DETECTION AND RESTORATION OF CLICK DEGRADED AUDIO BASED ON HIGH-ORDER SPARSE LINEAR PREDICTION

Bisrat Derebssa[1], Eneyew Adugna[2], Koen Eneman[3] and Toon van Waterschoot[4]
[1,2] School of Electrical and Computer Engineering, Addis Ababa Institute of Technology, Addis Ababa University, Ethiopia
[3,4]Department of Electrical Engineering, ESAT-STADIUS, KU Leuven, Belgium
Corresponding Author's Email bisrat@aait.edu.et

ABSTRACT

Clicks are short-duration defects that affect most archived audio media. Linear prediction (LP) modeling for the representation and restoration of audio signals that have been corrupted by click degradation has been extensively studied. The use of high-order sparse linear prediction for the restoration of click-degraded audio given the time location of samples affected by click degradation has been shown to lead to significant restoration improvement over conventional LP-based approaches. For the practical usage of such methods, the identification of the time location of samples affected by click degradation is critical. High-order sparse linear prediction has been shown to lead to better modeling of audio resulting in better restoration of click degraded archived audio. In this paper, the use of high-order sparse linear prediction for the detection and restoration of click degraded audio is proposed. Results in terms of click duration estimation, SNR improvement and perceptual audio quality show that the proposed approach based on high-order sparse linear prediction leads to better performance compared to state of the art LP-based approaches.

Index Terms: *Click degradation, Missing sample estimation, High-order sparse linear Prediction, linear prediction, Backward prediction*

INTRODUCTION

According to [1] click degradation refers to "localized artifacts which occur at random positions in an audio signal". These are often due to physical damages on medium and annoying to listen to. Clicks can be modeled as additive or as replacement degradation. An additive model, where the click degradation is assumed to be added to the underlying audio signal, has been shown to be acceptable for most surface defects in recording media, such as dust, dirt and small scratches [1]. A replacement model, where the degradation replaces the signal entirely for some short period of time, maybe applicable for breakages and large surface scratches which may completely destroy the underlying signal information. Generally, restoration of click-degraded audio can be seen as missing sample estimation if the underlying signal during the occurrence of the click is assumed to be lost and the time location of the click degradation is known. A method used for the restoration of click-degraded audio should only modify samples that are affected by click degradation by utilizing properties of the underrated signal before and after the degraded signal segment. To avoid unnecessary distortion of the sample values that are not degraded a

detection stages first carried out to locate samples that are affected by click degradation. Restoration is then carried out only for the samples on these detected sample locations.

The detection of click degraded samples, in short, click detection, can be cast in a statistical framework as the detection of samples that are not generated from the same random process as the underrated audio signal [1]. From this perspective, click detection becomes equivalent to outlier detection which is a widely researched problem in the field of statistical data analysis. Some of the most widely used click detection methods are based on linear filtering and autoregressive modeling.

- **Highpass Filtering**: This approach is based on the assumption that most audio signals contain little energy at high frequencies (greater than 10 kHz), while clicks have spectral content at all frequencies. Therefore, by using a high pass filter, clicks can be enhanced relative to the underlying signal [1]. Time domain power thresholding can be used after the filtering to detect those segments of the audio signal degraded by clicks. This method is one of the earliest click detection methods used in both analog and digital audio equipment [1]. It is simpleton implement with only the filter cutoff frequency and the detection threshold as parameters. The method will fail if the clicks are band-limited or if the signal has high frequency content, such as high-pitched musical instruments.
- **Autoregressive (AR) model-based click detection**: Model-based click detection methods use prior information about the underrated signal and the clicks into the detection procedure in the form of hypothesized signal models. In this approach, the underrated audio signal is assumed to be drawn from a short-term stationary process while the clicks are assumed to behave as impulsive noise. This AR modeling is very effective for human speech representation and is the basis for different audio signal representation schemes ranging from audio encoding, audio compression and audio feature extraction [2].

  For AR modeling of an underrated audio signal, the prediction error is expected to take on small values while the prediction error will be large if an impulsive noise that is not correlated with the underrated audio signal replaces the signal. Therefore, clicks can be detected by inverse filtering an audio signal using an AR model prediction error filter (PEF) and by thresholding the prediction error [1], [3], [4], [5], [6]. The limitations of this approach and researches conducted to address these are discussed below.

The PEF will spread a single impulse over future samples thereby creating interference with other impulses located in close proximity. This may make detection threshold selection problematic.

It is difficult to estimate the end time of a click due to the forward smearing effect of the PEF. Backward prediction has been used successfully to resolve this problem [1].

If the underlying audio signal is not produced by an AR process, the AR model may not well represent the signal and the prediction error may be large. In this case, false positives may be reported. This maybe the case for voiced speech and high-pitched musical notes where the AR model order may not be large enough. Autoregressive moving average (ARMA)modeling and high-order linear prediction have been proposed to better represent musical signals [1], [2],[7].

Several methods have been proposed for the restoration of click-degraded audio. The Least Squares (LS) estimation of the AR model coefficients, in the sequel referred to as linear prediction (LP) minimizes the square error (MSE) criterion assuming that the AR model excitation signal has a Gaussian distribution. It assumes that the underrated audio signal is generated by passing a white noise excitation through an all-pole filter and that the click-degraded samples are mutually independent and drawn from a Gaussian zero-mean process. The click-degraded samples can then be restored by LP-based interpolation from a priori knowledge of the LP coefficients of the undegraded audio signal, of the undegraded samples and of the time location of the click-degraded samples.

One of the limitations of the LP-based interpolator is the unavailability of the LP coefficients of the underrated signal. An iterative procedure for estimating the LP coefficients and then interpolating the missing samples was proposed by Janssen et al. [8] applying the Levinson-Durbin recursion in each iteration. Even though this approach works well for unvoiced speech [7], it is not suitable for music and voiced speech, where the AR model excitation is quasi-periodic and spiky [8]. For voiced speech and music, the minimization of the MSE, i.e., the *l2*-norm of the LP vector residual puts more emphasis on the periodic spikes of the residual [2].This problem could be resolved by including a pitch predictor in the AR model to estimate long-term correlation. However, this ignores the interaction between the long-term and short-term predictors, leading to a sub-optimal result. Joint optimization of the long-term and short-term predictors was proposed in [9]. Recently a method for the joint detection and restoration of click-degraded archived audio that uses a joint evaluation of signal prediction errors and leave-one-out signal interpolation errors was proposed [6]. It is based on thresholding the prediction error for click detection followed by multi-step ahead signal prediction. In this approach, the LP coefficients are estimated by the Levinson-Durbin recursion and restorations done by LS interpolation. The use of the conventional LP, i.e., short-term LP may limit the performance of this approach.

A better decoupling between the LP-based modeling of spectral envelope and pitch harmonics has been reported by using high-order sparse linear prediction (HOSpLP) [7],[10], [11]. In our previous work [12], [13] the use of *l1*-norm regularized and *l0*-norm regularized HOSpLP for the restoration of click-degraded audio given the time location of the click degradations has been investigated. Extensive simulations showed that the use of HOSpLP results in improved restoration performance compared to [8] in terms of signal-to-noise ratio (SNR) and perceptual evaluation of audio quality (PEAQ). In this paper the use of HOSpLP coefficients for the detection of click-degraded samples and restoration of these samples that works for both speech and music without priori on the type of audio is proposed. This will significantly decrease the need for manual annotation (speech vs. music) and segmentation (undegraded vs. degraded segments) needed for practical application.

