for the best fusion rules (e.g., using the machine learning techniques), and to avoid arduous listening tests. Unfortunately, none of the objective measures selects exactly the same fusion rules as experts do. The closest agreement can be observed for the 'v' measure, but also this measure fails when applied to C-patterns – it supports the "no alarm" ($\cap$) decision, which is a bad choice, as C-patterns usually correspond to short, low-energy but audible clicks that should be eliminated.

TABLE I
SUMS OF SQUARED INTERPOLATION ERRORS OBTAINED FOR THE COMPARED ADAPTIVE INTERPOLATION METHODS FOR 10 TEST RECORDINGS.

No.	Foward	Backward	Mixed
1	0.344	0.365	0.327
2	0.158	0.166	0.147
3	0.662	0.680	0.618
4	0.460	0.420	0.411
5	0.144	0.148	0.123
6	0.566	0.550	0.520
7	0.081	0.088	0.073
8	0.097	0.093	0.086
9	0.226	0.219	0.211
10	0.328	0.341	0.307
Average	0.307	0.307	0.282

TABLE III
ATOMIC FUSION RULES RECOMMENDED BY EXPERTS: "FRONT EDGE - FRONT EDGE" (FF), "COMPACTIFIED UNION" ($\sqcup$) AND "FRONT EDGE" (F).

Detection patterns	Fusion rule
A	FF
B	$\sqcup$
C	F
D	FF

TABLE II

EVALUATION OF DIFFERENT DETECTION FUSION RULES USING DIFFERENT PERFORMANCE MEASURES: DEGREE OF OVERFITTING (o), DEGREE OF UNDERFITTING (u), DISTURBANCE COVERAGE (c), SUM OF SQUARED INTERPOLATION ERRORS (MSE) AND THE NUMBER OF "LOCAL VICTORIES" (v). E1, E2 AND E3 DENOTE THE SCORES PROVIDED BY THREE EXPERTS: THE NUMBER OF TIMES WHERE THE EVALUATED RULE YIELDED THE BEST RESULTS WITHIN THE ANALYZED GROUP OF RECORDINGS (MORE THAN ONE FUSION RULE COULD BE NOMINATED IN EACH CATEGORY).

Pattern	Occurrence rate [%]	Rule	o[%]	u[%]	c[%]	MSE	v	E1	E2	E3	Experts' choice
A_1	3.39	$\cup/\cap$/FF	4.11	49.11	88.37	3.54E-02	607	10	10	10	×
A_2	8.38	$\cup$	61.98	25.59	96.22	3.74E-01	526	8	8	7	
		$\cap$	8.86	49.03	88.17	1.35E-01	641	0	0	3	
		FF	19.75	33.00	94.47	1.92E-01	885	9	8	8	×
A_3	7.51	$\cup$	61.95	27.10	95.00	1.89E-01	429	8	7	8	
		$\cap$	10.06	51.10	88.93	8.13E-02	587	0	0	1	
		FF	17.20	32.53	94.40	8.25E-02	818	10	9	10	×
A_4	40.15	$\cup$/FF	18.69	21.77	98.77	6.19E-01	4729	10	9	10	×
		$\cap$	0.86	73.18	88.36	5.27E-01	2461	0	1	0	
A_5	2.97	$\cup$	76.79	19.83	94.71	1.71E-01	161	7	8	6	
		$\cap$/FF	7.04	49.54	83.65	5.02E-02	371	8	7	10	×
B_1	3.81	$\sqcup$	19.54	14.81	98.46	8.90E-02	536	10	10	10	×
		$\cap$	0.00	100.00	0.00	1.45E-01	147	0	0	0	
B_2	0.07	$\sqcup$	153.52	21.13	49.68	1.92E-03	8	10	10	9	×
		$\cap$	0.00	100.00	0.00	1.15E-03	9	7	7	9	
C_1	13.10	$\cup$	198.24	46.24	79.66	1.33E+00	605	3	1	5	
		$\cap$	0.00	100.00	0.00	2.11E-01	1405	0	0	0	
		F	106.96	31.02	87.67	2.84E-01	387	9	9	9	×
C_2	12.53	$\cup$	119.58	46.29	76.54	6.90E-01	582	3	2	5	
		$\cap$	0.00	100.00	0.00	2.26E-01	1234	0	0	0	
		F	90.72	30.31	90.93	2.76E-01	448	10	8	8	×
D	8.08	$\sqcup$	88.00	5.27	99.68	5.85E-01	383	6	9	7	
		$\cap$	3.42	78.72	73.63	3.97E-01	292	0	0	0	
		FF	33.42	10.47	99.36	2.96E-01	913	10	7	9	×

