Addis Ababa University
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

The use of High-Order Sparse Linear Prediction for the Restoration of Archived Audio

by

Bisrat Derebssa Dufera

A PhD dissertation submitted to the School of Electrical and Computer Engineering, Addis Ababa Institute of Technology, Addis Ababa University in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Computer Engineering.

June, 2020

Addis Ababa University
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

By: ***Bisrat Derebssa Dufera***

This is to certify that the dissertation prepared by Bisrat Derebssa Dufera, titled ***The use of High-Order Sparse Linear Prediction for the Restoration of Archived Audio*** and submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy (PhD) in Computer Engineering complies with the regulations of the University and meets accepted standards with respect to originality and quality.

Approved and signed by board of Examining Committee

	Name	Signature	Date
Dean, School of Electrical and Computer Engineering	Dr. Yalemzewd Negash	_____	_____
Supervisor	Prof. Toon van Waterschoot	_____	20/04/2020
Supervisor	Prof. Koen Eneman	_____	14/05/2020
Supervisor	Dr. Eneyew Adugna	_____	_____
Internal Examiner		_____	_____
External Examiner		_____	_____

Declaration

I, the undersigned, declare that this dissertation submitted to the School of Electrical and Computer Engineering, Addis Ababa Institute of Technology, Addis Ababa University in partial fulfillment of the requirement for the degree of Doctor of Philosophy (PhD) in Computer Engineering is my original work and has not been presented for a degree in any other university and that all source materials used have been properly acknowledged.

Name: **Bisrat Derebssa Dufera**

Signature: _____

Date: _____

Acknowledgment

Firstly, I would like to express my sincere gratitude to my advisors Professor Toon van Waterschoot, Professor Koen Eneman and Dr. Eneyew Adugna for their guidance and continuous support throughout my Ph.D study. Their guidance helped me in all the time of research and writing this thesis.

I would also like to thank the Ministry of Education and HGPP project funded through GIZ GmbH for giving me the chance of research stay at KU Leuven, Belgium. I would like to thank Dr Dereje Hailemariam for his support and understanding in my stay at KU Leuven.

Above all I would like to thank my wife Melat Yeshaw for her love, inspiration and constant support throughout the past challenging years. To my kids Yannet and Makda, even though you may not realize it, you have been a great source of happiness and needed distraction from the stress of research.

Abstract

Since the invention of Gramophone by Thomas Edison in 1877, vast amounts of cultural, entertainment, educational and historical audio recordings have been recorded and stored throughout the world. Through natural aging and improper storage, the recorded signal degrades and loses its information in terms of quality and intelligibility. Degradation of audio signals is considered as any unwanted modification to the audio signal after it has been recorded. There are different degradations affecting recorded signals on analog storage media. The degradations that are often encountered are *clicks, hiss* and *'Wow and Flutter'*.

Several researches have been conducted in restoring degraded audio recordings. Most of the methods rely on some prior information of the underlying data and the degradation process. The success of these methods heavily depends on the prior information available. When such information is not available, a model of the underlying undegraded data can be used to generate such prior information. Linear prediction is one of the most widely used models to represent speech. However, linear prediction has limitations for voiced speech and music and as such restoration approaches that use linear prediction have limited success for voiced speech and music.

This research uses recent findings in linear prediction modeling in the restoration of click and 'wow and flutter'. Recent developments in efficient algorithms and computational capability have led to significant investigations on the usefulness of $\ell 1$-norm and $\ell 0$-norm regularization in the solution to the least squares problem. The use of high-order sparse linear prediction for overcoming the limitations posed by conventional linear prediction has been investigated by other researchers. This research investigates the use of high-order sparse linear prediction for the detection and restoration of degraded archived audio signals.

A method is developed that uses the high-order sparse linear prediction model to estimate the underlying audio signal without priori information on the type of audio and the details of the degradation. The model is then used for the detection of the degradations as well as for the restoration of the degraded sample values. The use of the model for two of the most widely encountered degradations in archived audio is investigated.

Results show that the use of high-order sparse linear prediction for the modeling of the underlying audio signal results in improved detection as well as restoration. Simulations are conducted for a wide range of audio signals including synthetic vowels, natural vowels, speech and music. The method was able to be used without prior information on the type of audio as well as without the need of pitch estimators. The performance was measured with respect to degradation characterization and restoration quality: in terms of signal-to-noise ratio of the restored signal versus the original undegraded signal and perceptual evaluation of audio quality for assessment of the subjective quality of the restored signal. Both results showed that the proposed framework achieves better quality of all types of audio signal. The computational time of the proposed framework was also investigated.

Contents

List of Figures

List of Tables

List of Acronyms

ADMM Alternating Direction of Multipliers

AR Autoregressive

CRB Crammer-Rao Bound

STDFT Short time Discrete Fourier Transform

EBC Ethiopian Broadcasting Corporation

ERTA Ethiopian Radio and Television Agency

EKF Extended Kalman Filter

HOSpLP High-Order Sparse Linear Prediction

IASA International Association of Sound and Audiovisual Archives

IES Institute of Ethiopian Studies

LP Linear Prediction

LS Least Square

MAP Maximum a Posteriori

MSE Mean Square Error

NMSE Normalized Mean Square Error

PEAQ Perceptual Evaluation of Audio Quality

PCM Pulse Code Modulation

PVC Polyvinyl Chloride

SLP Sparse Linear Prediction

SNR Signal to Noise Ratio

Chapter 1

Introduction

Sound recording is the process of transforming the acoustic energy of sound into some form in which it can be stored and reproduced at later time [1]. Thomas Edison was the first who successfully recorded sound that could be played back in 1877 by recording a human voice signal on a tinfoil wrapped around a cylinder. Since then recorded sound has been an important aspect of human life. Music which was a unique and live performance, experienced with a group in a public place, now with sound recording can be listened to at home over and over again. With sound recording, teachers could start bringing sound recordings to the classroom for improved delivery of the course. Another important aspect of recorded sound is the preservation of oral culture. Folklore and ethnographic audio collections, preserve audio windows into a range of geographical areas and cultures. Yet another aspect of recorded sound is radio broadcast archives and collections. "Perhaps no other sound medium has conveyed to listeners so much history and culture, through music, diverse entertainment programming, daily news and public affairs, and interview" [2]. However, due to physical and chemical properties of the storage medium and the storage environment, these media are at risk of degradation.

A survey conducted by Heritage Preservation Inc. in collaboration with the U.S. Institute of Museum and Library Services of public institutions estimated that there are around 46 million individual recordings in these institutions of which 44% of their audio collections were in "unknown condition" [2]. The Music Archive of the Ethnological Museum (Germany) houses 354 cylinder collections with approximately 16,800 recordings on slightly over

30,000 cylinders and approximately 3,500 magnetic tapes. The condition of the recordings is not well known [3]. The British Library estimates that of the 6.5 million sound recordings in its archive, 1 million of them are in risk of disappearing [4].

According to [5] the history of electronic media in Ethiopia starts with the first provisional radio station which was inaugurated in 1933 in a contract signed with an Italian Company. The most important audio-visual libraries in Ethiopia are the Ethiopian Broadcasting Corporation (EBC) audio-video library, National Library of Ethiopia, and the Institute of Ethiopian Studies (IES) library of Addis Ababa University. In addition, there are recordings made by travelers as early as 1900. It is difficult to obtain literature on the status of storage media in Ethiopian Libraries. This may stem from lack of understanding of the risks facing storage media, lack of funding or lack of experienced staff that can undertake the task of preservation.

Several types of storage media have been used according to advances in technology and demands of the market. Of the different storage media that have been used in the last century, the vast majority of historical, cultural, entertainment and research have been stored in the following storage media:

- Phonograph Cylinders,
- Disc Record,
- Magnetic Tape and
- Compact Disc.

The stored data in these storage media is faced with two important and unavoidable problems:

- Deterioration through natural aging and improper storage environment and
- Format obsolescence.

Therefore, libraries and archivists are faced with the task of copying these contents into more permanent media. Tasks that need to be taken by libraries and archivists can fall into the following categories:

- Preservation,

- Digitization, and
- Restoration.

Even though the long term solution to this set of problems is digitizing the archives and storing them in the latest format, most estimates agree that institutions have only a small window of time in which to complete high-quality digitization, depending on storage conditions, the storage media and other factors [6]. According to [6], the overall quality of digitization is dependent on several factors including:

- Original media quality and condition,
- Reproducer availability and quality and
- Analog-to-digital conversion quality.

While the analog-to digital conversion process is increasing in quality, the original tape quality, its condition and reproducer quality are deteriorating in time. The combination of these factors shows that we have limited time for high quality digitization.

Typical restoration may include removing the following degradations.

- **Clicks**: these are defects that of finite duration which occur at random positions in time with random amplitudes [1]. They are usually encountered in physical storage media due to dust, wear and groove damage. These defects cause undesirable deviations of short duration in the pickup system [7].
- **Hiss:** this is random additive noise composed of circuit noise, ambient noise from the recording and playback environment and irregularities in the storage media.
- **Wow and flutter:** these are pitch variation defects due to speed variation between playback and recording.
- **Low frequency noise:** these are usually associated with large scratches on the surface of a storage media. These scratches cause long-term resonance in the playback system.

These degradations can be split into two categories: *global degradation* and *local degradation*. In local degradation, the signal is modified or corrupted in short duration segments and restoration is only needed at these localized places. In global degradation, the signal is modified to some extent at all times. Clicks are categorized as local degradation; while, hiss, "wow and flutter" and low frequency noise are categorized as global degradation [8, 9].

The use of analogue restoration methods existed as early as magnetic tape. Techniques in the form of frequency domain equalization for background noise and manual cut-and-splice editing for clicks [1] have been used on magnetic tapes. Unfortunately, these methods "are not sophisticated enough to achieve a significant level of noise reduction without interfering with the underlying signal quality" [1]. Digital methods on the other hand enable a much greater degree of flexibility in processing, and therefore greater potential for noise reduction.

Digital audio restoration techniques are often performed on a copy of the digital audio data after digitizing the audio source from the analogue media. Once in the digital realm, recordings can be restored by using dedicated digital signal processing units or by implementing different digital signal processing algorithms on general purpose computers. As a result of the restoration effort, some undesired distortion may result in the audio signal. Hence, a restoration system should consider the trade off between the audibility of the degradation and perceived distortion of restoration processing. The aim of any audio restoration technique is therefore to remove only those artifacts which are audible to the listener.

Of these degradations the focus in this PhD research is on **clicks** and **'Wow and Flutter'** as these have significant impact on the quality of the audio and are more specific to archived media. The other two degradations are more common in other areas and significant research has already been done by other researchers.

1.1 Clicks

Click degradation refers to short duration artifacts which occur at random positions in an audio signal [1]. It is perceived by the listener as impulsive noise ranging from tiny 'tick' noises, 'scratch' to 'crackle' noise. These artifacts are due to physical defects on the medium from microscopic surface defects to physical breakage of the medium [9]. Clicks can be modeled as an

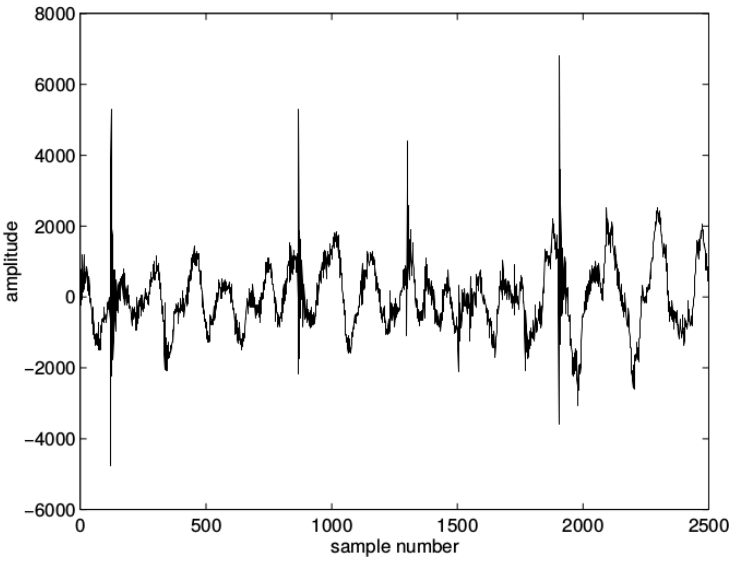


Figure 1.1: Typical Click Degradation.

additive or replacement degradation. An additive model, where the 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, may be appropriate for very large scratches and breakages which completely obliterate the underlying signal information.

A typical click degradation is shown in Fig. 1.1. The audio was digitized using a sampling frequency of 44100 Hz. There are four distinct occurrences of click degradation in Figure 1.1. They are located around sample numbers $121, 875, 1306$ and 1912. Each degradation is localized and has a very short duration. In addition, their amplitude is significantly higher than the maximum amplitude of the underlying signal.

Ideally, an interpolation or missing sample estimation technique used to restore click degraded audio signals should not modify the undegraded signal and restore only those samples which are degraded. Two tasks are therefore important for a successful restoration of click degraded audio signal: detection of degraded signal samples/segments and estimation of the underlying click degraded samples.

1.1.1 Click Detection

Click detection involves the "identification of samples which are not drawn from the same stochastic process as the underlying clean audio signal. In other words, they are drawn from some spurious outlier distribution" [1]. There has been a lot of research conducted in outlier detection in the field of statistical data analysis. Various criteria for detection are possible; however, as the aim of audio restoration is to remove those artifacts which are audible, click detection should detect those clicks that are audible.

Some of the most widely used click detection methods are based on frequency domain filtering and autoregressive modeling.

- **Highpass Filtering**: This approach is based on the assumption that most audio signals contain little information at high frequencies (greater than $8 - 16$ kHz), while clicks, like *impulses*, have spectral content at all frequencies. Therefore, by using a highpass filter, clicks can be enhanced relative to the underlying signal [10]. Thresholding can be used after the filtering to detect those sections of the audio signal degraded by click.

  This method was one of the first click detection methods used in both analog and digital equipment [1]. It has the advantage of being simple to implement and having no unknown system parameters, except the highpass filter cutoff frequency, ω_0 and the detection threshold. Taking a very high ω_0, and detection threshold will make the system insensitive to small clicks, while taking a low value will incorrectly detect part of the audio signal to be a click.