The contribution of this paper is twofold. First, we extend the use of HOSpLP, proposed in [12], [13] for the restoration of click-degraded audio, to click detection. Second, a unified detection and restoration method based on HOSpLP coefficients is proposed. Simulation results are included to show the superior performance of the proposed HOSpLP coefficients for detection as well as restoration of click-degraded audio in comparison to state-of -the-art LP-based approach. The organization of the paper is as follows. Section informally discusses the HOSpLP coefficients considering both *l1*-norm and *l0*-norm regularization to induce sparsity. Section III

discusses the problem of click detection and proposes two click detection approaches based on HOSpLPcoefficients. Section IV unifies the detection and restoration problem. Section V discusses the data, the artificial click degradation and the performance measures used in the simulations. Section VI presents simulation results on click detection and restoration and a comparative performance evaluation to state-of-the-art approaches. Finally, Section VII concludes the paper.

HIGH-ORDER SPARSE LINEAR PREDICTION

Linear prediction (LP) is a well-understood and widely used method for the analysis, modeling, and coding of speech signals [2]. Its success is due to its alignment with the source filter model of the speech generation process [14]. It has been shown that a slowly time-varying, low-order all-pole filter can be used to model the vocal tract. The glottal excitation is modeled as either an impulse train for voiced sounds or a white noise sequence for unvoiced sounds. The purpose of all-pole modeling through LP is to obtain a short-term predictor that characterizes the spectral envelope of the vocal tract response.

The LP coefficient vector **a** can be estimated for a frame of observed samples **x** by solving the following optimization problem [7]

$$\hat{\boldsymbol{a}} = \text{argmin}_{\boldsymbol{a}} \|\boldsymbol{x} - \boldsymbol{X}\boldsymbol{a}\|_p^p + \gamma \|\boldsymbol{a}\|_k^k \quad (1)$$

Where

$\boldsymbol{X}$	=	$\begin{bmatrix} x_{N_1-1} & \dots & x_{N_1-P} \\ \vdots & \ddots & \vdots \\ x_{N_2-1} & \dots & x_{N_2-P} \end{bmatrix}$
$\boldsymbol{a}$	=	$[a_1 \ \dots \ a_P]$
$\boldsymbol{x}$	=	$[x_{N_1} \ \dots \ x_{N_2}]$
P	is	the order of the LP model
N_1 and N_2	are	the start and end indices of the frame under consideration.
γ	is	a regularization parameter

The *lp*-norm operator $\|.\|p$ is defined as

$$= \left(\sum_{n=N_1}^{N_2} |x_n|^p\right)^{\frac{1}{p}} \quad (2)$$

For conventional LP solved via the Levinson-Durbin algorithm, the *l*2-norm is used, $p = 2$, and no structure on the coefficient vector is imposed, $\gamma = 0$. Furthermore, the prediction order is usually set to a small value corresponding to twice of the number of formant frequencies to be modeled. Even though such modeling works well for unvoiced speech where the excitation can be modeled as white noise [7], it is not a good model for music and voiced speech, where the excitation is quasi-periodic and spiky [8]. For voiced speech, the excitation is appropriately modeled as periodic pulse train corresponding to the glottal output. As such the minimization of the *l*2-norm of the residual puts more emphasis on the periodic spikes of the residual [2].As a result, it tradeoff short-term prediction, i.e., spectral envelope, estimation accuracy against the long-term prediction, i.e., pitch estimation accuracy [2]. As the aim of conventional LP is to model the vocal tract and not the glottal excitation, this leads to a suboptimal solution.

For musical sounds or tonal audio for which the signal contains a finite number of dominant frequency components, the LP model is much less popular than in speech analysis as the generation of musical sounds is dependent on the instruments used [2]. This makes it hard to use a generic audio signal generation model [2]. In addition, each polyphonic audio signal should be modeled using multiple source-filter models [2], [14]. In the absence of noise, by using a

model order which is twice the number of tonal components, LP can be used to estimate the spectral peaks. In practice, noise is always present that may be due to imperfections in the tonal behavior, or lack of tonal behavior, finite precision arithmetic, finite-length data windowing or background or sensor noise. Therefore, such LP signal estimates are very often poor. In [2] extensive simulations were conducted to assess the performance of conventional and alternative LP models for tonal audio analysis in the presence of noise. It was reported that high-order all-pole models are better suited to the audio LP problem albeit being impractically complex in many applications due to the excessive number of LP coefficients.

One of the most recent approaches to LP is sparse linear prediction (SpLP), which takes into consideration the sparsity of the residual and the LP coefficients. When applied to high-order all-pole models, SpLP is referred to as high-order sparse linear prediction (HOSpLP). A better decoupling between the spectral envelope and pitch estimation has been reported by using HOSpLP [7], [10], [11], [12], [13]. While the high-order all-pole method used in [2] minimize the *l2*-norm of the residual to obtain the LP coefficients, the HOSpLP methods impose sparsity of the residual and the coefficient vector in the optimization problem.

A. l1-norm regularized HOSpLP

By imposing sparsity of the residual in the LP problem formulation the emphasis on outliers in the solution to (1) can be decreased [7]. That is by considering a sparsity-inducing norm of the residual vector instead of the *l2*-norm. The convex relaxation of the '*l0*-norm' cardinality problem has been proposed to lead to a sparser residual [7]

In addition, by using a high-order all-pole model and imposing sparsely of the coefficient vector in (1), by setting γ=0 and k = 1, joint estimation of the short-term predictor and the long-term predictor can be achieved [7] as in (3).

$$\widehat{\boldsymbol{a}} = \text{argmin}_{\boldsymbol{a}} \|\boldsymbol{x} - \boldsymbol{X}\boldsymbol{a}\|_1 \qquad (3)$$

This results from the observation that a cascade of a long-term and short-term predictor results in a filter that has few non-zero coefficients [14]. Therefore, the sparsity of the coefficient vector can be used to regularize the solution. The purpose of the HOSpLP coefficients obtained by solving (4) is to model the whole spectrum, i.e., the pitch related harmonics and the spectral envelope.

$$= \text{argmin}_{\boldsymbol{a}} \|\boldsymbol{x} - \boldsymbol{X}\boldsymbol{a}\|_1^1 + \gamma\|\boldsymbol{a}\|_1^1 \qquad (4)$$

The problem in (4) is convex but not differentiable. However, it can be solved via splitting methods such as the alternating direction method of multipliers (ADMM) by reformulating the problem as a basis pursuit problem [15]. The regularization parameter, γ, determines the trade-off between the sparsity of the predictor coefficients and the sparsity of the residual. The modified L-curve [16] has been used to obtain an optimum value for the regularization parameter in[11]. In [11] an adaptive algorithm was proposed for estimating the regularization parameter based on the observation that the optimal γ is related to the pitch gain.

To solve the problem of obtaining the short-term and long-term predictors from a HOSpLP coefficient vector, **a**, the first few, *Nf*, coefficients of the HOSpLP coefficient vector been used to represent the short-term predictor in [7]. After this, a polynomial factorization can be carried out to obtain the long-term predictor after selection of the number of taps in the long-term predictor, typically *Np*= 1 or *Np*= 3.