F. Validation of Fusion Rules

Validation of the proposed approach was based on two sets of recordings – 10 obtained by adding to clean audio files noise pulses extracted from an archive recording, and 10 authentic. All recordings, along with the results of their processing, are available through the website: https://www.eti.pg.gda.pl/katedry/ksa/IEEE-TASL.html.

During validation tests, the rules recommended by the experts were evaluated *en block* (all atomic rules were applied jointly) by 20 test persons. All auditions were made using the same equipment, and in particular – using the same set of high-quality headsets. Every test person could play the compared variants of processing (or their arbitrarily selected fragments) as many times as needed to make up his/her mind. All compared recordings were displayed simultaneously on the screen in a random order, without revealing their identity.

1) Artificially Corrupted Audio Files: The artificially generated database was obtained by adding noise pulses to clean audio signals. The same set of audio recordings was used as that incorporated for selection of fusion rules, but the impulsive disturbances were extracted from another archive recording – see Fig. 6. Hence, the performance of the proposed declicking procedure was checked on a different data set than that used earlier for training purposes. All processing parameters were the same as those used during the training session.

The occurrence rates of different detection patterns observed during validation tests are shown in Table IV. Note that even though the numbers differ from those displayed in the second column of Table II, the general tendency remains the same – the most frequently observed configurations of detection alarms are those classified as A_4, C_1, C_2 and D (86% of all the cases).

Table V shows the results of comparison of 4 approaches to elimination of impulsive disturbances: the approach based on forward-time processing (traditional), the approach based on backward-time processing, the mixture approach of Canazza *et al.* [20], and the proposed bidirectional approach. The results of the ordering test show clearly superiority of the proposed method.

Some overall performance statistics of the bidirectional algorithm are shown in Table VI. Since all recordings were corrupted by the same sequence of noise pulses, the differences in scores are caused by the fact that, depending on the musical "background", the same disturbance may be easy or difficult to detect and localize.

2) Archive recordings: Even though experiments with artificially corrupted audio files have some obvious advantages – they provide access to the "ground truth", allowing one to evaluate various objective quality measures – the ultimate performance tests should be always made using real archive audio recordings. Such recordings are usually more demanding as, in addition to impulsive disturbances, they contain wideband noise (such as a surface noise of a gramophone record). Note that wideband noise was not incorporated in the measurement model (1). The data base consisted of 10 gramophone recordings (some mono and some stereo), sampled at the 44.1 kHz rate: 5 heavily corrupted (1,2,8,9,10) and 5 moderately corrupted (3,4,5,6,7). They cover a wide range

TABLE IV
OCCURRENCE RATE OF DIFFERENT DETECTION PATTERNS IN RECORDINGS USED FOR VALIDATION PURPOSES.

Pattern	A_1	A_2	A_3	A_4	A_5	B_1	B_2	C_1	C_2	D
Occurrence rate [%]	0.64	3.80	3.13	32.63	1.04	5.43	0.02	19.19	16.95	17.16