  The method will fail if the clicks are band limited or if the signal itself has a high frequency content, such as high pitched musical instruments. In addition, these methods are dependent on assumptions that cannot be verified or rejected during detection. As such, even though their implementation is one of the easiest, their applicability is very limited.

- **Autoregressive (AR) model-based click detection**: Model-based click detection methods use prior information about the undegraded signal and the clicks into the detection procedure in the form of hypothesized signal models. In this approach, the undegraded audio signal is assumed to be drawn from a short-term stationary AR 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 [11].

For AR modeling of an undegraded 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 undegraded 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], [12], [13], [14], [15]. 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 the 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 may be 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], [11], [16].

1.1.2 Estimation of click degraded samples

Several approaches for the restoration of audio signals degraded by clicks have been proposed. A Maximum a Posteriori (MAP) Interpolator that uses Bayesian inference as a means to incorporate prior information about the restoration problem has been proposed in [8]. In this method, no assumption is made about the underlying signal generation process except that the click degradations samples are generated from a zero-mean multivariate Gaussian

process. It combines the click detection and estimation procedures by maximizing the a posteriori probability of the missing samples given the click degradation configuration. It has the appeal of being the 'most probable' solution of all the possible solutions. However, the problem of click detection and estimation involves searching through all possible click degradation states. This is computationally impractical as there are a total of 2^N different click states for a signal of length N. Different approaches have been proposed to overcome this limitation of the MAP interpolator. The methods rely on incorporating additional prior information in the interpolation. Linear Prediction (LP) based methods that incorporate the source-filter model of speech production have been proposed and used extensively in the literature for restoration of click degraded speech and audio signals [1], [8], [17], [18].

The Least Square (LS) autoregressive interpolator [1] is based on the assumption that the underlying audio signal is generated by passing an excitation through an all-pole filter and the click degradation samples are mutually independent and drawn from a Gaussian zero-mean process. The missing samples then can be obtained from prior knowledge of LP coefficients of the underlying signal, the time range of known samples and the time range of degraded samples.

One of the limitations of the LS solution is the unavailability of the AR coefficients, short-term predictors, of the undegraded underlying signal. An iterative method for the estimation of the AR coefficients and unknown samples was proposed by Janssen *et. al.* [19]. The method minimizes a sum of squared residual errors involving the LP coefficients, known samples and the unknown samples from a sufficiently large neighborhood as a function of the unknown samples and the unknown LP coefficients. Minimization with respect to the LP coefficients followed by minimization with respect to the unknown samples are performed in each iteration. The $\ell 2$-norm of the residual, the difference between the actual and predicted signal, is minimized to estimate the AR coefficients. This works well for unvoiced speech; that is, when the samples of the excitation are Gaussian and independent identically distributed random processes [16]. However, the use of the $\ell 2$-norm of the residual vector is not suited for voiced speech, as the excitation is spiky and quasi-periodic, and music [19]. In this case, the $\ell 2$-norm gives more emphasis to the periodic spikes of the residual [11]. As a result, it trades off short-term predictor, spectral envelope, estimation accuracy to estimating the long-term predictors, harmonics, due to the excitation [11]. Several methods have been proposed to alleviate these limitations. The use of Sparse Linear Predic-

tion (SLP), that tries to maximize the sparsity of the residual as well as the prediction coefficients, has been shown to lead to less emphasis on the periodic spikes of the residual. By using High-Order Sparse Linear Prediction (HOSpLP) a more efficient decoupling between the pitch harmonics and the spectral envelope has been achieved in [16], [20], [21]. Even though it was originally used for speech processing purposes, it has found applications in many fields such as radar processing [22], geology and general signal representations [23].

Another limitation of the LS interpolator is the need of a pitch predictor to predict long-term correlation.

For tonal audio analysis, that is, signals containing 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 [11]. This makes it hard to propose a generic audio signal generation model [11]. In addition, "each polyphonic audio signal should be analyzed using multiple source-filter models, which seems to be rather impractical" [11].

1.1.2.1 Adaptive Click Estimation

An adaptive signal processing scheme has been proposed for the removal of clicks by [9]. It uses least mean square and normalized least mean square approaches and has obtained good click removal results. The adaptive filter coefficients are updated depending on the error between the observed signal and the restored signal. This is an iterative procedure and it is stopped when either the error is very small or the filter coefficient update is small.

Improved adaptive algorithms have been proposed. Some of the most prominent are discussed below.

- **Kalman Filtering [24]:** This method tries to simultaneously solve the problems of filtering, parameter tracking and elimination of the outliers ("clicks"), by using the Extended Kalman Filter (EKF) theory. In this solution, the audio signal is modeled by a time varying autoregressive model of order P. The problem of filtering and parameter tracking are strictly tied together and solved jointly. This is essentially an iterative update procedure. This iterative procedure is computationally intensive. Some improvements proposed by [25] are the following.

- **Bootstrap Procedure**: the first $100ms$ of the signal are time-reversed and fed to the filter so that the parameters for the initialization of the model are estimated properly.
- **Stability Check**: the estimated time-varying AR-model is not guaranteed to be stable at all times. Therefore stability is checked at each time by checking the magnitude of the reflection coefficients by using the Levinson recursion.

The performance of the system was tested on clean audio recording that was intentionally corrupted with clicks and scratches. There are two modes of operation: *without fine-tuning the EKF filter parameters* and *with fine-tuning the EKF parameters*. Listening tests that were conducted to compare the effectiveness of the EKF system with that of state of the art showed that the proposed method yields comparable result to the state of the art systems.

1.1.2.2 Fusing Multiple Copies Click Removal

A technique for removing clicks from audio signals that fuses multiple copies of the same recording has been proposed [7]. If there are multiple copies of a given recording all from the same master recording, this algorithm exploits the fact most degradation in audio signals are record-dependent. The first step is to align the signals in an optimal non-rigid alignment that is robust to the presence of sparse outliers with arbitrary magnitude. By assuming an isolated interval of missing components in the signal set and using dynamic time wrapping algorithms, they have derived an equation the restored signal should minimize. They tested this approach on three copies of the same recording. The method was able to precisely align long segments of audio. In addition, the click removal method was applied successfully on a long duration click-degraded signal segment (> 100) samples to demonstrate its advantage.

1.2 Wow and Flutter

Wow and Flutter refers to overall pitch variations which are not present in the original audio signal [1]. It is perceived as an "undesired frequency modulation in the range of approximately 0.5 to $6Hz$" [26]. This variation can be due to the following [1].

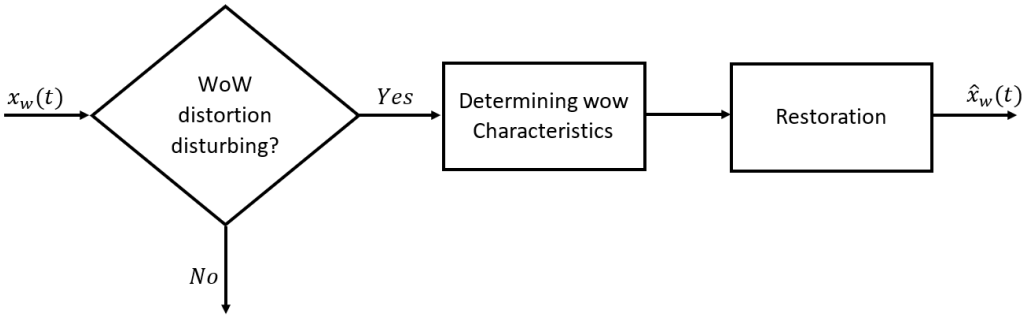


Figure 1.2: Typical wow restoration approaches.

- Variation of rotational speed of recording medium,
- Physical deformation of storage medium,
- Uneven stretching of magnetic tape and others.

Preventive measures can be taken to minimize the effect these degradations in the analog domain, on the physical storage media itself. For example, in principle it is possible to correct a poorly punched disc center hole. However, these measures can put the physical media in danger of damage and cannot completely remove the degradation [26]. Therefore, the distortion can be found on wax cylinders, disks and magnetic tapes. Digital signal processing techniques provide an additional means for the detection and removal of these degradations once the analog signal is converted to digital. Some research has been done to model and remove these degradations [27] [28], [26], [29].

Wow and flutter degradation is generally modeled as a non-uniform warping of the time axis [1]. The main task of a system used for the restoration of audio signals degraded by wow degradation is therefore the estimation of the time warping function or the pitch variation function. If the time-warping function is known and invertible then the original signal can be obtained by using time-domain non-uniform re-sampling approaches as shown in Figure 1.2.

An approach for the detection and correction of wow degradation that assumes no prior of the pitch variation curve was proposed in [8]. It is assumed that a pitch variation vector is drawn from some random process which has prior distribution. A pre-processing stage that transforms the raw data into a time-frequency 'map' by using the Short time Discrete Fourier Trans-

form (STDFT) and then tracks the principle frequency components in the data ('frequency tracking') is adopted. The idea behind this is the observation that most musical signals are composed of combinations of steady tones each comprising a fundamental pitch and overtones [1]. Pitch variations that are common to all tones are assumed to be due to wow degradation, while pitch variations that are not common to all tones are assumed to be due to genuine pitch variations in the musical performance. The pitch variation curve is then estimated from these frequency tracks using the Bayesian algorithm into a single pitch variation vector. The final restoration is then approximated by non-uniform digital re-sampling.

The use of the STDFT in obtaining the spectral peaks does not take into account the a priori model of the signal. By incorporating a priori model of the underlying signal a better peak identification and tracking can be achieved.

1.3 Objective

The general objective of this research is the restoration of audio signals stored in old audio storage media. Of the types of possible degradation found in audio archives this research focuses on Clicks and 'Wow and Flutter' degradation. The selection of these degradations is based on the fact that they have the most annoying impact on the listener and is due to the vast amount of research conducted on other degradations. The use of a recently developed high-order sparse linear prediction model for the identification, characterization and restoration of different degradation on old archived audio is investigated.

1.3.1 Specific Objective

The specific objectives of the PhD research are the following.

- To review degradations affecting archived audio media, Ethiopian audio archives and to compare with state of the art preservation techniques.
- To propose a framework for the restoration of degraded audio that can work for unvoiced speech, voiced speech and music without a prior on the type of audio and degradation characteristics. The following type of degradations are considered:

  - Click degradation,
  - Wow degradation.

- To test the proposed framework on different types of audio in the presence of click and wow degradation for clean as well as noisy scenarios.

- To implement an application program that can automate the detection and restoration of these two types degradations.

1.4 Publications

As a result of this research the following research outputs were published in peer-reviewed publications containing some parts of the results.

- Peer-reviewed Journal

  - Bisrat Derebssa Dufera, "Review of early storage media degradation factors, preservation techniques and trends in Ethiopia," ZEDE, vol. 37, pp. 27 - 38, May 2019.
  - Submitted for publication
    * Bisrat Derebssa Dufera, Eneyew Adugna, Koen Eneman, and Toon van Waterschoot, "Detection and Restoration of Click Degraded Audio Based on High-Order Sparse Linear Prediction", IEEE Journal of Selected Topics in Signal Processing

- Peer-reviewed conference proceeding

  - Bisrat Derebssa Dufera, Koen Eneman, and Toon van Waterschoot, "Missing sample estimation based on high-order sparse linear prediction for audio signals," in 26th European Signal Processing Conference, EUSIPCO 2018, pp. 2464 - 2468, September 3-7, Roma, Italy, 2018.
  - Bisrat Derebssa Dufera, Eneyew Adugna, Koen Eneman, and Toon van Waterschoot, "Restoration of click degraded speech and music based on high order sparse linear prediction," in IEEE AFRICON 2019, Accra, Ghana, September 25-27, 2019.

1.5 Document Organization

The PhD dissertation document has been organized in six chapters as follows.

- **Chapter 1: Introduction**
  - Briefly discusses the problem, objective of the PhD research and the degradations to be restored.
- **Chapter 2: Degradation of Storage Media**
  - To understand the challenge of restoration of archived audio, this chapter briefly discusses the different types of audio storage media used over the years, their degradation factors, the preservation techniques and the context of audio archival in Ethiopia.
- **Chapter 3: High-Order Sparse Linear Prediction**
  - Presents the high-order sparse linear prediction model and discusses the different levels of sparsity of the coefficient vector.
- **Chapter 4: Restoration of Click Degraded Samples**
  - Discusses click degradation, the HOSpLP framework for the detection and restoration of click degraded samples and extensive results to assess the performance of the proposed framework.
- **Chapter 5: Restoration of Wow and Flutter Degraded Audio**
  - Discusses Wow and flutter degradation, the HOSpLP framework for the identification and characterization of wow degradation and extensive results to assess the performance of the proposed framework.
- **Chapter 6: Conclusion**
  - Concludes the dissertation.

Chapter 2

Degradation of Storage Media

Of the different media that have been used for audio recording, the phonograph cylinder, flat disc, magnetic tape and compact disc have the vast majority of the data. Of these, the compact disk is the most recent and stable. A recent study on accelerated aging tests of the compact disc has reported the average life expectancy of the compact disc under room temperature and controlled humidity to be 776 years [30].
[1]

2.1 Storage media types and degradation factors

Due to the different physical and chemical properties of the storage media that have been used throughout the years, their degradation factors are also different.

2.1.1 Phonograph Cylinder

Phonograph cylinders, classified as a mechanical carrier, are the earliest commercial medium for recording and reproducing sound [32]. The air pressure variation due to sound was captured by a horn that is connected to a membrane. The moving membrane is then connected via levers to a cutting

[1] Partial results of the presented work have been published in [31].

Figure 2.1: Phonograph cylinder

stylus. The stylus then engraves these movements into the surface of a rotating cylinder. The reproduction of sound worked in the opposite way. A stylus was moved by the modulated groove on the rotating cylinder, driving the membrane via levers. The vibrations of the membrane are then amplified by the horn.

Cylinders may be divided into the following types and their degradation is also affected to a different degree.

2.1.1.1 Brown-Wax Cylinders

Brown wax is a "metallic soap" combined with lesser percentages of natural waxes. This material was soft enough to make a good direct recording.

Playback of any wax cylinders at a temperature exceeding 25°C should be absolutely avoided as wear may increase significantly with temperature. In addition, rapid temperature changes may cause hairline cracks and total loss of a record [33]. Because of their highly organic composition, they are unfortunately prone to fungus attack. Humidity and temperature significantly increase the rate of fungus infection. Often, they are found in an unplayable condition as a result of this fungus damage as seen in Figure 2.2.