The use of the *l1*-norm in HOSpLP has been shown to outperform conventional LP in the estimation of spectral envelope, sparse LP coefficients and sparse residual [7]. With regards to stability of the obtained short-term filters, it has been shown in [7] that the percentage of unstable filters is very low (around 2%) with "mild" instability.

B. l0-norm regularized HOSpLP

The prior knowledge of the structure of the coefficient vector resulting from cascading a long-term and short-term predictors can also be incorporated in the HOSpLP optimization problem as (5) [13],

$$\widehat{\boldsymbol{a}} = \text{argmin}_{\boldsymbol{a}} \|\boldsymbol{x} - \boldsymbol{X}\boldsymbol{a}\|_2^2 \text{s.t.} \|\boldsymbol{a}\|_0 \leq \Psi \quad (5)$$

Where Ψ is the sum of the filter order of the long-term and short-term predictors.

This formulation does not impose a restrictive structure on the coefficient vector except that the coefficient vector has a fixed maximum number of non-zero coefficients. As such, it can give emphasis to the formant filter coefficients if the signaling the frame is composed of unvoiced speech and to the pitch or tonal components if the frame is composed of voiced speech or music. In [13] it was shown that the coefficients obtained by solving (5) correspond to the short-term and long-term predictor. As the location of the non-zero coefficients is neither incorporated into (5) nor dependent on a pitch predictor, prior information regarding the type of signal is not needed. In addition, the structure of the coefficient vector can change from frame to frame if the signal is composed of both speech and music.

It should be mentioned that the use of the *l*1-norm of the residual in (5) is expected to lead to better results as compared to *l2*-norm. However, *l1*-norm of the residual in (5) is difficult to solve efficiently. Problem (5) is non-convex [17] which means that it a may have several local minima and its convex relaxation, the least absolute shrinkage and selection operation (LASSO), obtained with *p*=2 and *k*=1 in (1), is typically solved instead [17],[18]. Nevertheless, proximal gradient methods can efficiently solve (5) if a good initialization is given, e.g., the solution of LASSO [18]. In recent work, Antonello et. al [18] developed the Structured Optimization package for the Julia programming language that can solve (5) in a reasonable time. This package is used in this work to obtain *l0*-norm regularized HOSpLP coefficient vector.

CLICK DETECTION

In practice the time location of the click degradation is not known a priori, therefore click detection methods are needed. One of the most widely used click detection approaches consists in energy thresholding of the LP residual [1]. This approach is based on the assumption that the click degradations not generated from the same AR random process as the undegraded audio signal. Therefore, in the presence of click degradation the energy of the LP residual in that time frame will be much larger than the energy of the residual when click degradation is not present. It has been shown in other applications that significant improvement in noise detectability can be achieved by transforming the noisy speech to the excitation domain of the speech signal [19].

In LP-based click detection methods, the energy of the LP residual at each sample is compared with an average residual energy of the frame as follows,

For $n = 1$ to N

1) Calculate LP residual: $\epsilon_n = x_n - \sum_{j=1}^{P} \hat{a}_j x_{n-j}$
2) if $|\epsilon_n| \geq K\sigma_e$, then $\mathbf{i}_n = 1$, else $\mathbf{i}_n = 0$

Where

σ_e^2	is the variance of the LP residual,
K	is a detection threshold,
N	is the frame length,
i	is a vector representing the presence or absence of click degradation at each sample value, $\mathbf{i}_n$= 1 represents presence and $\mathbf{i}_n$=0 represents absence of click degradation at the n^{th} sample.

In this approach the start of click degradation is accurately estimated [1]. However, the end of a click degradation cannot be accurately estimated due to the forward smearing effect over P + 1 samples, where P is the order of the AR model .To detect the end of a click, a moving average filter can be applied to see when the residual variance in a local window has energy lower than the threshold (or some scaled version of the threshold). However, this requires a precise tuning o the threshold and local window size to detect the end of a click degradation.

When impulses are present in close vicinity to each other their impulse responses resulting from filtering with the PEF may add constructively to give a false detection or cancel one another out [1].In general, threshold selection is difficult when impulses of differing amplitudes are present. The use of the backward prediction error for the detection of clicks has been proposed in [1], [20].

This method takes advantage of the accurate LP-based start click identification. In this approach, once a click is detected and its start location identified, the backward prediction error is then used to detect the end of the click. By assuming that the time-reversed signal can be reasonably modeled as an AR process, the energy of the LP residual of the time-reversed signal near the identified click start location is evaluated to detect the end of the click degradation. The backward prediction error is defined as

$$\epsilon_n^b = x_n - \sum_{i=1}^{P} b_i x_{n+i} \quad (6)$$

When these coefficients are obtained by using the conventional LP, the backward prediction error is composed of spikes due to the quasi-periodic excitation for voiced speech and music. This makes it difficult to select a threshold for the detection of the end of clicks without incorrectly selecting spikes due to the quasi-periodic excitation.

In this paper, the HOSpLP coefficients are used in click detection, see Algorithm 1, by exploiting the fact that the short-term and the long-term predictors can be jointly estimated using HOSpLP leading to a residual that has less spiky nature due to the quasi-periodic excitation [7].

As such, the backward prediction error in a local window near the identified click start can be used to estimate the end of the click without significantly being affected by a spiky residual. To avoid mislabeling un degraded samples between two click degradations that are close together, the backward prediction error is checked to be greater than the threshold in local window around the detected click start.

Algorithm 1 Backward prediction using HOSpLP model

procedure BACKWARD_PRED_HOSPLP
  Input: $\mathbf{x}, P, \gamma, R, K, N$
  Output: $\mathbf{c}$
  $\hat{\mathbf{a}}$ =COEFFICIENT($\mathbf{x}, P, R, \gamma, \zeta$);
  $\epsilon^{\mathbf{x}}$=RESIDUE(x,$\hat{\mathbf{a}}$);
  σ_e =STANDARD_DEVIATION($\epsilon^{\mathbf{x}}$);
  for $n = 1$ to N **do**
    if ($|\epsilon_n| \leq K\sigma_e$), break;
    else $\mathbf{c}_n = 1$;
    $\hat{\mathbf{b}}$ =COEFFICIENTS($\mathbf{x}^B, P, R, \gamma, \zeta$);
    ϵ^b=RESIDUE($\mathbf{x}^B$,$\hat{\mathbf{b}}$);
    for $l = n$ to $n + k_{max}$ **do**
      if ($|\epsilon|_l^b \geq K\sigma_e$) $\mathbf{c}_l = 1$; continue;
      for $j = l \leq l + W$ **do**
        if ($|\epsilon|_j^b \geq K\sigma_e$) $\mathbf{c}_{l:j} = 1$; $l = j$; break;
      end
      if ($j \geq l + W$) $n = l$; continue;
    end
  end
  Return

Where,

x	is the click degraded signal vector;
$\boldsymbol{x_B}$	is the time-reversed click degraded signal vector;
I	is the estimated location of click;
K	is the threshold value;
N	is the number of samples in each frame;
R	is the maximum number of ADMM iterations for $l1$-norm HOSpLP;
W	is a local window size;
γ	is the regularization parameter for $l1$-norm HOSpLP;
ζ	is the residual stopping criterion for ADMM algorithm in $l1$-norm HOSpLP.