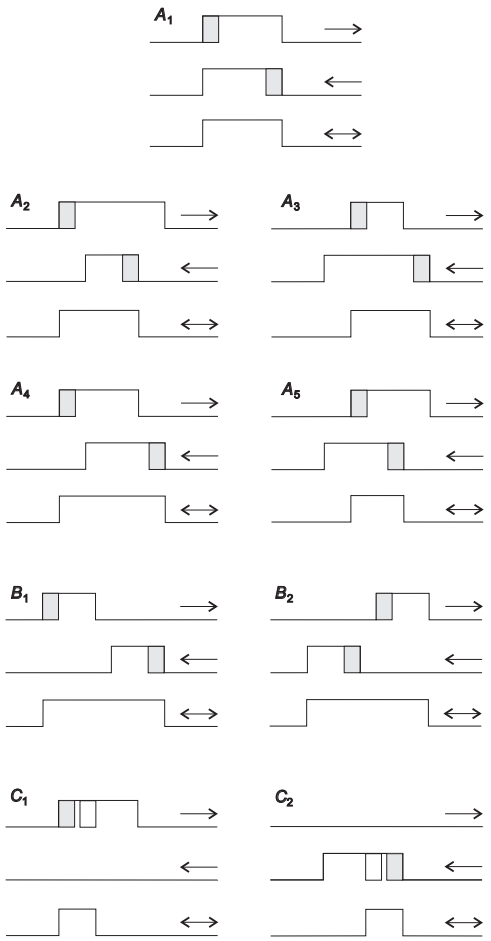

Fig. 7. Atomic fusion rules selected by experts. The plots show the results of forward detection (→), backward detection (←) and bidirectional detection (↔) for all elementary detection patterns.

TABLE V
COMPARISON OF DECLICKING ALGORITHMS BASED ON: FORWARD-TIME PROCESSING, BACKWARD-TIME PROCESSING, COMBINATION OF FORWARD-TIME AND BACKWARD-TIME PROCESSING, AND THE PROPOSED BIDIRECTIONAL PROCESSING. THE SCORES SHOW THE NUMBER OF TIMES WHERE THE EVALUATED ALGORITHM YIELDED THE BEST RESULTS WITHIN THE ANALYZED GROUP OF RECORDINGS (MORE THAN ONE RECORDING COULD BE NOMINATED).

Recording	Foward	Backward	Mixed	Proposed
1	0	0	0	20
2	0	1	0	20
3	0	0	0	20
4	0	0	0	20
5	1	0	0	20
6	0	1	0	19
7	0	1	0	19
8	0	0	0	20
9	0	0	1	19
10	0	0	0	20

TABLE VI
OVERALL PERFORMANCE STATISTICS OF THE BIDIRECTIONAL ALGORITHM OBSERVED DURING VALIDATION TESTS.

Recording	o[%]	u[%]	c[%]	MSE
1	59.19	9.10	98.32	2.20E-01
2	93.73	3.34	99.27	4.16E-01
3	39.94	17.10	94.77	3.12E-01
4	48.57	15.52	96.67	1.74E-01
5	46.62	10.78	97.51	4.90E-01
6	39.60	12.56	92.49	7.18E-01
7	92.88	1.95	99.73	4.09E-01
8	131.86	4.51	99.71	5.18E-01
9	51.65	4.81	99.38	7.71E-01
10	66.63	11.47	97.81	3.84E-01
Average	67.07	9.11	97.57	2.12E-01

of musical styles, from classical music (6,8) and opera (3,4), to pop (1,2,5,7) and blues (9,10).

One of the things we wanted to check was whether the proposed local case-dependent fusion rules yield better results than the case-independent local rules, such as the intersection rule or union rule. The results of such a comparison are shown in Table VII. Since the results obtained for the intersection rule were regarded as the worst ones by all 20 test persons for all 10 test recordings, the corresponding (zero) scores were not shown in Table VII. The advantages of using case-dependent rules are evident. Note that neither audio signals nor impulsive disturbances incorporated in this test were earlier used for training purposes.

Our last experiment aimed at comparing our results with those offered by a good commercial application. As a competitor to the proposed algorithm we have chosen CEDAR - a commercial audio restoration package known of its very good declicking capabilities.[4] The Auto Declick tool offered by CEDAR is a fully automatic procedure which does not require selection of any user-dependent parameters. All details of the outlier elimination algorithm incorporated in CEDAR are proprietary.