Figure 2.2: Degradation of brown wax cylinder

2.1.1.2 Black-Wax Cylinders

Black-Wax cylinders are made from metal soaps (lead stearate) with carbon with additional hardening agents. The major advantage of these cylinders over their predecessors is that they can be mass produced by molding process [34]. Another advantage is that "black-wax records play at a higher speed than most brown-wax records and thus, can produce a louder and better sound quality" [33]. Even though black-wax is more durable than brown-wax cylinder, it gets more brittle with age.

2.1.1.3 Celluloid Cylinders

Celluloid Cylinders were made from plaster core with celluloid plastic play-back surface layer. Celluloid records have many advantages over brown-wax and black-wax cylinder records. They can be played many more times with less wear, are less brittle and rarely get infected with fungus [35].

Loss of Camphor, a plasticizer in the celluloid, over time leads to tube shrinkage in length and diameter. This can lead to "end splits" if the celluloid is constrained by a core due to the increasing stresses [33]. Low temperature can be a high risk of breakage, as the celluloid will become significantly more brittle [35]. Another problem is that with moisture and age the core materials (plaster and cardboard) can swell preventing the cylinder from fitting all the way on the phonograph mandrel [35].

Figure 2.3: Flat disc

2.1.2 Flat Discs

A gramophone record, commonly known as a disc, is a mechanical sound storage medium in the form of a disc with a carved, modulated spiral groove. Its principle of storage is by physically plotting the waveform of the audio signal on a disc. Discs may be divided into the following types and their degradation is also affected to a different degree.

2.1.2.1 Acetate Disk

Instantaneous recordings were usually made on acetate discs until this medium was supplanted for this purpose by magnetic tape. "A lacquer coating, consisting mainly of cellulose nitrate, carries the information. The substrate or support of the discs is generally made from metal but some are of glass" [32].

Unfortunately, cellulose nitrate is an unstable material. The most im-

Figure 2.4: Delamination and Cracking of Lacquer disc.

portant degradation reactions are thermal, photo-oxidative, and hydrolytic decomposition which result in chain scission or reduction of molecular size. This results in delamination and cracking of the lacquer coating as seen in Figure 2.4. Castor oil is used as a softener to provide the physical properties needed for engraving the disk. The loss of Castor oil causes shrinkage of the lacquer coating. Because the coating is bonded to the core and cannot shrink, the result is the creation of internal stresses which result in cracking and peeling of the coating [36]. Acetate discs are also affected by mold.

2.1.2.2 Shellac Discs

Shellac disc has been used as early as 1890 and it has been used until 1950, after which it was replaced by vinyl discs. The shellac type disc is made of 70 percent or more of a mineral or cellulose aggregate which is bound together and protected by resins and waxes as fillers. The nature of the binder and fillers is important in determining the physical properties and response to aging of these discs, but these qualities are more dependent on the other constituents than on the binder [36]. As a result, determining the causes of shellac degradation is difficult. While the shellac is fungus resistant, the organic materials in the aggregates are susceptible to fungus attack.

The curing process during shellac manufacturing generates a condensation

reaction between its organic compounds. Due to this reaction the shellac shrinks, making it more dense and brittle. It also causes it to shed a fine powder after each playback.

2.1.2.3 Vinyl Discs

Vinyl discs are made of primarily Polyvinyl Chloride (PVC) and a small percentage of stabilizer, anti-static substances, "fillers", pigment, etc. [37]. The vinyl disc has proven to be the most stable. However, its life is not indefinite. According to [38], no systemic degradation of these materials is expected in the near future.

However, when PVC is exposed to ultraviolet light or to heat, it degrades chemically [34]. Compared to shellac discs, Vinyl discs are relatively soft. This makes them more susceptible to mechanical damage, such as scratches.

2.1.3 Magnetic Tapes

"From its introduction and its worldwide rise to the primary medium for sound recording, magnetic tape earned a deserved reputation as a reliable and high-quality storage medium" [1]. The recording mechanism relies on a magnetic layer, binder, that is capable of storing information by the retention of the magnetism impressed on them by the recording head. This layer is sustained over a structural support layer, base film. The magnetic layer and the structural layer are attached by a polymeric binder.

The formulation of magnetic tape is prone to degradation in storage due to several factors; such as: type of magnetic tape, acidity, humidity and temperature of the storage environment.

2.1.3.1 Magnetic Tape Binder

The audio quality, noise level, tape-to-head contact, and friction of a magnetic tape are affected by the chemical composition of the binder, binder uniformity and smoothness of application. These also affect the tape's aging properties.

Polyester polyurethane is the most common binder resin while gamma ferric oxide (Fe_3O_2) is the most commonly used ferromagnetic particle [37]. According to [6] two major binder coating failure modes that have been identified are the following.

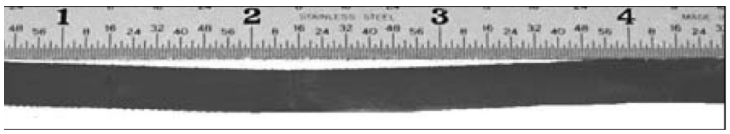


Figure 2.5: Country laning.

- Sticky shed syndrome: "The binder resin, which is made of ester, reacts with water drawn from humidity in the air to liberate carboxylic acid and alcohol. This results in the binder shedding a gummy and tacky material which causes tape layers to stick together" [39]. This degradation is sometimes temporarily treated with incubation or baking.

- Loss of lubricant: It is the failure of a tape with sticky shed syndrome to be restored to play-ability after a normal incubation or baking cycle. It has been recommended by Richard L. Hess [6] that this terminology be properly referred as "Soft Binder Syndrome (SBS)" as the cause of the failure is not really a loss of lubricant. "Playback is accompanied by squealing sound and, in some cases, the tape sticks strongly enough to the fixed surfaces that it will stop the tape transport" [6].

Hydrolysis also weakens the bond between the binder and the base, which results in shedding or possible detachment.

2.1.3.2 Magnetic Tape Base

The Base is the structural support of the magnetic tape. It provides mechanical stability against stress during playback and storage. The base film consists, in historical sequence, of acetate cellulose, PVC and polyester terephthalate, generally called polyester, which has been in use since the late 1950s [40].

The common degradation effects on the base are the following [6].

- Country laning: a tape deformation in which the tape is wavy and does not lie straight as seen in Figure 2.5. This is in most cases the result of bad slitting during manufacture or by a poor wind.

- Winding defects: usually due to sloppy winding. In this case the tape has popped strands, can cinch, be jammed against a flange or have a portion of the pack slip [41] as shown in Figure 2.6.

Figure 2.6: Winding defect.

- Edge frilling: This seems to be caused by heat damage during storage or playback or mechanical damage.

- Storage at high temperature and/or relative humidity levels may lead to fungus infection [40].

2.1.3.3 Acetate

These plasticizer additives in the Acetate base film used for suppleness evaporate and crystallize over time. This results in brittleness and drying. "Acetate tapes are very susceptible to linear expansion in humid and/or warm conditions. Due to the different properties of the binder and base, the absorption of humidity and heat result in tape curling and edge fluttering" [37].

When acetate decomposes it forms acetic acid leading to Vinegar syndrome. High temperature and high humidity levels, the presence of iron oxide, and lack of ventilation all accelerate this irreversible process [40]. This is usually observed in tapes stored in metallic containers.

2.1.3.4 Polyester

Polyester replaced cellulose acetate in the early 1960s for magnetic tape backing. Accelerated aging tests have shown polyester to be a stable material.

However, its high tensile strength make polyester-based tape prone to stretching irreparably instead of breaking cleanly. Such stretched tape cannot be repaired as easily as acetate-backed tape which breaks cleanly.

2.2 Preservation Methods

Preservation, as defined by [37], is "all actions taken to retard deterioration of, or to prevent damage to, cultural property". "Preservation involves controlling the environment and conditions of use, and may include treatment in order to maintain a cultural property, as nearly as possible, in an unchanging state" [37]. Since audiovisual recordings are stored on storage media that cannot retain the data on them permanently, digitization and continued copying of the digital files is necessary to ensure the long-time survival of the recordings [38]. The tasks that should be undertaken by archivists and libraries to preserve the media can fall into either of the following categories.

- Improvement of methods to identify and treat endangered media,
- Improvement of methods and material to decrease the rate of degradation, and
- Digital Preservation.

2.2.1 Identification of Endangered Media

There are different factors determining the life expectancy of media. Some of the factors are storage media type, age, and storage history.

Considering its longest age, it is straightforward to deduce that the wax-cylinder is the most endangered medium of the three media formats. The reason for this can be one of the following.

- Fast degradation with age,
- Instability of the material used and
- Format obsolescence.

It is reasonable to expect any older recording to be in a higher risk of degradation and information loss than a recent one. Even though this assumption holds in most cases, other factors should also be considered. For

magnetic tape, age by itself is not considered a major concern by archivists and preservation engineers [40]. Those that experience problems are usually plagued by issues unrelated to the aging process itself.

Considering the significant impact of humidity and temperature on the degradation of almost all storage media, it is reasonable to expect a recording stored in an environment with reliable climate control to be in a better condition than a recording that is stored in uncontrolled environment. As such, the presence of the following factors in the storage history are important identifiers of a recording in risk.

- Humidity,
- Temperature,
- Magnetic fields,
- Light and
- Dust.

2.2.2 Methods to Decrease the Rate of Degradation

In order to decrease the rate of degradation, the storage environment should be controlled to avoid degradation risks. Audio collections should be stored in locations with the following conditions.

- Reliable climate control: Audio recordings should be kept at a consistently low temperature and humidity. Magnetic tape should not be stored below 8°C, and "no audio format should be stored at or below freezing temperatures" [34]. For long-term storage it is recommended between 8°C-12°C and 25% - 35% relative humidity,
- Good ventilation: to filter out any dust and foreign matter that may accumulate,
- Sufficient floor loading capacity,
- Fire suppression systems and
- Security.

Recordings should also be protected against damage from

- Light: Materials made of PVC are susceptible to degradation in the presence of light, especially UV light.
- Magnetic fields: Magnetic fields may disrupt magnetic audio recordings.

2.2.3 Digital Audio Preservation

Any storage medium used for sound recording will deteriorate in time. The process of preservation using analog video tape was through migration. In the context of digital preservation, "the production of high-quality digital audio files means that subsequent copies produced in digital migrations will be bit-for-bit identical with their predecessors" [42]. This makes digital migration retain the quality of the original digital recording. But there are risks associated with the digital environment: loss of bits in storage or in digital migration. In order to mitigate these risks good data-management practices should be followed.

Preserving digital audio recordings requires the active maintenance of the recordings. The Technical Committee of the International International Association of Sound and Audiovisual Archives (IASA) has published two guides to audio preservation, referenced as "TC-04" [42]. These guidelines, have become the benchmark of digital audio preservation best practices [43].

- "Digital storage medium is the preferred medium of preservation.
- Digital audio files should be transparent, i.e., audibly indistinguishable from the original. Pulse Code Modulation (PCM) file format is the preferred encoding method.
- Preservation transfers must be flat, that is, without any imposed equalization of the frequency range or use of restoration techniques.
- Digital audio preservation files must be produced at high sampling and bit rates, and be uncompressed.
- Storage must be planned for the long term. Long-term preservation requires the migration of digital files over time and creation of digital-repositories.

- Rich meta-data related to content, format, and other attributes of the audio files must accompany preservation files.
- Professionalism is an essential component of audio preservation. Digital preservation requires trained professional staff."

2.3 Ethiopian Context

In order to understand the current condition of the storage media in Ethiopia, EBC library and IES library were selected for assessment. The libraries were visited to assess their facility in terms of types of storage media, type of data, preservation efforts, assessment methods, and ongoing digitization efforts. A brief finding on the visit has been summarized into the following three categories.

- Types of storage media employed,
- Storage history and
- Current preservation efforts.

2.3.1 Ethiopian Broadcasting Corporation

The EBC, formerly known as Ethiopian Radio and Television Agency (ERTA), is the state run broadcasting corporation in Ethiopia. It started electronic radio and television transmission in 1950s. EBC has recorded and archived a wide range of important cultural, historic, academic and entertainment content throughout it rich history.

2.3.1.1 Types of Storage Media

The audio-video media in EBC library are stored in magnetic tape. None of the data was stored in cylinder or flat disc formats. More recent recordings are stored in either CD/DVD or flash memory.

2.3.1.2 Storage History

The storage history of the recordings in relation to the above stated risk factors is reviewed. The following were observed about the audio-video library.

- The recordings are well documented and cataloged.
- The magnetic tapes are stored in cardboard boxes and plastic boxes. No recordings were stored in a metallic container that would increase the risk of degradation.
- Analog migration has been conducted periodically. The master copy is kept in a separate location from the other copies.
- The library does not have climate control. However, considering the average temperature and humidity in Addis Ababa is 15.9^0C and 60.7% respectively, the absence of climate control doesn't pose a significant degradation risk. However, for continued storage of these recordings, the usage of climate control should be considered.
- The library does not have air filtering. Considering the relative closeness of the EBC building to a main road, fumes and dust particles from vehicles may be a degradation risk.

2.3.1.3 Current Preservation Techniques Used

For preservation, continued usage and to aid in easy access, EBC is digitizing all the media in its libraries. There are 35 thousand hours of analog video/TV data and 82 years worth of audio/radio broadcast data at the time of the assessment. The radio and video digitization are carried out separately. The digitization efforts and their alignment with international guidelines on digital preservation are discussed below.

- Audio/Radio
  - Almost all the radio broadcast analog audio media data has been digitized and ingested to a central digital archive along with important meta-data.
  - Most of the analog media were in a good condition and as a result most of the digital audio files have good quality. However, there are very few recordings that could not be digitized due to significant degradation and format obsolescence. Further investigation and restoration efforts should be considered on the recordings that were too degraded to be digitized.

  - The meta-data management system is not comprehensive. The management of meta-data should be considered seriously.
  - The digital files are stored in WAV format without compression.The WAV format uses linear PCM which is in line with IASA guidelines.