The function COEFFICIENTS(**x,** *P, R,* γ) obtains the LP coefficients as follows. The function RESIDUE(**x**,**â**) obtains the residual error by inverse filtering the signal with a AR filter with coefficients **â**.

• ***l*1-norm regularized HOSpLP**: the ADMM algorithm for solving the *l1*-norm regularized problem [15] is used to obtain the HOSpLP coefficients [12].

• ***l*0-norm regularized HOSpLP**: the *l0*-norm regularized problem (5) is solved via the Structured Optimization Julia package to obtain the HOSpLP coefficients [18].

IV. UNIFIED APPROACH FOR DETECTION ANDRESTORATION OF CLICK-DEGRADEDAUDIO

In this section, a unified approach is proposed that detects the location of click degraded-samples and restores these samples by using the HOSpLP coefficients without a prior knowledge on the type of audio and the time location and duration of the click degradation.

A. Detection and restoration by using backward prediction and Janssen iteration

Initially, the backward prediction based on *l0*-norm regularized HOSpLP coefficients is used to detect samples degraded by click degradation. Then these samples are restored by an iterative algorithm, see Algorithm 2, similar to the Janssen iteration [8], [17] but using *l0*-norm regularized HOSpLP for the restoration as this is shown to provide the best signal restoration performance [13].

Algorithm 2 Detection and Restoration using backward prediction and Janssen restoration based on HOSpLP model

procedure RESTORATION_HOSPLP
  Input: $\mathbf{x}, P, \gamma, R, K, N, L, Q, \zeta$
  Output: $\mathbf{y}$
  $\hat{\mathbf{c}}$=BACKWARD_PRED_HOSpLP($\mathbf{x}, P, \gamma, R, Q, N$);
  $h = 1; g = 1$
  for $i = 1 : N$ **do**
    if ($\hat{\mathbf{c}}_i == 1$) $\mathbf{v}_h = i; h = h + 1;$
    else $\mathbf{u}_g = i; g = g + 1;$
  end
  $\boldsymbol{\Theta} = |\mathbf{v}\mathbf{1}_{1\times N} - \mathbf{1}_{M\times 1}[1, 2, \cdots N]|;$
  $\hat{\mathbf{x}}_\mathbf{u} = \mathbf{x}_\mathbf{u}$; $\hat{\mathbf{x}}_\mathbf{v} = 0$; $\boldsymbol{\Phi} = \mathbf{0}_{M\times N}$; $l = 0$;
  for $l \leq L$ **do**
    $\hat{\mathbf{a}}$ =COEFFICIENT($\hat{\mathbf{x}}$,P,γ,R,ζ);
    $\mathbf{b} = \begin{bmatrix} 1 & -\hat{\mathbf{a}}^T \end{bmatrix} \mathbf{A}$;
    $\boldsymbol{\Phi}_{i,j} = \mathbf{b}_{\boldsymbol{\Theta}_{i,j}+1}$; $\forall i, j : \boldsymbol{\Theta}_{i,j} \leq P$
    $\hat{\mathbf{x}} = -\boldsymbol{\Phi}^{-1}_{(1:M,\mathbf{c})} \boldsymbol{\Phi}_{(1:M,\mathbf{v})} \mathbf{s}_\mathbf{v}$;
    $l = l + 1$;
  end
  Return

Where

$$\mathbf{A} = \begin{bmatrix} 1 & -\hat{a}_1 & -\hat{a}_2 & \cdots & -\hat{a}_P \\ -\hat{a}_1 & -\hat{a}_2 & \cdots & -\hat{a}_P & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\hat{a}_P & 0 & 0 & \cdots & 0 \end{bmatrix};$$

L is the number of Janssen iterations.

B. Benchmark method incorporating HOSpLP coefficients

As a comparison, a recently proposed method by Ciołeket.al. [6] for the joint detection and restoration of click-degraded archived audio that uses a joint evaluation of signal prediction errors and leave-one-out signal interpolation errors is used. It is based on thresholding the forward prediction error for click detection followed by multi-step-ahead prediction for restoration.

A click start is detected when the absolute prediction error is larger than and a click end is detected if the residual at *k*0iteration is smaller than a threshold and consecutive residuals are smaller than same threshold. In this approach, the LP coefficients are estimated by the Levinson-Durbin recursion and restoration is done by LS interpolation [21]. The use of the conventional LP may limit the performance of this approach due to the limited capability to model pitch and tonal components. We propose to use HOSpLP coefficients in this method by using the *l*1-norm regularized HOSpLP coefficients instead of using the conventional LP coefficients solved via the Levinson-Durbin recursion. Algorithm 3 shows a simplified algorithm to illustrate where the HOSpLP coefficients to be used. The code for the original implementation is available in [22]. The reason the *l*1-norm regularized HOSpLP is used instead of *l*0-norm regularized HOSpLP is due to the fact that the *l0*-norm regularized HOSpLP coefficients are solved by using the Structured Optimization package of Julia programming language, whereas the original code for Ciołek's method is in MATLAB. It should be mentioned that the use of HOSpLP coefficients in this method leads to significant computational cost as it yields to a solution to an iterative problem nested in another iterative problem, i.e., re-estimating the HOSpLP coefficients. The restoration is done by using the LS interpolation method as used in their original work.

The function COEFFICIENTS($\hat{\boldsymbol{x}}$, *P, M, γ, ζ*) obtains the LP coefficients using Levinson-Durbin in the original method[6] and using ADMM in our proposed *l*1-norm regularized HOSpLP variation of [6].

Algorithm 3 Iterative detection and restoration via leave-one-out interpolation [6] by incorporating HOSpLP model

procedure CIOLEK_HOSPLP
  Input: $\mathbf{x}, P, M, \gamma, K, N$
  Output: $\mathbf{y}, I$
  $\hat{\mathbf{x}} = \mathbf{x}$;
  $\hat{\mathbf{a}}$ =COEFFICIENT$(\mathbf{x}, P, R, \gamma, \zeta)$;
  $\epsilon^{\mathbf{x}}$=RESIDUE(x,â);
  σ_ϵ=STANDARD_DEVIATION$(\epsilon^{\mathbf{x}})$;
  for $n = 1$ to N **do**
    if $(|\epsilon_n| \leq K\sigma_\epsilon)$ continue;
    $\mathbf{i}_n = 1$;
    x̂=Leave_One_Out_Interpolation(x̂,n,â);
    â =COEFFICIENTS(x̂, P, M, γ, ζ);
    $\epsilon_n = \hat{x}_n - \sum_{j=1}^{P} \hat{a}_j \hat{x}_{n-j}$;
    if $\exists l \in \{0, ..., k_0\}: |\epsilon_{n-l}| \geq K\sigma_\epsilon$ continue;
  end
  Return

SIMULATION SETUP

A. Data used

To fairly assess the detection and restoration performance of the proposed methods, the experiments were conducted using speech (male and female) and music (singing voice and instrumental) from the Archimedes dataset [23]. In order to have comparable degradations among all signals, each signal is normalized so that the maximum amplitude is 1.Five male and five female speech from different speakers are taken. For each speech simulation is done on 100 frames each 32.5 ms. The result is then averaged among these. Similarly, for music 2 male singing voices, 2 female singing voice, 2instrumental audio and 4 audio consisting of singing voice and instrument are used.