Since CEDAR works with 44.1 kHz audio files, some of the settings of the bidirectional algorithm were modified to accommodate the change from 22.05 kHz to 44.1 kHz sampling: the quantities $k_{\max}$ (the maximum length of detection alarms) and $T_{\min}/T_{\max}$ (the minimum/maximum value of the fundamental period), similarly as effective widths of all local analysis windows used during signal identification, were doubled. All other parameters (including r, μ and ϵ) remained unchanged. The results of comparison, shown in Table VIII, are case-dependent and hence partially inconclusive. In 2 cases (2, 8)

[4]CEDAR was originally developed at the Cambridge University for the British Library National Sound Archive (BLNSA). It turned out to be the best audio restoration system among 8 commercial products compared in [20].

TABLE VII
COMPARISON OF THE RESULTS OF DECLICKING BASED ON THE PROPOSED LOCAL CASE-DEPENDENT ALARM FUSION RULES WITH THE ANALOGOUS RESULTS OBTAINED USING THE GLOBAL CASE-INDEPENDENT UNION RULE. ALL TESTS WERE PERFORMED ON FRAGMENTS OF REAL ARCHIVE GRAMOPHONE RECORDINGS.

Recording	Advantage Union	Advantage Proposed	Deuce
1	3	12	5
2	2	18	0
3	0	18	2
4	0	19	1
5	5	8	7
6	1	12	7
7	3	13	4
8	3	6	11
9	1	16	3
10	3	15	2

the proposed algorithm yielded results rated by the majority of listeners as better than those produced by CEDAR, in one case (10) the scores were identical, and in the remaining 7 cases CEDAR was rated higher. All listeners stressed, however, that the differences were subtle – note a relatively large number of neutral decisions (almost 25%).

Even though the overall rating of CEDAR was higher, it should be noted that the proposed algorithm was run with default settings. In particular, no attempt was made to optimize its performance by selecting the detection multiplier μ, more carefully. We have noted that in many cases considerably better results can be obtained when μ is trimmed to the particular recording at hand. This leaves the room for further improvements. Automatic selection of μ will be a subject of our further research.

TABLE VIII
COMPARISON OF THE RESULTS YIELDED BY THE AUTO DECLICK CEDAR TOOL WITH THOSE PRODUCED BY THE PROPOSED BIDIRECTIONAL ALGORITHM. ALL TESTS WERE PERFORMED ON FRAGMENTS OF REAL ARCHIVE GRAMOPHONE RECORDINGS.

Recording	Advantege CEDAR	Advantege Proposed	Deuce
1	11	5	4
2	8	9	3
3	8	1	11
4	14	2	4
5	8	7	5
6	8	4	8
7	14	1	5
8	6	8	6
9	10	2	8
10	8	8	4

G. Universality Versus Specificity

Validation tests have shown that even though trained to perform well on a particular realization of impulsive disturbances, the proposed fusion rules are pretty universal, i.e. they work satisfactorily when applied to a large variety of archive recordings. It should be stressed, however, that rather than the concrete set of decision rules, the main contribution of this paper is the *procedure* (including preparation of the test data files) for their selection and validation. Applying this procedure to more specialized training data, one can easily arrive at new rules, better "matched" to the particular problem at hand, e.g. to a particular class of audio recordings and/or disturbances.

V. CONCLUSION

It was shown that impulsive disturbances can be eliminated from archive audio recordings more efficiently if the results of traditional, forward-time outlier detection are combined with the analogous results of backward-time detection. The set of local fusion rules, allowing one to combine forward/backward detection alarms, was established and validated experimentally, using both artificially corrupted audio files and real archive gramophone recordings. The new bidirectional approach offers performance improvements compared to the classical unidirectional approach and yields results comparable with those produced by the state-of-the-art commercial declicking software.

REFERENCES

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

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

[3] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, pp. 197–210, 1978.

[4] D.M.Y. Ephraim, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, pp. 1109–1121, 1984.

[5] S. Böll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 113–120, 1979.

[6] R.J. McAulay and M.L. Malpass, "Speech enhancement using a soft decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, pp. 137–145, 1980.

[7] P. Vary, "Noise suppression by spectral amplitude estimation – mechanism and theoretical limits," *Signal Process.*, vol. 8, pp. 387–400, 1985.

[8] D.F. Rosenthal, and H.G. Okuno, *Computational Auditory Scene Analysis*, Mahwah, 1998.