- Video/TV
  - Of the 35 thousand hours of video data in the video/TV library 11 thousand hours has been digitized and ingested into the digital archive. The remaining video data is in the process of digitization.
  - As with the Audio/Radio digitization, the meta-data management in the Video/TV system is not comprehensive. The management of meta-data should be considered seriously.
  - The video is stored in MXF format using encoders to change different file formats, flv, avi, mp3, mp4 and others. Encoding and compression techniques that are used in archiving should be carefully considered as changes in technology may render the data unreadable.
  - There are some video recordings that are degraded and were difficult to read and digitize. Restoration efforts should be considered on these recordings.
  - There are some recordings that could not be digitized because of format obsolescence. More efforts in obtaining spare parts or new professional equipment that can play these recordings should be investigated.

2.3.2 Institute of Ethiopian Studies

The IES officially established in 1963 has its origins in the 1950s when it began as a repository of artifacts collected by scholars working with the Ethnological Society of the University College of Addis Ababa. Its audio-video archives include the audio-video files of Emperor Haile Selassie of Ethiopia.

2.3.2.1 Types of Storage Media

The audio-video media in IES library are stored in flat disc and different magnetic tape formats. More recent recordings are stored in either CD/DVD

or flash memory.

2.3.2.2 Storage History

The recordings are stored in audio-video libraries. The following were observed about the library.

- The recordings are not well documented and cataloged.
- The storage media are not stored systematically. The media are scattered in different rooms and the storage history in each of these rooms is not documented. This makes determination of quality of the recordings and identifying degradation risk very difficult.
- The recordings have not been copied to newer media periodically to retain their quality. As analog migration has not been conducted, the condition of the data is in a significant risk.
- The main library has air conditioning. However, not all rooms where the audio-video recordings are located have air conditioning. For continued storage of these recordings, the usage of climate control should be considered.
- There is no air filtering. The main library is cleaned frequently and there was no noticeable dust on the shelves. However, the other rooms where the audio-video recordings are located are not cleaned frequently. As a result, the recordings in these rooms are in significant risk of degradation. As a result, air filtering should be considered in the future.
- Security camera are used to deter potential damage and theft.

2.3.2.3 Current Preservation Efforts

None of the audio video files have been digitized. The media are stored in their original storage media. Considering their age and uncontrolled environment they are stored in, the audio-video media are in great risk of damage. It is anticipated that most of the recordings have some if not significant amount of degradation. The quality and extent of degradation of the media is not known at this time.

IES is in pre-digitization. It is identifying the following factors that should be investigated before digitization.

- Type of media in storage,
- Type of instrument that plays each of the media,
- Condition of media in storage and
- Meta-data information gathering.

Preservation efforts are being carried out currently. Some of the preservation efforts are the following.

- Storage condition improvement,
- New building construction to be dedicated to AV preservation.

Chapter 3

High-Order Sparse Linear Prediction

Linear prediction is a well-understood and commonly used method for the modeling, coding and analysis of speech signals [11]. Its success in representing speech signals is due to its alignment with the source-filter model of the speech generation process [44]. It has been shown that the vocal tract can be modeled by a slowly time-varying, low-order all-pole filter [45]. The glottal excitation for voiced sounds due to the periodic vibration of the vocal chords is modeled as impulse train whereas for unvoiced sounds it is modeled as a white noise sequence. The purpose of all-pole modeling through LP is to obtain a spectral envelope that characterizes the vocal tract. This can then be used for a multitude of applications including: speech compression, speech coding, speech recognition, speaker recognition and speech synthesis.

The LP coefficient vector $\mathbf{a}$ can be obtained from a set of observed samples $\mathbf{x}$ by the following optimization problem [16].

$$\mathbf{a} = \arg\min_{\mathbf{a}} \left|\left|\mathbf{x} - \mathbf{X}\mathbf{a}\right|\right|_l^l + \gamma \left|\left|\mathbf{a}\right|\right|_k^k \tag{3.1}$$

where,

$$\mathbf{X} = \begin{bmatrix} x(N_1 - 1) & \cdots & x(N_1 - P) \\ \vdots & \ddots & \vdots \\ x(N_2 - 1) & \cdots & x(N_2 - P) \end{bmatrix}$$

N_1 and N_2 are the start and end indexes of the frame,
P is the prediction order,
γ is the regularization parameter;

The ℓp-norm operator $||.||_p$ is defined as

$$\big|\big|\mathbf{x}\big|\big|_p = \left(\sum_{n=N_1}^{N_2} |x(n)|^p \right)^{\frac{1}{p}} \tag{3.2}$$

In conventional LP, the $\ell 2$-norm of the residual is minimized, $l = 2$. In addition, no structure on the coefficient vector is imposed, $\gamma = 0$, except that the prediction order is set to a small value corresponding to twice the number of format frequencies to be modeled.

Even though such modeling works well for unvoiced speech where the excitation can be modeled as white noise [16], it is not a good model for music and voiced speech, where the excitation is quasi-periodic and spiky [19]. For voiced speech, the excitation is appropriately modeled as a periodic pulse train corresponding to the glottal output. As such the minimization of the $\ell 2$-norm of the residual vector puts more emphasis on the periodic spikes of the residual [11]. As a result, it trades off short-term predictor, spectral envelope, estimation accuracy against estimating the long-term predictors, harmonics [11]. One of the most recent methods proposed to mitigate this limitation is the use of sparse linear prediction (SLP), which also incorporates the sparsity of the LP coefficient vector into the optimization problem (3.1). A better decoupling between the spectral envelope and pitch harmonics has been reported by using HOSpLP [16], [20], [21].

Another limitation of conventional LP that it needs a pitch predictor to estimate long-term correlation. In [46], a SLP approach was proposed that minimizes the $\ell 2$-norm to jointly model the long-term and short-term correlations. Similar work was proposed in [13] that uses sparse AR modeling to construct a cascade of pitch filter and formant filter to eliminate impulsive disturbances from archived speech signals. In [47] the joint optimization of the formant and pitch predictors has been proposed. It poses the estimation of the formant and pitch filter as a single LP problem given the prior knowledge of the intermediate residual signal, i.e. the output of prediction error filter. It is an iterative optimization algorithm where the intermediate residual signal of a previous iteration is used to jointly estimate the formant and pitch predictor filters. These methods however require prior knowledge of the pitch period. In [48] a pseudo-multi tap pitch filter for low bit-rate speech coding was proposed, that can estimate the pitch lag and pitch prediction taps jointly. The pitch lag is chosen as that value of lag that maximizes a normalized correlation value. In the simplest case of one coefficient pitch

predictor, this normalized correlation is in fact the reflection coefficient determined from Burg formulation.

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" [11]. This makes it hard to use a generic audio signal generation model [11]. In addition, each polyphonic audio signal should be modeled using multiple source-filter models, which seems to be rather impractical [11]. 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, a signal that is not tonal in nature, finite precision arithmetic, finite-length data windowing or noise in general. "While a sum of N sinusoids can be exactly modeled using an all-pole model of order $2N$, a sum of N sinusoids plus noise should instead be modeled using an autoregressive moving-average or pole-zero model with $2N$ zeros and $2N$ poles" [11]. Therefore, such LP signal estimates are very often poor.

3.1 $\ell 1$-norm regularized HOSpLP

The motivation for using sparse linear prediction is to give less emphasis on the quasi-periodic spikes of the residual so that estimation of the short-term predictor is less affected by the quasi-periodic excitation [16]. To achieve this, the sparsity of the residual vector, i.e. by minimizing the '$\ell 0$-norm'[1], can be used instead of the $\ell 2$-norm. "Even though, the '$\ell 0$-norm' is an ideal candidate for measure of sparsity, it results in a combinatorial problem that is NP hard" [49]. To mitigate this problem the $\ell 1$-norm, $l = 1$ in (3.1), has been used as a convex relaxation of the '$\ell 0$-norm' [16].

$$\mathbf{a} = \arg\min_{\mathbf{a}} \left|\left|\mathbf{x} - \mathbf{X}\mathbf{a}\right|\right|_1 \tag{3.3}$$

Compared to the minimization of the $\ell 2$-norm, the $\ell 1$-norm minimization gives less emphasis on the spiky underlying excitation associated with voiced speech or music on the solution.

By also incorporating the sparsity of the coefficient vector, $\gamma \neq 0$, using a high-order linear predictor and taking the $\ell 1$-norm of the residual vector,

[1] The '$\ell 0$-norm' is a pseudo-norm rather than a norm.

$k = 1$, in (3.1) joint estimation of the long-term predictor and the short-term predictor can be achieved [16].

$$\mathbf{a} = \arg\min_{\mathbf{a}} \left|\left|\mathbf{x} - \mathbf{X}\mathbf{a}\right|\right|_1^1 + \gamma\left|\left|\mathbf{a}\right|\right|_1^1 \tag{3.4}$$

The justification for incorporating the sparsity of the coefficient vector in solution to (3.4) is from the observation that a cascade of short-term and long-term predictor filters results in a filter that has high-order and few non-zero coefficients [44]. It is included as a regularization term in the cost function. The purpose of the high order sparse predictor obtained by solving equation (3.4) "is to model the whole spectrum, i.e., the spectral envelope and the spectral harmonics. This can be achieved due to the ability of high-order LP to resolve closely spaced sinusoids" [50].

The equation in (3.4) is convex but not differentiable. However, it can be solved via Alternating Direction of Multipliers (ADMM) by "reformulating the problem as a basis pursuit problem" [23].

There are mainly two challenges associated with utilizing the modeling properties of the sparse high-order predictor.

- Determining an suitable value of γ to solve (3.4) and
- Factorization of the HOSpLP coefficient vector to obtain the long-term and short-term predictors.

The regularization parameter, γ, controls the trade-off between the sparsity of the residual and the sparsity LP coefficients. Different approaches for obtaining the regularization parameter have been proposed. The use of the "modified L-curve to find the value of γ as the point of maximum curvature of the curve $(||\mathbf{x} - \mathbf{X}\mathbf{a}||_1, ||\mathbf{a}||_1)$" was proposed in [51]. An adaptive update algorithm for estimating the regularization parameter based on the observation that the optimal γ is related to the pitch gain was proposed in [21].

To solve the problem of obtaining the short-term and the long-term predictors from a high order LP coefficient, $\mathbf{a}$, the first few, N_f, coefficients of the HOSpLP predictor has been used to represent the short-term predictor in [16]. After this, a simple polynomial factorization can be used to obtain the long-term predictor after selection of the number of taps in the long-term predictor, typically either $N_p = 1$ or $N_p = 3$.

It has been shown in [16] that the use of the $\ell 1$-norm HOSpLP outperforms conventional LP in sparsity of the LP coefficients, sparsity of prediction

residual and spectral envelope estimation, for voiced speech and music. With regards to the issue of stability of the obtained short-term filters, it has been shown in [16] that the percentage of unstable filters is very low (around 2%) with "mild" instability.

3.2 $\ell 0$-norm regularized HOSpLP

The prior knowledge of the structure of the coefficient vector resulting from cascade of long-term and short-term predictor filters can also be incorporated as the following optimization problem.

$$\mathbf{a} = \arg\min_{\mathbf{a}} \left|\left|\mathbf{x} - \mathbf{X}\mathbf{a}\right|\right|_2^2 \quad s.t. \quad \left|\left|\mathbf{a}\right|\right|_0 \leq Q \tag{3.5}$$

Where Q is the sum of the filter order of the long-term and short-term predictor filters.

Problem (3.5) is non-convex which means that it a may have several local minima and its convex relaxation (LASSO), (3.1) with $p = 2$ and $k = 1$ is typically solved instead [52]. Nevertheless, proximal gradient methods can solve (3.5) if a good initialization is given, e.g., the solution of LASSO [52]. In recent work, Antonello et. al [52] developed the StructuredOptimization package for Julia programing language that can solve (3.5) in a reasonable time. This package is used in this research to obtain the $\ell 0$-norm regularized HOSpLP coefficient vector.

In this optimization problem, no prior structure on the coefficient vector is set except that the high-order LP coefficient vector has a fixed maximum number of non-zero coefficients. As such, the optimization problem in (3.5) can give emphasis to the tonal components or formant filter coefficients if the frame is composed of music or speech respectively. To show this, the HOSpLP coefficient vector resulting from a solution to equation (3.5) for natural male vowel and music is shown in Fig. 3.1 and Fig. 3.2 respectively with Q set to 16. To solve (3.5) the Julia StructuredOptimization package developed by [52] is used using the solution to the LASSO problem as initialization.

Fig. 3.1 shows that for a speech vowel the coefficient vector obtained by solving (3.5) gives more emphasis to the formant filter with very few non-zero coefficients located around 90 corresponding to the pitch period of $11.25msec$. For music, Fig. 3.2 shows that the coefficient vector has few non-zero coefficients for the short-term predictor while having more non-zero

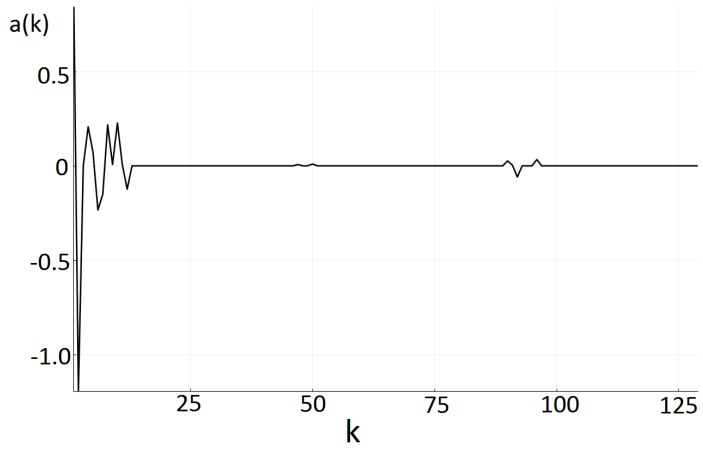


Figure 3.1: Coefficient vector for male vowel.

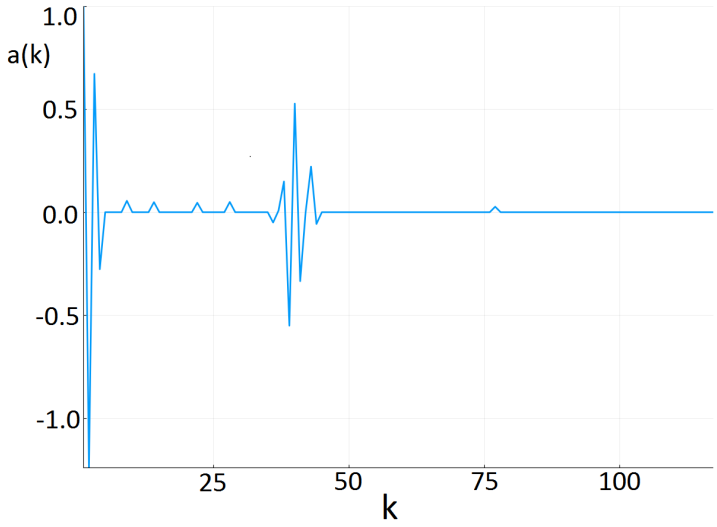


Figure 3.2: Coefficient vector for music.

coefficients for the long-term predictor distributed over the whole coefficient vector length.