B. Click Degradation Model

Usually, the start, duration and amplitude of each click degradation is modeled probabilistically. Different probability distributions for the time between impulses and for their amplitudes can be used [1], [24]. In this

work, the time location of click degradation was assumed to be uniformly distributed as the causes of click degradation are not correlated with the audio signal. As such, click degradations can occur at any location irrespective of previous click degradation location and the samples during the occurrence of click were replaced with zero-mean Gaussian noise to obtain a click degraded signal. The standard deviation of the click degradation is set as twice the standard deviation of the audio signal. The impact of the click degradation variance on the detection and restoration performance of the various methods is investigated in Section VI.

C. Performance Measures

To evaluate click detection accuracy, the normalized MSE in click duration estimation for each data set and for a given click duration as shown in (7) is used.

$$NMSE = \sum_{h=1}^{H} \frac{|T_{click}(h) - \hat{T}_{click}(h)|^2}{|T_{click}(h)|^2} \qquad (7)$$

where

Tclick	is the actual click duration;
$\hat{T}click$	is the estimated click duration;
H	is the total number of audio files for each dataset.

To evaluate the restoration performance, the Signal-to-noise ratio (SNR) and perceptual evaluation of audio quality (PEAQ)are used. The SNR is evaluated over the entire duration of the signal to also take into account unnecessary interpolation that may result from incorrect click detection.

$$SNR = \sum_{h=1}^{H} 10\log_{10} \frac{|\boldsymbol{x}(h)|^2}{|\boldsymbol{x}(h) - \hat{\boldsymbol{x}}(h)|^2} \qquad (8)$$

Where $\boldsymbol{x}$is a vector of the undegraded audio and $\hat{\boldsymbol{x}}$is a vector of the restored audio.

PEAQ is used to assess the subjective quality of the restored audio signal [25]. It predicts the basic audio quality of a signal with respect to a reference signal by modeling the psychoacoustic properties of the human auditory system. It has a range of 0 to -4: 0 representing imperceptible distortion while -4 means very annoying distortion. PEAQ has been used for the assessment of click-degraded audio restoration in [5]and [6]. The PEAQ implementation in [26] is used in this research.

RESULTS AND DISCUSSION

The backward prediction and iterative forward prediction methods are based on thresholding the absolute value residual, backward prediction error and forward prediction error respectively, where the threshold values is not signal dependent and does not require rigorous tuning. In both cases, different threshold values were tested and a value of $K = 3$ led to the best results in agreement with the "3-sigma" rule [6]. The parameters used during the simulations are shown in Table I.

Table I: Simulation Parameters

No	Description	Value
1	Sampling frequency	44.1kHz and 8kHz
2	Frame size	32.5 ms
3	Conventional LP order	12
4	HOSpLP order	half of frame size
5	Number of non-zero *l0*-norm regularized HOSpLP coefficients	20
6	Artificial click duration	0.4536ms - 2.268ms
7	Local window size, *kmax*	5

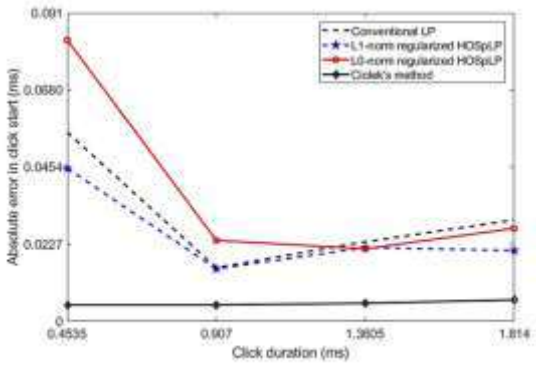
A. Click Detection Performance

1) Estimation of start of click: The backward prediction based click detection is heavily dependent on correct estimation of the start of the click degradation. To evaluate the performance of the backward prediction based click detection in the estimation of the start of the click, average absolute error in estimating the click start is shown in Figure 1 by using conventional LP and HOSpLP coefficients in the backward prediction method. The method proposed by Ciołek et.al. is also taken as a benchmark.

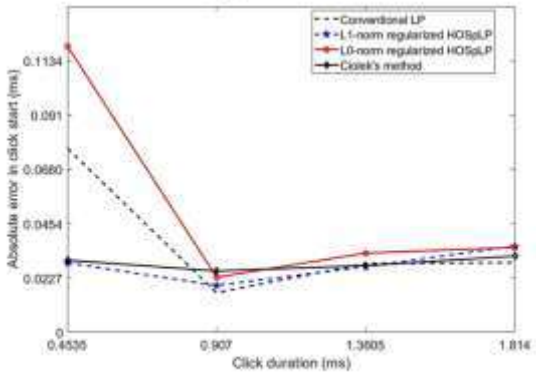
It is seen that Ciołek's method leads to the best estimation of the start of the click. However, note that at 44.1 kHz sampling frequency, 0.0227 ms is 1 samples, as such the backward prediction method on average leads to click start error of 1 samples only. The conventional LP and HOSpLP coefficients perform similarly in the estimation of the start of the click. The absolute error of estimation is on average 0.0227 ms, i.e. 1 sample at 44.1 kHz sampling frequency, for click degradation of duration up to 2.268 ms or 100 samples.

2) Estimation of click duration: Figure 2 shows the NMSE for click duration estimation for speech and music by using backward prediction based on conventional LP and HOSpLP coefficients and by using Ciołek's method. It is observed that for click duration less than 1 ms, the backward prediction based click detection fails entirely. However, for longer click durations the backward prediction based on HOSpLP leads to superior click duration estimation performance for music.

This is in agreement with the modeling assumption made regarding the HOSpLP coefficients for music. It is noted that for music at 44.1 kHz sampling frequency, even though Ciołek's method leads to superior identification of the click start, its estimation of the click duration is inferior to the backward prediction based method for both conventional LP and HOSpLP coefficients.



(a) Music



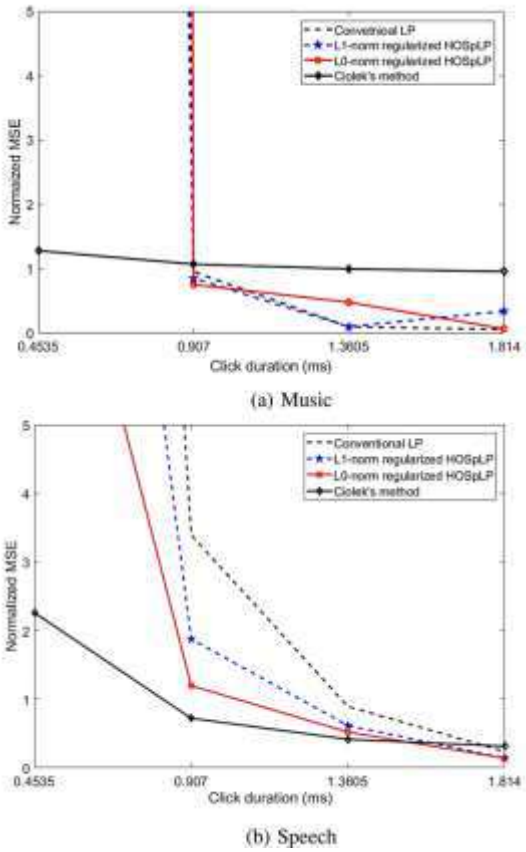
(b) Speech

Figure 1: Absolute error in click start estimation using backward prediction using HOSpLP coefficients.