[9] D.L. Wang and G.J. Brown, "Separation of speech from interfering sounds based on oscillatory correlations," *IEEE Trans. Neural Networks*, vol. 10, pp. 684–697, 1999.

[10] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[11] D.L. Wang and G.J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms and applications*, Wiley, 2006.

[12] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech, Audio Process.*, vol. 9, pp. 504–512, 2001.

[13] D.E. Tsoukalas, J.N. Mourjopoulos and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech, Audio Process.*, vol. 5, pp. 497–513, 1997.

[14] S. Canazza, G. Coraddu, G. De Poli and G.A. Mian, "Objective and subjective comparison of audio restoration systems," *Proc. Int. Cultural Heritage Informatics Meeting*, ICHIM'01, pp. 273–281, 2001.

[15] S.V. Vaseghi and R. Frayling-Cork, "Restoration of old gramophone recordings," *J. Audio Eng. Soc.*, vol. 40, pp. 791–801, 1992.

[16] 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, pp. 267–278, 1995.

[17] 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, pp. 352–372, 1995.

[18] M. Niedźwiecki, and K. Cisowski, "Adaptive scheme for elimination of broadband noise and impulsive disturbances from audio signals," *Proc. Quatrozieme Colloque GRETSI*, pp. 519–522, 1993.

[19] M. Niedźwiecki, and K. Cisowski, "Adaptive scheme for elimination of broadband noise and impulsive disturbances from AR and ARMA signals," *IEEE Transactions on Signal Processing*, vol. 44, pp. 528–537, 1996.
[20] 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. 41, pp. 1107-1115, 2010.
[21] A.M. Zoubir, V. Koivunen, Y. Chakhchoukh and M. Muma, "Robust estimation in signal processing," *IEEE Signal Processing Magazine*, vol. 29, pp. 61–80, 2012.
[22] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): Objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs," International Telecommunication Union, Geneva, Switzerland, 2005.
[23] J.G Beerends, A. Hekstra, A. Rix, and M. Hollier, "Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part II – Psychoacoustic Model ," *J. Audio Eng. Soc.*, vol. 50, pp. 765–778, 2002.
[24] P.A.A Esqef, L.W.P. Biscainho, L.O. Nunes, B. Lee, A. Said, T. Kalker, and R.W. Schafer "Quality assessment of audio: increasing applicability scope of objective methods via prior identification of impairment type," *Proc. IEEE Int. Workshop Multimadia Signal Process.*, pp. 1–6, 2009.
[25] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, 1979.
[26] M. Niedźwiecki, *Identification of Time-varying Processes*, Wiley, 2001.
[27] P. Stoica, "Multistep prediction of autoregressive signals," *Electronics Letters*, vol. 29, pp. 554–555, 1993.
[28] J. Wolfe, M. Garnier and J. Smith, "Vocal tract resonances in speech, singing, and playing musical instruments," *HFSP J.*, vol. 3, pp. 6–23, 2009.
[29] S.V. Vaseghi and P.J.W. Rayner, "Detection and suppression of impulsive noise in speech communication systems," *IEE Proceedings*, vol. 137, pp. 38–46, 1990.
[30] M. Niedźwiecki and M. Ciołek, "Elimination of clicks from archive speech signals using sparse autoregressive modeling," Proc. 20th European Signal Processing Conference, Bucharest, Romania, 2012.
[31] P.R. Ramachandran, and P. Kabal, "Stability and performance analysis of pitch filters in speech coders," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, pp. 937–946, 1987.
[32] P.R. Ramachandran, and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 467–478, 1989.
[33] M. Niedźwiecki,"Locally adaptive cooperative Kalman smoothing and its application to idfentification of nonstationary stochastic systems," *IEEE Trans. Signal Process.*, vol. 60, pp. 48–59, 2012.
[34] F. Lewis, *Optimal Estimation*. Wiley, 1986.
[35] M. Niedźwiecki,"Statistical reconstruction of multivariate time series," *IEEE Trans. Signal Process.*, vol. 41, pp. 451–457, 1993.