To illustrate that the coefficients solved via equation (3.5) correspond to the different tones in the tonal audio, a tonal audio consisting of tones at $300Hz$, $600Hz$, $1000Hz$ and $2000Hz$ was synthetically constructed, a white noise was added to this signal and a segment of this tonal audio was artificially degraded with click degradation. The pole plot of conventional

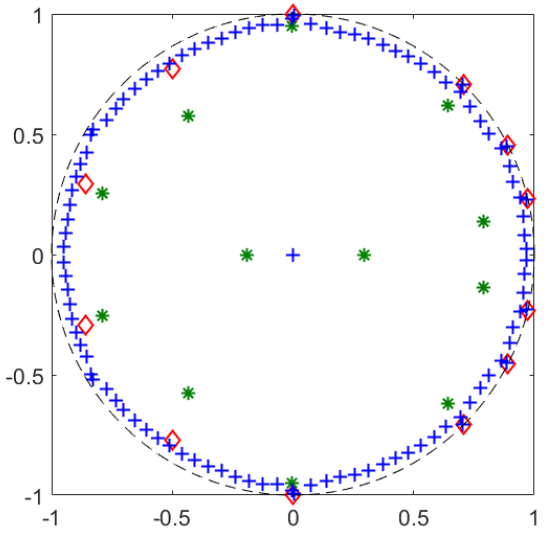


Figure 3.3: Pole-zero plot of click degraded tonal audio. $\diamond$ represent actual pole locations for the original tonal audio. $*$ represent poles of conventional LP for the click degraded noisy tonal audio. $+$ represent poles of $\ell 0$-norm regularized HOSpLP filter solution to (3.5) with 16 non-zero coefficients for the click degraded noisy tonal audio.

LP and $\ell 0$-norm regularized HOSpLP is shown in Fig. 3.3. It is seen from Fig. 3.3 that in the presence of noise and click degradation, the poles of the conventional LP drifts away from the actual poles. On the other hand, the poles of the $\ell 0$-norm regularized HOSpLP that lie on the unit circle are accurate in the presence of noise and click degradation. As such, multiple tones in the music signal are represented more accurately by using the $\ell 0$-norm regularized HOSpLP.

As the time-location of the non-zero coefficients is neither incorporated into equation (3.5) 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.

Chapter 4

Restoration of Click Degraded Samples

In this chapter the use of the HOSpLP coefficients for the detection and restoration of click degraded audio that does not require priori knowledge on the type of audio and location of click degradation is investigated. First, a restoration framework is developed that can be used with any LP based prior model of the underlying signal given the location of click degradation. Then this framework is used for the restoration of click degraded audio by using the proposed HOSpLP coefficients and two benchmarking LP models. Furthermore, the proposed HOSpLP coefficients is incorporated into a click detection methodology. To assess the performance of the proposed HOSpLP coefficients for the detection and restoration of click degraded audio, a wide range of experiments are conducted to assess the following.

- Restoration performance in clean audio,
- Restoration performance and LP coefficient properties in the presence of background noise,
- Restoration performance in the presence of noise cancellation methods,
- Restoration performance in terms of subjective quality,
- Computational complexity and
- Detection of click degraded samples.

4.1 Restoration framework

In this research, a method of restoration of click degraded audio signals that works for speech, tonal audio and music is proposed that uses high-order sparse linear prediction coefficients in the iterative Janssen algorithm for estimating the missing samples. The click degraded segment location and duration are assumed to be known a priori in this section. The problem of detecting the location of click degraded samples is investigated in section 4.4.

"Janssen et. al. [19] proposed a method that minimizes a sum of squared residual errors involving the unknown samples, the LP coefficients, and the known samples from a sufficiently large neighborhood as a function of the unknown samples and the unknown LP coefficients. It is an iterative method whereby in each iteration minimization with respect to the LP coefficients and, subsequently, minimization with respect to the unknown samples are performed. The LP coefficients of the filter are identified by minimizing the '2-norm' of the residual, the difference between the actual and predicted signal" [49]. The Janssen algorithm shown in Algorithm 1 is used as a framework for the implementation of the different high-order sparse linear prediction based restoration approaches proposed. The procedure COEFFICIENT determines the LP coefficient vector by either using the proposed $\ell 1$-norm regularized HOSpLP and $\ell 0$-norm regularized HOSpLP coefficients as well as conventional LP and Joint optimization of LP coefficients for comparison.

Algorithm 1 Framework for missing sample estimation. Modified from [53] by incorporating pitch information

procedure JANSSENINTERPOLATION_HOSPLP
  Input: $\mathbf{x}, \mathbf{v}_m, \mathbf{v}_{Obs}, P, L, \gamma, K, \epsilon$
  Output: $\hat{\mathbf{x}}$
  $\mathbf{\Theta} = \mathbf{v}_m \mathbf{1}_{1\times N} - \mathbf{1}_{M\times 1}[1, 2, \cdots, N]$;
  $\hat{\mathbf{x}}_{\mathbf{v}_{Obs}} = \mathbf{x}_{\mathbf{v}_{Obs}}$; $\hat{\mathbf{x}}_{\mathbf{v}_m} = \mathbf{0}$; $\mathbf{\Phi} = \mathbf{0}_{M\times N}$; $l = 0$;
  for $l \leq L - 1$ **do**
    $\hat{\mathbf{a}}$ =COEFFICIENT($\hat{\mathbf{x}}$,P,γ,K,ϵ);
    $\mathbf{b} = \begin{bmatrix} 1 & -\hat{\mathbf{a}}^T \end{bmatrix} \mathbf{A}$;
    $\mathbf{\Phi}_{i,j} = \mathbf{b}_{\mathbf{\Theta}_{i,j}+1}$, $\forall i, j : \mathbf{\Theta}_{i,j} > P$
    $\hat{\mathbf{x}} = -\mathbf{\Phi}^{-1}_{(1:M,\mathbf{v}_m)} \mathbf{\Phi}_{(1:M,\mathbf{v}_{Obs})} \mathbf{x}_{\mathbf{v}_{Obs}}$;
    $l \leftarrow l + 1$;
  Return

Where,

$\mathbf{x}$	is the click degraded audio signal to be restored,
$\mathbf{v}_m$	is a vector of the index of click degraded samples,
$\mathbf{v}_{Obs}$	index of undegraded samples,
N	is the length of the signal,
L	is the number of Janssen iterations,
K	is the maximum number of ADMM or joint optimization iterations,
M	is the number of missing samples,
γ	is the regularization parameter,
$\mathbf{A}$	$= \begin{bmatrix} 1 & -a_1 & -a_2 & \cdots & -a_P \\ -a_1 & -a_2 & \cdots & -a_P & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_P & 0 & 0 & \cdots & 0 \end{bmatrix}$;
P	is the LP order;
ϵ	is the residual stopping criterion for ADMM algorithm in $\ell 1$-norm HOSpLP.

In Algorithm 1, a maximum fixed number of iterations is first set, L. In each iteration, the coefficients are estimated from the restored signal of the previous iteration and in turn these coefficients are used to restore the signal. This is repeated for a fixed number of iterations. The pitch information is incorporated by using high-order LP coefficients obtained by using the following four approches for solving the LP coefficient vectors.

4.1.1 Conventional LP

This is the conventional Janssen algorithm where the coefficients in each iteration are determined by the Levinson-Durbin recursion [54]. This is used as a baseline.

4.1.2 Joint optimization of linear predictors

In this approach, the estimation of the short-term and long-term predictor filter is formulated as a single LP problem given a prior knowledge of the intermediate residual signal after the inverse formant filter [47]. Given a speech signal $\big[x(n-1), \cdots, x(n-P)\big]$ and intermediate residual signal

$\big[d(n-M), \cdots, d(n-M-N_P+1)\big]$, the windowed mean-square of the final residual is minimized resulting in the following system of equations [47] to obtain the short-term prediction filter $\big[a_1, \cdots, a_P\big]$ and long-term predictor $\big[\beta_1, \cdots, \beta_{N_P}\ \big]$.

$$\mathbf{\Phi c} = \alpha \tag{4.1}$$

Where,

$\mathbf{c}^T = \big[a_1, \cdots, a_P, \beta_1, \cdots, \beta_{N_P}\ \big]$,

$\mathbf{\Phi} = \sum_{n=0}^{N-1} \mathbf{u}^{(n)}\mathbf{u}^{(n)^T}$,

$\alpha = \sum_{n=0}^{N-1} x(n)\mathbf{u}^{(n)}$,

$\mathbf{u}^{(n)^T} = \big[x(n-1), \cdots, x(n-P), d(n-\varrho), \cdots, d(n-\varrho-N_P+1)\big]$,

ϱ is the pitch lag and

N_P is the number of pitch predictor taps.

As the short-term predictor filter coefficients are not calculated directly from the intermediate residual, the short-term predictor filter can ignore the pitch pulses in the intermediate signal [47]. The pitch lag, ϱ, is chosen as that value of lag that maximizes a normalized correlation value.

In practice, the intermediate residual signal is not available. Therefore, the problem is posed as an iterative combined approach where the intermediate residual signal of a previous iteration is used to jointly estimate the formant and pitch predictor filters [47]. For the first iteration, the pitch prediction coefficients are set to zero, the formant filter is estimated and an intermediate residual obtained. Afterwards, the intermediate residual from the previous iteration is used to jointly solve equation (4.1).

The iterative combined approach does not guarantee that the overall error decreases monotonically over the iterations. If the error increases during a particular operation, the iteration has to be stopped prematurely. It only guarantees a mean-square error that is never worse than conventional sequential solution which is limited by the quasi-periodic nature of the residual signal [47].

4.1.3 $\ell 1$-norm regularized HOSpLP

The ADMM algorithm for solving $\ell 1$-regularized linear regression problems [23] is used to obtain the HOSpLP coefficients. "It starts from the conventional LP coefficients and iteratively minimizing the sum of the $\ell 1$-norm of the estimation error and $\ell 1$-norm of the coefficients" [23].

Algorithm 2 ADMM [23]

procedure L1L1_SLP_ADMM
  Input: $\mathbf{x}$, P, γ, K, ϵ
  Output: $\mathbf{a}$
  $\mathbf{y}, \mathbf{z}, \mathbf{u} = 0_{(N+2*P)\times 1}$; $Iter = 0$; $\mathbf{a} = 0_{P\times 1}$
  $\mathbf{R_x} = autocorrelation(\mathbf{x})$;
  $\mathbf{a} = levinsondurbin(\mathbf{R_x})$;
  $\mathbf{H} = \left(\begin{bmatrix} -\gamma \mathbf{I}^T_{P\times P} & \mathbf{X}^T \end{bmatrix}^T \right)^+$
  while $\{E_n > \epsilon \;\&\&\; C_n > \epsilon \;\&\&\; Iter < K\}$ **do**
    $\mathbf{a} = \mathbf{a} - \mathbf{H}(\mathbf{y} - \mathbf{u})$;
    $\mathbf{e} = \mathbf{X}[\mathbf{1}, -\mathbf{a}]$;
    $\mathbf{z} = \begin{bmatrix} \gamma\ \mathbf{a}^T & \mathbf{e}^T \end{bmatrix}^T$;
    $\mathbf{y} = Sm(\mathbf{z} + \mathbf{u}, \rho)$;
    $\mathbf{u} = \mathbf{u} + \mathbf{z} - \mathbf{y}$;
    $E_n = ||\mathbf{e}||_1$; $C_n = ||\mathbf{a}||_1$;
    $Iter = Iter + 1$;
  $\mathbf{a} = \mathbf{y}_{1:P}/\gamma$;
  Return

Where:

$\mathbf{x}$	is the click degraded signal to be restored;
M	is the number of missing samples;
$N = N_2 - N_1$	is the number of samples in each frame;
$\mathbf{v}_{Obs}$	is a vector index of known samples;
$\mathbf{v}_m$	is a vector index of click degraded samples;
$\mathbf{X}$	$= \begin{bmatrix} x(0) & 0 & \cdots & 0 \\ x(1) & x(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x(N-1) & x(N-2) & \cdots & x(N-P) \end{bmatrix}$,
K	is the maximum number of ADMM iterations
$()^+$	is the Moore-penrose pseudo-inverse and
$Sm()$	is a soft thresholding operator and
ρ	is augmented Lagrangian parameter.

When the ADMM algorithm is used to estimate the LP coefficients in Algorithm 1, it leads to a computationally expensive iteration as a result of the ADMM iteration at step 7 of Algorithm 1. This ADMM iteration inside the Janssen iteration makes the proposed algorithm computational complex. This computational complexity can be decreased by merging the two iterations into one. "This can be achieved by re-estimating the missing samples inside the ADMM iteration. The ADMM algorithm iteratively minimizes the sum of $\ell 1$-norm of the coefficients and estimation error by starting from the conventional LP coefficients as shown in Algorithm 2. In this algorithm, the residual vector and the HOSpLP coefficients are made sparser by every iteration [49]. Therefore, it seems straightforward to do the restoration at each iteration by using the most recent HOSpLP coefficients. A restoration method merging the two algorithms is shown in Algorithm 3. The number of Janssen iterations at each iteration of the ADMM algorithm is limited to one as the HOSpLP coefficients are fixed in each ADMM iteration" [49].

The computational complexity of Algorithms 1 and 3 are analyzed and shown to be as follows. In the derivations shown in [49] only multiplications were considered. Let:

Γ = cost of each iteration of the ADMM algorithm.