To see the impact of the sampling frequency on the click estimation of the methods, similar experiments were conducted for audio sampled at 8 kHz. Figure 3 shows the NMSE for click duration estimation for speech and music by using backward prediction based on conventional LP and HOSpLP and by using Ciołek's method for a wide range of click durations.

It is observed that for long click durations (longer than 4 ms), all methods yield similar detection performance. However, as the click duration decreases, the conventional LP and $l1$-norm regularized HOSpLP accuracy decreases significantly.

The use of backward prediction with $l0$-norm regularized HOSpLP coefficients leads to the best click duration estimation results for all click durations, except for very short click durations (less than 1 ms) where all methods fail. For music, it is seen that the $l1$-norm regularized HOSpLP performs best for long click durations. This performance of the backward prediction



(a) Music

(b) Speech

Figure 2: Performance of click duration estimation at 44.1 kHz sampling frequency.

method with HOSpLP is consistent at both sampling frequencies where as Ciołek's method leads to inferior performance as the sampling frequency is increased.

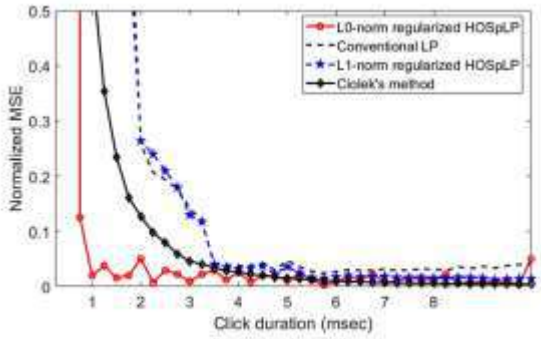
B. Detection and Restoration performance

To measure the unified detection and restoration performance, the artificially click degraded audio was restored by using the proposed Algorithm 2 and state-of-the-art Algorithm 3 then the SNR was computed and averaged for each dataset. No information regarding the location and duration of the click degradation is used in any of the methods. Figure 4 shows the results of the detection and restoration for audio sampled at 44.1 kHz.

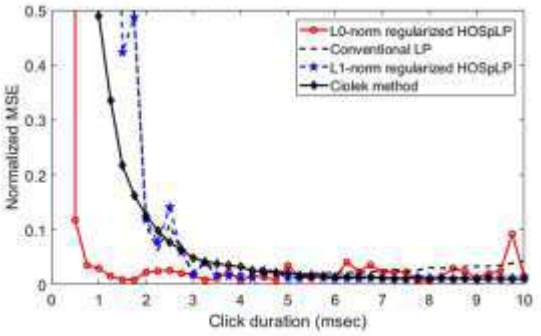
For audio sampled at frequency of 44.1 kHz the backward prediction method with HOSpLP coefficients leads to superior restoration performance as compared to Ciołek's method. This is attributed to the superior click duration estimation performance of the proposed backward prediction method with HOSpLP coefficients as compared to Ciołek's method.

It is also noted that the use of HOSpLP coefficients in Ciołek's method leads to improvement in restoration performance as compared to conventional LP in Ciołek's method. The improvement in SNR by the HOSpLP based methods is observed to be higher in music as compared to speech.

This can also be attributed to the superior modeling capability of HOSpLP coefficients in the case of music. This has been also seen to be the case in HOSpLP coefficient based restoration methods as reported in our previous works [13] and [12].

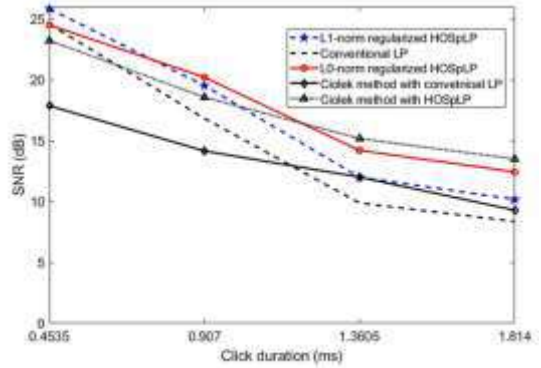


(a) Speech

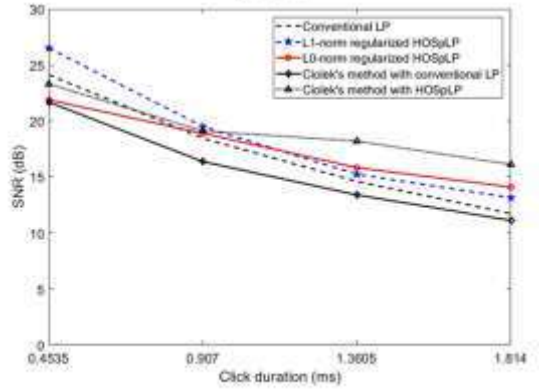


(b) Music

Figure 3: Performance of click duration estimation at 8 kHzsampling frequency.



(a) Music



(b) Speech

Figure 4: SNR of restored audio by using detection and restoration without any a priori knowledge on location and duration of click degradation.

Figure 5 show the SNR improvement obtained by using the backward prediction method with HOSpLP coefficients and Ciołek's method for the detection and restoration of click degraded audio sampled at 44.1 kHz. This is the difference between the SNR of the restored audio and the SNR of the click-degraded audio.

It is seen that all restoration methods achieve significant SNR improvement over the click-degraded audio. The proposed backward prediction method with HOSpLP coefficients for click detection and restoration is observed to lead to SNR improvement up to 4.5dB over Ciołek's method using conventional LP. On average both backward prediction and Ciołek's method performs similarly when using HOSpLP coefficients. This seems to indicate that the use of HOSpLP coefficients in both approaches is the reason for the improvement in restoration performance.

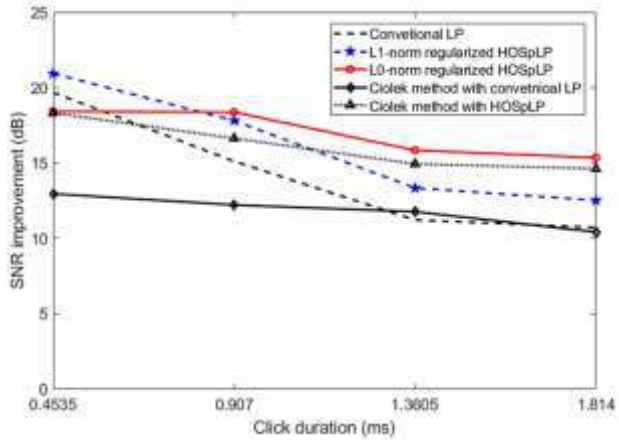


Figure 5: SNR improvement by detection and restoration without any a priori knowledge on location and duration of click degradation.