- **Algorithm 1:**

Algorithm 3 HOSpLP-based missing sample estimation with merging of ADMM iterations and Janssen iterations

procedure ADMM_HOSPLP_ITER
  Input: $\mathbf{x}, \gamma, K, \epsilon, \mathbf{v}_m, \mathbf{v}_{Obs}$
  Output: $\hat{\mathbf{x}}$
  $\mathbf{y}, \mathbf{z}, \mathbf{u} = 0_{(N+2*P)\times 1}$;
  $\mathbf{\Theta} = \left|\mathbf{v}_m \mathbf{1}_{1\times N} - \mathbf{1}_{M\times 1}[1, 2, \cdots N]\right|$;
  $\mathbf{\Phi} = \mathbf{0}_{M\times N}$; $\hat{\mathbf{x}}_{\mathbf{v}_{Obs}} = \mathbf{x}_{\mathbf{v}_{Obs}}$; $\hat{\mathbf{x}}_{\mathbf{v}_m} = 0$; $Iter = 0$;
  $\mathbf{R}_{\mathbf{x}} = autocorrelation(\mathbf{x})$;
  $\mathbf{a} = levinsondurbin(\mathbf{R}_{\mathbf{x}})$;
  $\mathbf{H} = \left(\begin{bmatrix} -\gamma \mathbf{I}_{P\times P}^T & \mathbf{X}^T \end{bmatrix}^T \right)^+$
  while $\{E_n > \epsilon \quad \&\& \quad C_n > \epsilon \quad \&\& \quad Iter < K\}$ **do**
    $\mathbf{a} = \mathbf{a} - \mathbf{H}(\mathbf{y} - \mathbf{u})$;
    $\mathbf{e} = \mathbf{X}[\mathbf{1}, -\mathbf{a}]$;
    $\mathbf{z} = \begin{bmatrix} \gamma \mathbf{a}^T & \mathbf{e}^T \end{bmatrix}^T$;
    $\mathbf{y} = Sm(\mathbf{z} + \mathbf{u}, \rho)$;
    $\mathbf{u} = \mathbf{u} + \mathbf{z} - \mathbf{y}$;
    $E_n = ||\mathbf{e}||_1$; $C_n = ||\mathbf{a}||_1$;
    $\mathbf{b} = \begin{bmatrix} 1 & -\mathbf{a}^T \end{bmatrix} \mathbf{A}$;
    $\forall i, j : \mathbf{\Theta}_{i,j} > P \ \mathbf{\Phi}_{i,j} = \mathbf{b}_{\mathbf{\Theta}_{i,j}+1}$;
    $\hat{\mathbf{x}} = -\mathbf{\Phi}_{(1:M,\mathbf{v}_m)}^{-1} \mathbf{\Phi}_{(1:M,\mathbf{v}_{Obs})} \mathbf{x}_{\mathbf{v}_{Obs}}$;
    $Iter = Iter + 1$;
  Return

$$L\frac{M^3}{3} + L\frac{N^2}{2} + LK\Gamma \tag{4.2}$$

- **Algorithm 3:**

$$K\frac{M^3}{3} + K\frac{N^2}{2} + K\Gamma \tag{4.3}$$

Equations (4.2) and (4.3) cannot be compared directly due to the different number of ADMM iterations in each algorithm. After conducting experiments to estimate the different variables in the equations, the following observations have been made in [49]

- $K \gg L$
- "The cost of the ADMM algorithm, the most computationally expensive part, is decreased by a factor of L in Algorithm 3 as compared to Algorithm 1".

Even though computational time saving is achieved by using Algorithm 3, it was observed by simulation on different types of audio that the restoration performance is not as good as Algorithm 1. As the application at hand is the restoration of archived media, computational time is not our priority. Therefore, Algorithm 1 was used in the restoration steps although Algorithm 3 could be used to get a faster restoration at the expense of restoration quality.

4.1.4 $\ell 0$-norm regularized HOSpLP

In this approach (3.5) is solved via the StructuredOptimization Julia package [52] to obtain the $\ell 0$-norm regularized HOSpLP coefficients. The implementation is shown in Algorithm 4. Where the @minimize function minimizes the $\ell 0$-norm by starting from an initial solution. These coefficients are then used for restoration using Janssen algorithm.

In line 7 of Algorithm 4, the $\ell 2$-norm of the residual is solved under the constraint of the number of non-zero coefficients of $\mathbf{a}$ is less than or equal to a fixed number set a prior, Q. The julia package solves this by using proximal gradient methods if a good initialization is given, e.g., the solution of LASSO [52]. Solutions of cardinality problems are inherently time consuming and computational time consumption should be investigated for practical use of this algorithm.

Algorithm 4 $\ell 0$-norm regularized HOSpLP

1: **procedure** L0_NORM_REGULARIZED_COEFFICIENT
2: **Input:** $\mathbf{x}$, P, Q
3: **Output:** $\mathbf{a}$
4: $\mathbf{X} = \begin{bmatrix} x(0) & 0 & \cdots & 0 \\ x(1) & x(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x(N-1) & x(N-2) & \cdots & x(N-Q) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x(N-1) \end{bmatrix}$
5: $\mathbf{y} = [\mathbf{x}, 0_{1\times P}]^T$;
6: $\mathbf{a} = \mathbf{0}_{P\times 1}$;
7: $\underset{\mathbf{a}}{\text{minimize}} \, ||\mathbf{Xa} - \mathbf{y}||^2$ s.t. $||\mathbf{a}||_0 \leq Q$;
8: Return

4.2 Data used, Click Noise Model and Performance Measure

4.2.1 Data

To fairly assess the restoration performance of the proposed methods on any kind of audio the experiments are conducted using different datasets. However, it should be noted that each of the methods did not take any prior on the type of audio data.

- **Synthetic male and female vowels**: ten synthetic vowels synthesized using the Klatt speech synthesizer [55].
  - /b*i*t/, /b*a*t/, /b*ea*t/, /b*e*t/, /p*a*rt/, /b*oo*t/, /p*o*t/, /b*u*t/, /b*oo*k/, /p*e*rt/
- **Natural male and female vowels**: same ten vowels from human speakers from Western University of Michigan dataset [56];
- **Male and female speech**: ten male and female speech signals from Voxforge dataset [57]; and

- **Music**: ten excerpts consisting of male singing voice, female singing voice and instrument from Sparse Models, Algorithms and Learning for Large-scale data (SMALL) dataset [58].

In order to have comparable degradation among all signals, each signal is normalized so that the maximum amplitude is 1.

4.2.2 Click Degradation Model

The onset, duration and amplitude of each click degradation is usually modeled probabilistically. Different distributions for the time between impulses and for their amplitudes can be used [1], [17], [59]. The location of the click degradation was selected randomly and click degradation with durations from $0.25msec$ to $10msec$ was used to see the performance of the methods for different click degradation duration. The samples in this click duration were replaced with zero-mean Gaussian noise to obtain a click degraded signal.

4.2.3 Performance Measures

To evaluate the restoration performance of the methods Signal to Noise Ratio (SNR) and Perceptual Evaluation of Audio Quality (PEAQ) are used. The SNR is computed on the click duration of the click-degraded fragment. As the ultimate goal of any audio restoration is the improved quality perception by a human listener, perceptually motivated evaluation of the results is conducted by using PEAQ. "PEAQ is a standardized algorithm for objectively measuring perceived audio quality, developed in 1994-1998 by a joint venture of experts within Task Group 6Q of the International Telecommunication Union's Radiocommunication Sector (ITU-R)" [60]. "It predicts the basic audio quality of a signal with respect to a reference signal by modeling the psycho-acoustic principles of the human auditory system" [60]. PEAQ has been used for the assessment of click-degraded audio restoration in [14] and [15]. It has a range of 0 to -4 as described below.

- 0: Imperceptible distortion,
- -1: Perceptible but not annoying distortion,
- -2: Slightly annoying,
- -3: Annoying and

Table 4.1: Simulation Parameters

No	Description	Value
1	Sampling frequency	8 kHz
2	Frame size	256 samples
3	Conventional LP order	12
4	HOSpLP order	128
5	Number of datasets	3
6	Number of 10 second audio files for each dataset	10
7	Number of simulations for each audio file	100
8	Artificial click duration	0.25 msec - 10 msec

- -4: Very annoying.

4.3 Results

The artificially click-degraded audio and speech data were restored using the Janssen algorithm as shown in Algorithm 1 using the four LP coefficient estimation methods discussed in section 4.1.1 till 4.1.4. The SNR is computed in the click degraded sample time range and averaged over all the audio data for each click duration. The parameters used during the simulations are shown in Table 4.1.
[1]

4.3.1 Joint optimization of LP coefficients

The Janssen algorithm was used to restore the click degraded signal by using the short-term and long-term predictor obtained by using the joint optimization of LP approach. The order of the short-term predictor is an important parameter for the performance of the restoration algorithm. Fig. 4.1 show the restoration performance of the joint optimization approach for speech. It

[1]Partial results of the presented work have been published in [49] and [61].

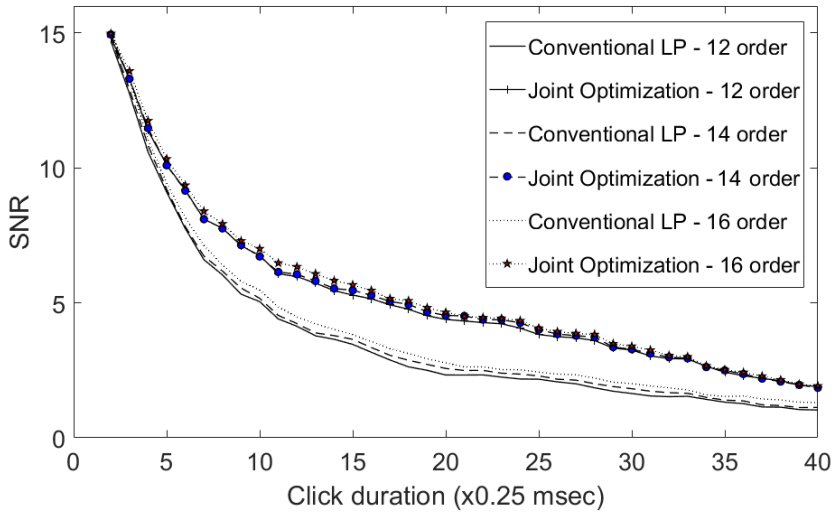


Figure 4.1: SNR of restored signal using iterative combined joint optimization of LP coefficients for speech.

is observed that the inclusion of the pitch predictor shows better SNR over conventional LP.

The restoration performance for music is shown in Figure 4.2. For this experiment conventional LP of order 12 is used as increasing the conventional LP order is not observed to lead to noticeable restoration improvement.

For music, the SNR improvement obtained by the inclusion of the jointly optimized LP predictor is not as good as that obtained for speech. This is expected as music is not well modeled with a single pitch period and significant improvement should not be expected by the inclusion of a single pitch prediction.

The order of the short-term predictor is not observed to increase the restoration SNR significantly. However, a small increase is observed. This is expected for speech as only a few formant frequencies are needed to properly characterize the vocal tract transfer function. For music on the other hand, it is seen that better restoration SNR is seen when the order of the formant predictor is increased.

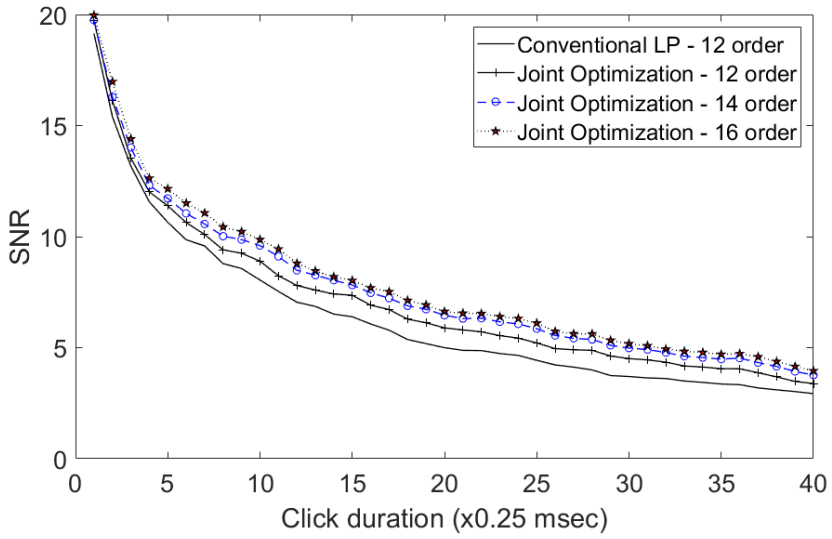


Figure 4.2: SNR of restored signal using iterative combined joint optimization of LP coefficients for music.

4.3.2 High-Order Sparse Linear Prediction

The Janssen algorithm was used to restore the click degraded signal by using the formant and pitch predictor obtained by using the $\ell 1$-norm regularized HOSpLP and $\ell 0$-norm regularized HOSpLP Algorithm 2 and 4 respectively. One of the important parameters for the $\ell 0$-norm regularized HOSpLP approach is the number of non-zero coefficients, Q. To see the effect of this parameter on the performance of the approach, different values were used and the results for music are shown on Fig. 4.3.