To see the impact of the sampling frequency on the restoration performance of the backward prediction method with HOSpLP and Ciołek's method, similar experiments were conducted for audio sampled at 8 kHz. Figure 6 shows the results of the detection and restoration for audio sampled at 8 kHz. At this sampling frequency the backward prediction method with HOSpLP coefficients leads to higher SNR for most click durations. The use of HOSpLP coefficients in Ciołek's method is observed to lead to better SNR as compared to conventional LP for higher click durations. This also shows the superior

C. Perceptual evaluation of audio quality

PEAQ was used to estimate the subjective quality of the audio signal that is restored by using the proposed backward prediction method with HOSpLP coefficients and Ciołek's method. The PEAQ was calculated for each audio fragment as the original clean signal is available.

The result of each fragment was then averaged for each type of audio. Table III and III show the PEAQ evaluation obtained for by using the backward prediction method with HOSpLP, Ciołek's method and Ciołek's method with HOSpLP for music and speech respectively.

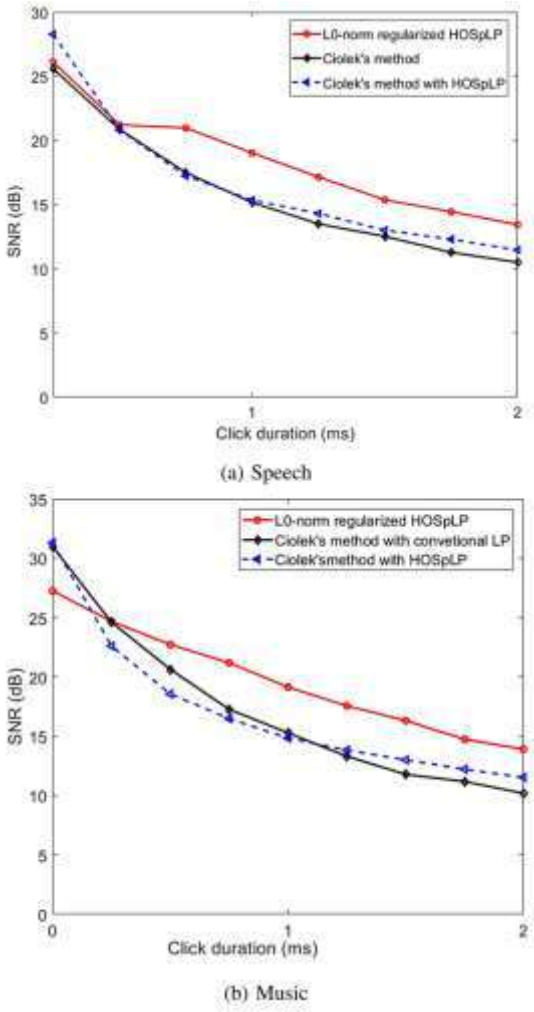


Figure 6: SNR of restored audio sampled at 8 kHz by using detection and restoration without any a priori on location and duration of click degradation.

It is seen that, the use of *l0*-norm and *l1*-norm regularized HOSpLP coefficients in the backward prediction click detection and then restoration leads to better PEAQ results as compared to conventional LP. However, it is noted that the *l*1-norm regularized HOSpLP coefficients lead to higher PEAQ results as compared to *l*0-norm regularized HOSpLP coefficients even though in terms of SNR *l0*-norm regularized HOSpLP coefficients lead to better results. This may be attributed to the better modeling capabilities of *l*1-norm regularized HOSpLP

coefficients especially for music. For speech, the use of HOSpLP coefficients in Ciołek's method is not observed to lead to significant improvement in PEAQ as compared to conventional LP Ciołek's method. However, for music the use of HOSpLP coefficients in Ciołek's method leads to significant improvement in PEAQ as compared to conventional LP. This again, shows the better modeling capability of HOSpLP coefficients for music.

Table II: PEAQ evaluation for Music

Method	**Click duration in ms**				
	0.454	0.907	1.361	1.814	2.268
Backward prediction with Conventional LP	-0.84	-1.13	-0.98	-1.29	-1.55
Backward prediction with *l1*-norm HOSpLP	-0.67	-0.99	-0.85	-1.24	-1.34
Backward prediction with *l*0-norm HOSpLP	-0.68	-0.81	-0.97	-1.27	-1.47
Ciolek's method	-1.13	-1.14	-0.93	-0.90	-0.95
Ciolek's method with *l1*-norm HOSpLP	-0.65	-0.75	-0.62	-0.68	-0.91

Table III: PEAQ evaluation for speech

Method	**Click duration in ms**				
	0.454	0.907	1.361	1.814	2.268
Backward prediction with Conventional LP	-0.54	-0.64	-0.76	-0.79	-0.89
Backward prediction with*l1*-norm HOSpLP	-0.37	-0.65	-0.75	-0.60	-0.77
Backward prediction with *l0*-norm HOSpLP	-0.44	-0.56	-0.68	-0.76	-0.81
Ciolek's method	-0.67	-0.41	-0.49	-0.57	-0.59
Ciolek's method with *l1*-norm HOSpLP	-0.38	-0.47	-0.46	-0.49	-0.54

D. Impact of amplitude of click degradation

A challenge for the click detection that has not been discussed is the amplitude of the click degradation, represented here by the variance of the assumed click generating-random process, σ_c^2. As the causes of click degradation are very diverse it is quite difficult to assume a single value for the variance of the click-generating random process. As such, even in a single recording, click degradation with very different amplitudes will be present. To evaluate the performance of the proposed HOSpLP-based click detection and restoration method for click degradations of different variance, the SNR improvement is evaluated by degrading the audio with click degradations having variance the same as the audio signal ($\sigma_c^2 = \sigma_s^2$) andquarter of the audio signal ($\sigma_c^2 = \frac{\sigma_s^2}{4}$).

Figure 7 shows the SNR improvement by the backward prediction method with HOSpLP and Ciołek's method with HOSpLP when the variance of the click generating random process is varied for speech and audio sampled at 8 kHz. It is seen that the three methods achieve significant SNR improvement. For click durations more than 0.5 ms, the proposed backward prediction method with HOSpLP and Ciołek's method with HOSpLP lead to a much better SNR improvement as the variance of the click-generating process decreases. However, for very short click durations, the backward prediction method with HOSpLP is inferior to Ciołek's method. It is also noted that as the variance of the click-generating random process decreases, Ciołek's method with HOSpLP leads to significant improvement as compared to the other two.

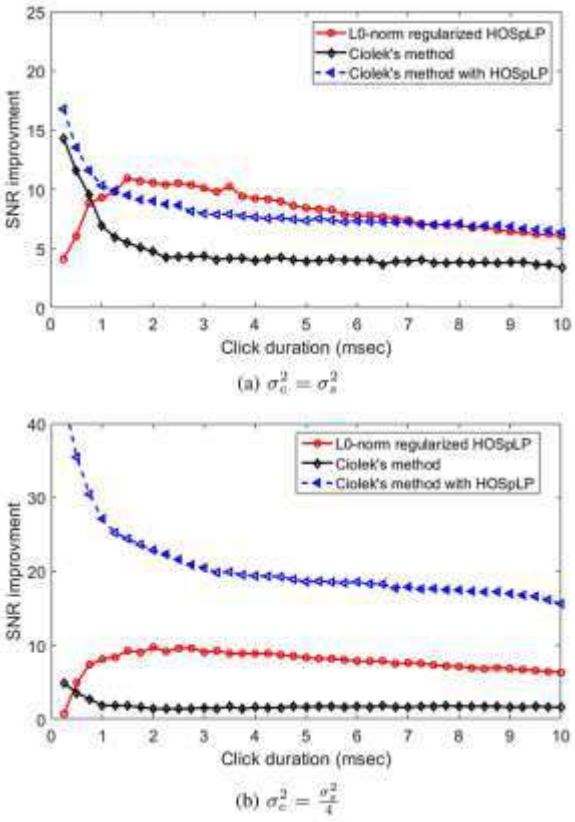


Figure 7: SNR improvement by detection and restoration without any a priori knowledge on location and duration of click degradation for music for different click degradation variance.