It is observed that for very short duration of click degradation, restoration by using the $\ell 1$-norm regularized HOSpLP achieves the highest SNR and the $\ell 0$-norm regularized HOSpLP is observed to approach the performance of the $\ell 1$-norm regularized HOSpLP as the number of non-zero coefficients, Q, is increased. Both HOSpLP approaches achieve better SNR as compared to conventional LP. However, as the duration of the click degradation increases, the restoration by using $\ell 0$-norm regularized HOSpLP shows a significant SNR improvement over both conventional LP and $\ell 1$-norm regularized HOSpLP. The restoration by using $\ell 1$-norm regularized HOSpLP approaches the SNR of conventional LP as the click duration is increased. This shows that as the click duration increases, the $\ell 1$-norm regularized HOSpLP does not

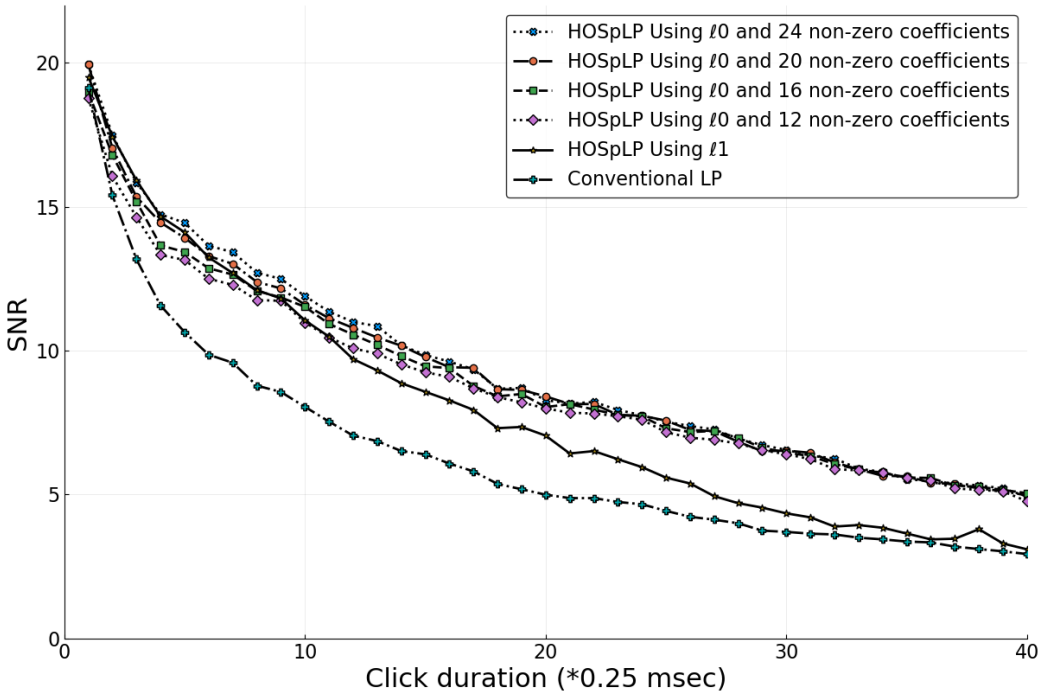


Figure 4.3: SNR of restored signal for music using $\ell 1$-norm and $\ell 0$-norm sparsity.

model the pitch information accurately.

It is also observed that as the number of non-zero coefficients, Q, is increased the restoration performance of the $\ell 0$-norm regularized HOSpLP coefficients increases. However, the improvement in restoration of using 20 and 24 non-zero coefficients is small. As such, the number of non-zero coefficients is set to 20 for further experiments.

The same result is obtained for synthetic and natural vowels as shown in Fig. 4.4 and Fig. 4.5 respectively. Similar results are observed for speech.

4.3.3 Combination of ADMM and Janssen iterations

The missing sample estimation performance of Algorithm 1 and Algorithm 3 as compared to the conventional LP-based iterative filtering for music, male speech and female speech Fig. 4.6. Algorithm 3 was implemented for different values of regularization parameter and it was observed that the regularization parameter γ has the same impact for both algorithms. In order to compare

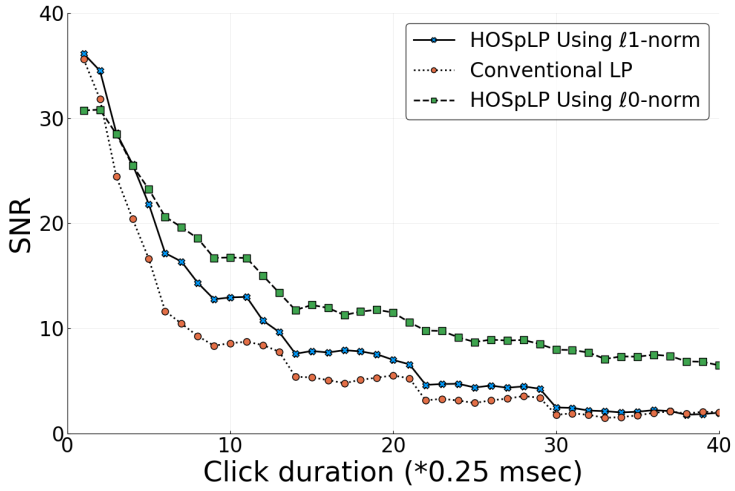


Figure 4.4: SNR of restored signal for synthetic vowels using $\ell 1$-norm and $\ell 0$-norm regularized HOSpLP.

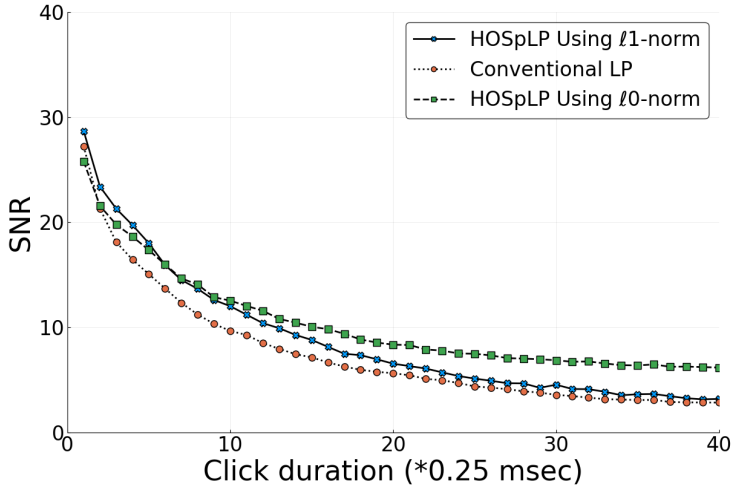


Figure 4.5: SNR of restored signal for natural vowels using $\ell 1$-norm and $\ell 0$-norm regularized HOSpLP.

the computational time taken by the two algorithms, they were implemented in MATLAB and their execution time was measured. Although the obtained result from this execution may not be representative for all scenarios, it is an indicator of a comparison between the two algorithms.

- **Number of ADMM iterations:-** When using Algorithm 3 the num-

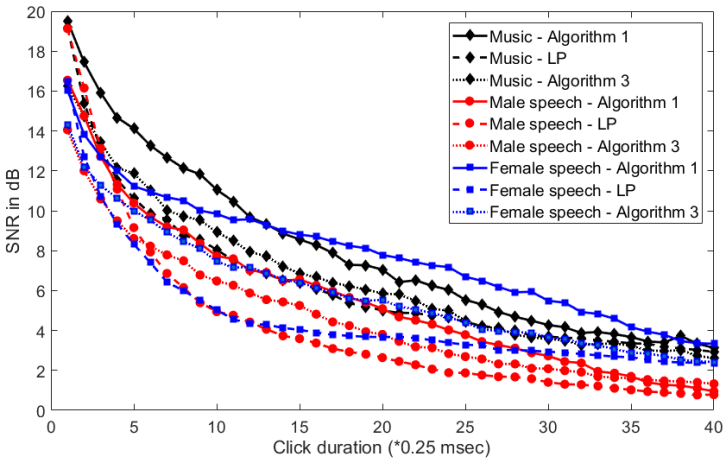


Figure 4.6: SNR of of algorithm 1 and 3 vs conventional LP.

ber of ADMM iterations, K, is observed to be much lower than Algorithm 1. This can be traced to the re-estimation of the signal in each ADMM iteration cycle.

- **Overall computational time:-** Algorithm 3 is 3.95 times faster than Algorithm 1.

Fig. 4.6 shows that Algorithm 1 achieves the best result albeit with more computational cost. On the other hand, Algorithm 3 achieves a result inferior to Algorithm 1 but better than conventional LP for the three data types.

4.3.4 Comparison of Joint-optimization of LP coefficients and high-order sparse linear prediction

A comparison of restoration of click degraded audio by the joint optimization of LP coefficients approach, $\ell 1$-norm regularized HOSpLP and $\ell 0$-norm regularized HOSpLP with 20 non-zero coefficients is shown in Fig. 4.7 and Fig. 4.8 for speech and music respectively.

Restoration by using $\ell 0$-norm regularized HOSpLP shows the best SNR over most click degradation durations except for very low click degradation duration for both speech and music. The restoration SNR of joint-optimization of LP coefficients is seen to be much lower than $\ell 0$-norm regu-

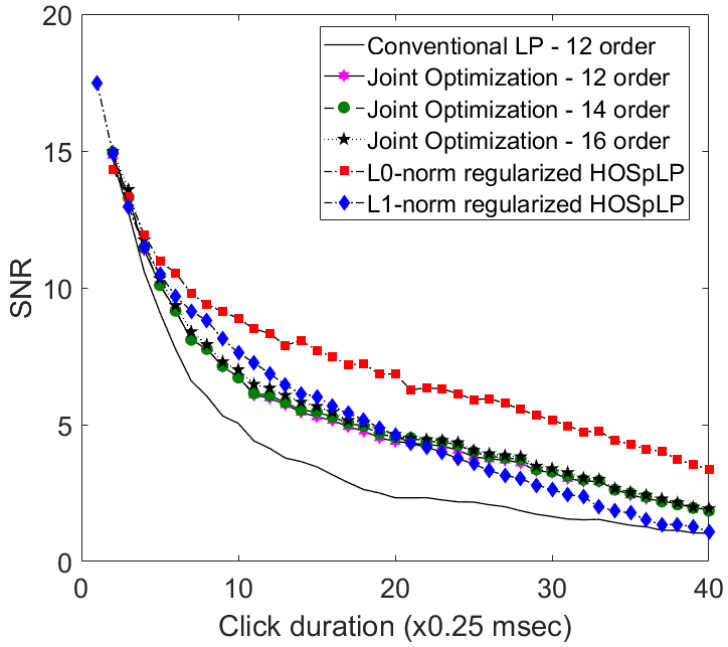


Figure 4.7: SNR of restored signal for speech using conventional LP, Jointly optimized LP, $\ell 1$-norm and $\ell 0$-norm regularized HOSpLP.

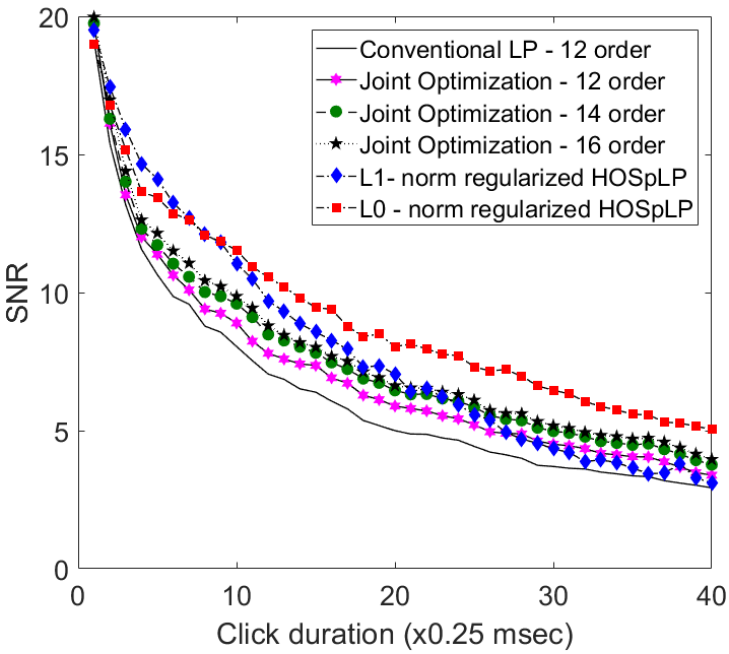


Figure 4.8: SNR of restored signal for music using conventional LP, Jointly optimized LP, $\ell 1$-norm and $\ell 0$-norm regularized HOSpLP.

larized HOSpLP even for speech. This seems to go against the Crammer-Rao Bound (CRB) for estimating a sparse parameter vector. It was shown in [62] that "the CRB bound is equal to the CRB of an estimator with knowledge of the support set, for almost all feasible parameter values. Consequently, in the unbiased case, the bound is identical to the Mean Square Error (MSE) of the oracle estimator." The decrease from expected performance of the joint optimization of LP coefficients observed in our case can be attributed to the fact that the iterative combined approach does not guarantee that the overall error decreases monotonically over the iterations. If the error increases during a particular operation, the iteration has to be stopped prematurely. The joint optimization of LP coefficients only guarantees a mean-square error that is never worse than conventional sequential solution which is limited by the quasi-periodic nature of the intermediate residual signal.

For speech, the joint optimization of LP coefficients is observed to perform better than $\ell 1$-norm regularized HOSpLP from moderate to long click duration. For music on the other hand, $\ell 1$-norm regularized HOSpLP achieves better SNR as compared to joint optimization of LP coefficients except for very long click duration.

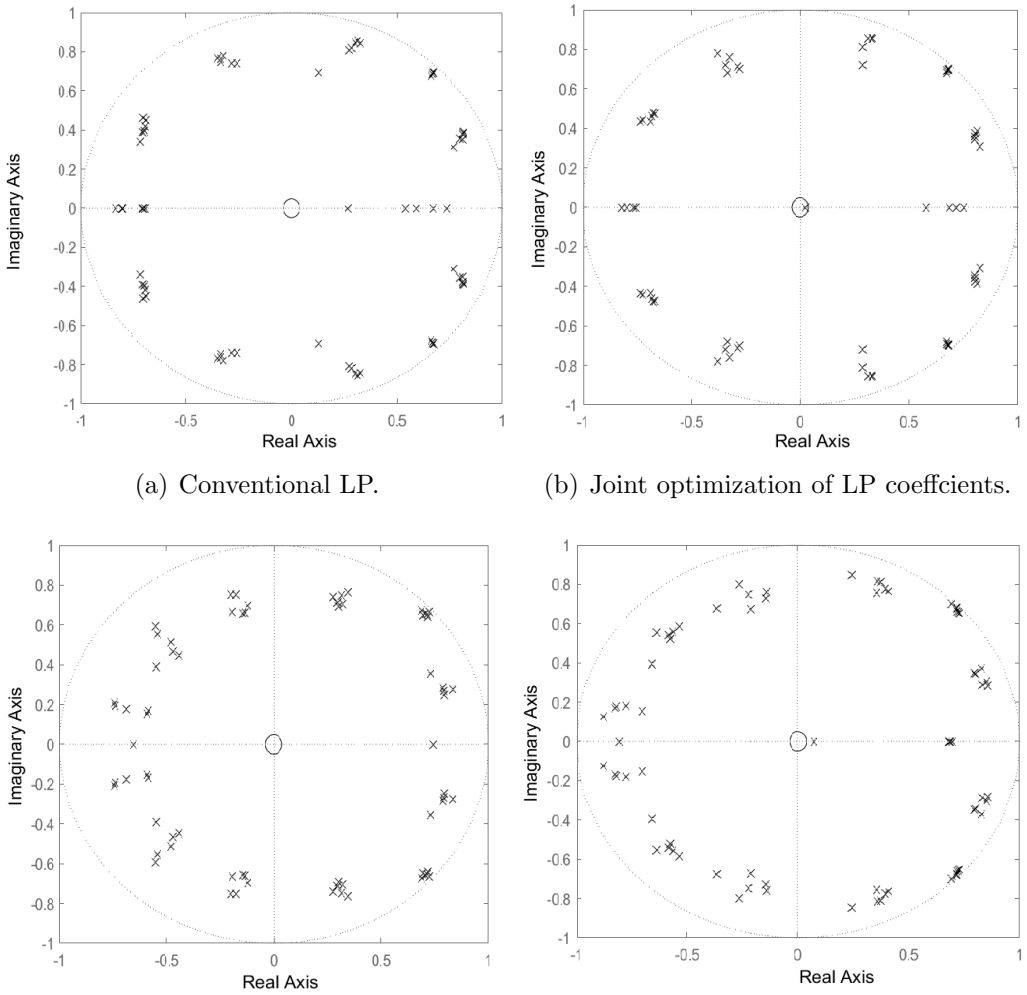
4.3.5 Noise Robustness

To assess the noise robustness of the proposed methods, additive white noise was added so that the SNR of the signal is 10dB, 20dB and 30dB. The four restoration methods were then used to remove the click degradation in the presence of background noise.