CONCLUSIONS

In this paper, the use of high-order sparse linear predictions proposed for the detection of clicks and restoration of audio corrupted by click degradation. The use of the HOSpLP coefficients is suitable for both speech and tonal audio without a prior knowledge about the type of signal or click degradation. Several experiments were conducted to assess the performance of the proposed method in terms of click detection, restoration performance and robustness to the degrading click variance. The proposed methods achieved an improvement in SNR over conventional LP and a recently proposed method that also jointly detects and restores click degraded audio for speech and music. Even though both *l1*-norm and *l0*-norm regularized HOSpLP-based methods are not real-time, by using efficient ADMM and proximal gradient algorithm, the computation time can be limited to 2-3 times the duration of the frame under consideration on current general purpose computer. Considering the application at hand is for the restoration of archived audio media, the computational time is not expected to be a significant limitation.

Only artificial click degradation was considered in our experiments. A next step is to evaluate the proposed methods under real-life click degradation conditions. However, as the click-degraded samples are first discarded before restoration, working with real click degradations will only affect the detection and not the restoration performance.

REFERENCES

[1] GodsilS J and Rayner, P. J. W. *Digital audio restoration: a statistical model based approach*. Springer, January 1998.

[2] Van Waterschoot T. and Moonen, M. Moonen, "Comparison of linear prediction models for audio signals," *EURASIP J. Audio, Speech, Music Process*., vol. 20(5), pp. 1644–1657, July 2008.

[3] Ruandaidh J. O. and Fitzgerald, W. Fitzgerald, "Interpolation of missing samples for audio restoration," *IEEE Electronics Letters*, vol. 30(8), pp. 622–623, April 1994.

[4] Niedzwiecki M. and CiołlekM., "Elimination of clicks from archive speech signals using sparse autoregressive modeling," *Proc. 21nd European Signal Process. Conf.*

(EUSIPCO l12), pp. 2615–2619, August2012.

[5] NiedzwieckiM., Cioek, M. and Cisowski, K. "Elimination of impulsive disturbances from stereo audio recordings using vector autoregressive modeling and variable-order Kalman filtering," *IEEE Trans*. *Audio Speech Lang*. *Process*., vol. 23(6), pp. 970–981, June 2015.

[6] CiołekM. andNiedzwieckiM., "Detection of impulsive disturbances in archive audio signals," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (New Orleans, LA,USA), March 2017.

[7] GiacobelloD., Christensen M. G., Murthi M. N., S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," *IEEE Trans*. *Audio Speech Lang*. *Process*., vol. 20(5),pp. 1644–1657, July 2012.

[8] Janssen A., Veldhuis R., and Vries L., "Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes," *IEEE Trans., Acoust., Speech, Signal Process*., vol. 34(2), pp. 317–330,Apr. 1986.

[9] Kabal P. and Ramachandran R. P., "Joint optimization of linear predictors in speech coders," *IEEE Trans*. *On Acoust., Speech and Signal Processing*, vol. 37(5), p. 642 650, May 1989.

[10] Shi L., Jensen J. R., and Christensen M. G., "Least 1-norm pole zero modeling with sparse deconvolution for speech analysis," in *Proc*. *2017 IEEE Int*. *Conf*. *Acoust., Speech, Signal Process*. *(ICASSP 17)*, (New Orleans, LA, USA), June 2017.

[11] Giacobello D., Christensen M. Dahl G., J., Jensen S. H., and MoonenM.,"Joint estimation of short-term and long-term predictors in speech coders," in *Proc*. *2009 IEEE Int*. *Conf*. *Acoust., Speech, Signal Process.(ICASSP)*, (Taipei, Taiwan), pp. 409–412, IEEE, Apr. 2009.

[12] Dufera B., Eneman D., K., and van WaterschootT., "Missing sample estimation based on high-order sparse linear prediction for audio signals," in *26th European Signal Processing Conference, EUSIPCO 2018*,(Roma, Italy), pp. 2464–2468, September 3-7, 2018.

[13] Dufera B. D., Adugna E., Eneman K., and van Waterschoot T., "Restoration of click degraded speech and music based on high order sparse linear prediction," in *IEEE AFRICON 2019*, (Accra, Ghana), September25-27, 2019.

[14] Giacobello D., van Water schoot T., Christensen M. G., Jensen S. H., andMoonen M., "High-order sparse linear predictors for audio *Process*. *Conf*. *(EUSIPCO l10)*, (Aalborg, Denmark), pp. 234–238, August 2010.

[15] Jensen T. L., GiacobelloD., van Waterschoot T., and M. G. Christensen, "Fast algorithms for high-order sparse linear prediction with applications to speech processing," *Speech Communication*, vol. 76(5), pp. 143–156,July 2016.

[16] Hansen P. C., "Analysis of discrete ill-posed problems by means of thel-curve," *SIAM Review*, vol. 34(4), pp. 561–580, Dec. 1992.

[17] Toledano D. T., Gimenez A. O., Teixeira A., Rodruıguez J. G., Gomez L. H., Hernandez R. S. S., and Castro D. R., "*Advances in Speech and Language Technologies for Iberian Languages"*. Springer, November2012.

[18] Antonello N., Stella L., Patrinos P., and van Waterschoot T., "Proximal gradient algorithms: Applications in signal processing," *arXiv:1803.01621*, March 2018.

[19] Vaseghi S. V. and Rayner P. J. W., "Detection and suppression ofimpulsive noise," in *speech communication systems*. *IEE Proceedings*, pp. 38–46, 1990.

[20] Niedzwiecki M. and Ciołek M., "Renovation of archive audio recordings using sparse autoregressive modeling and bidirectional processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, (Vancouver, BC, Canada), May 2013.

[21] Niedzwieck M. i and Cisowski K., "Adaptive scheme for elimination of broadband noise and impulsive disturbances from ar and arma signals," *IEEE Trans. on Audio, Speech, Lang. Processing*, vol. 14(1), pp. 967–982, March 1996.

[22] Ciołek M. And NiedzwieckiM..http://eti.pg.edu.pl/katedra-systemow-automatyki/ICASSP2017.

[23] Bang and Olufsen, "Music for archimedes,"

[24] Avila F. R. and Biscainho L. W. P., "Bayesian restoration of audio signals degraded by impulsive noise modeled as individual pulses," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20(9), pp. 2470–2480, November 2012.

[25] ITU-R, "Method for objective measurements of perceived audio quality," Recommendation 1387-1, International Telecommunication Union,1998-2001.

[26] Kabal P., "An examination and interpretation of itu-r bs.1387: Perceptual evaluation of audio quality," tsp lab technical report, Dept. Electrical and Computer Engineering, McGill University, May 2002.