4.3.5.1 Impact of noise on the LP coefficients

To characterize the impact of background noise on the LP coefficients obtained by the four methods (conventional LP, Joint optimization LP, $\ell 1$-norm regularized HOSpLP, $\ell 0$-norm regularized HOSpLP), a plot of the location of the poles of the LP model solved by the four methods in the presence of background noise between 30dB and 5dB is shown in Figure 4.9(a) to Figure 4.9(d). From the plot of the location of poles of the LP models, the impact of noise on the LP coefficients is not clear.

Furthermore, to characterize the spread of the location of the poles from their original position due to the presence of background noise, the magnitude and phase of the poles from the origin were modeled as a Gaussian



(a) Conventional LP.

(b) Joint optimization of LP coeffcients.

(c) $\ell 1$-norm regularized HOSpLP coeffcients.

(d) $\ell 0$-norm regularized HOSpLP coeffcients.

Figure 4.9: Pole-zero plot obtained by using the four LP coefficient estimates for natural male vowel with background white noise added so that SNR is 5dB.

distribution and the variance of the magnitude and phase was plotted to see how the four LP coefficients performed in the presence of noise. The plots from Figure 4.10(a) to Figure 4.10(d) show the variance of the magnitude and phase in linear as well as magnitude scale.

From these experiments it was observed that the use of HOSpLP decreases the spread of the magnitude of the poles at a cost of more spread of the phase.

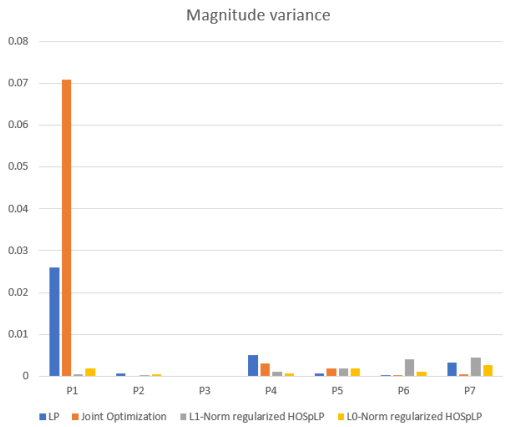
4.3.5.2 Objective quality of restoration by using the proposed LP coefficients

To assess the objective quality of the different restoration schemes proposed, the SNR was used. The four restoration methods were then used to remove the click degradation in the presence of background noise. Fig. 4.11 shows the SNR of the restored noisy signal for male speech.

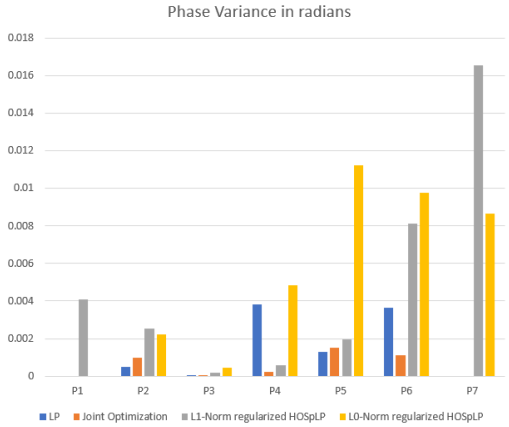
It is seen that even though the performance of all the restoration approaches decrease with the addition of noise, the degradation in performance is graceful. It is also seen that the $\ell 0$-norm regularized HOSpLP performs better for high-SNR background noise cases, while $\ell 1$-norm regularized HOSpLP method seems to perform better for the low-SNR cases (10 dB and 20 dB). This seems to indicate that the assumption that the solution to the optimization problem of $\ell 0$-norm regularized HOSpLP gives emphasis for the short-term and long-term predictors versus the tonal components depending on the type of audio may not hold in low SNR cases.

4.3.5.3 Impact of noise reduction methods on the performance of the proposed method

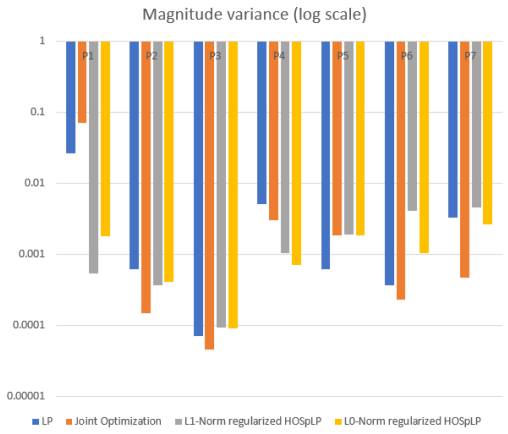
In the presence of background noise, the usual first task is to use noise reduction methods such as spectral subtraction, Wiener filtering and others to first reduce the impact of the background noise [63]. Such preprocessing steps should not affect subsequent steps in a negative way. In this experiment the impact of noise cancellation methods on the performance of the proposed HOSpLP framework for the restoration of click degraded audio is investigated. The investigation primarily focuses on finding out if such methods result in significantly lower restoration performance of the proposed HOSpLP framework. Noise cancellation methods were first used before click removal. Figure 4.12 shows the effect of power subtraction and spectral subtraction noise cancellation methods [45] on $\ell 1$-norm regularized HOSpLP and $\ell 0$-norm



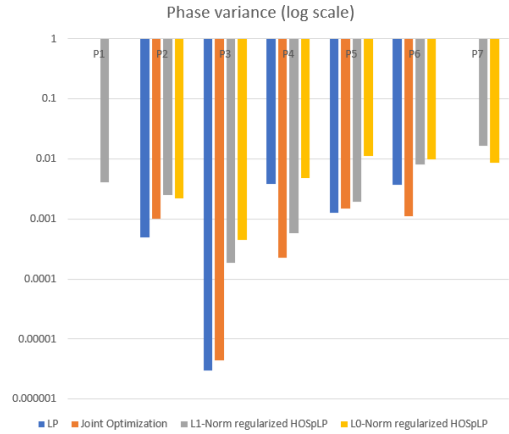
(a) Magnitude variance of the poles obtained by using the four LP methods.



(b) Phase variance in radians of the poles obtained by using the four LP methods.



(c) Magnitude variance of the poles obtained by using the four LP methods on logarithm scale.



(d) Phase variance of the poles obtained by using the four LP methods on logarithm scale.

Figure 4.10: Magnitude and Phase variance of the poles (P1, P2, P3, . . ., P7) obtained by using the four LP methods.

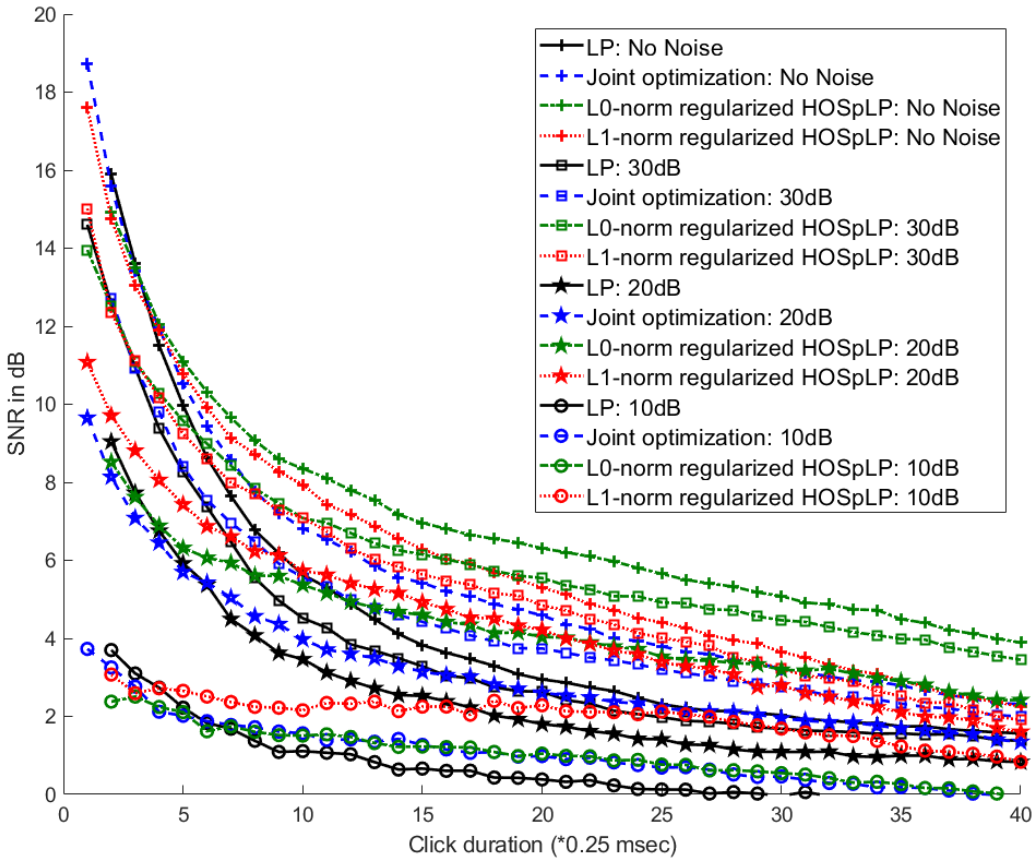
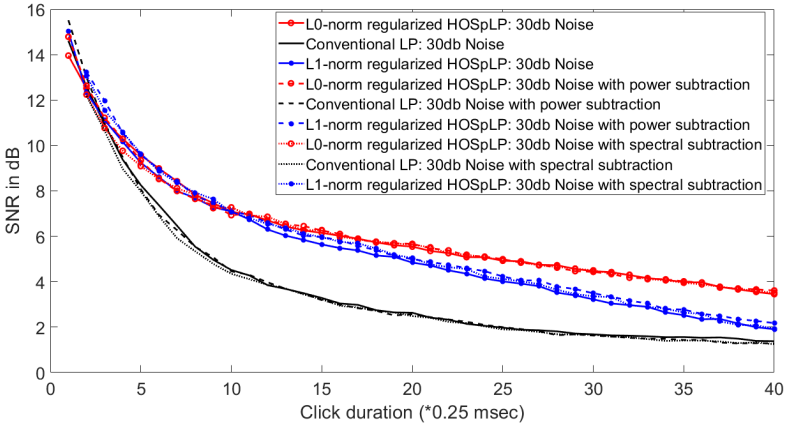


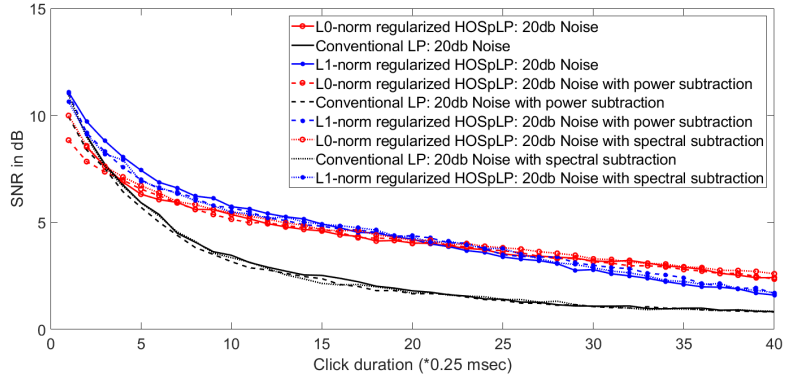
Figure 4.11: SNR of the restored signal for male speech in the presence of background noise.

regularized HOSpLP for male speech that has been degraded by background noise so that the SNR is 10dB, 20dB and 30dB. It is seen from the results that the proposed restoration methods can be used after using background noise reduction methods without significant performance degradation.

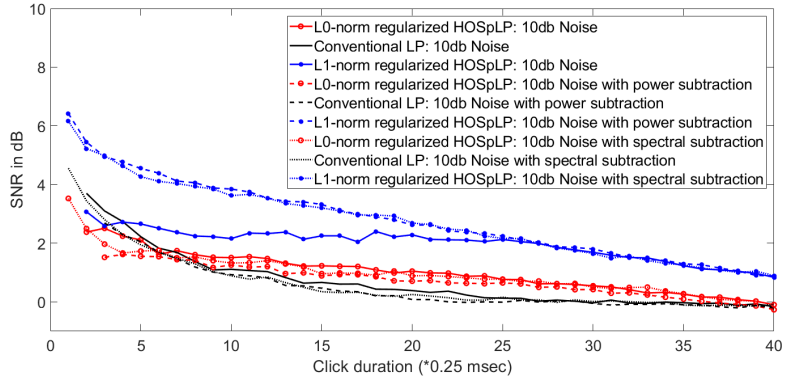
Similar experiments were conducted for female vowel, natural male speech, female speech and music by using power subtraction and spectral subtraction to reduce background noise. The results showed that the use of background noise reduction methods improve the restoration performance as compared to the signal that has been degraded by background noise. Furthermore, it is seen that the use of background noise reduction methods before restoration does not break the proposed restoration methods.



(a) Effect of noise reduction methods for background noise of SNR of 30dB



(b) Effect of noise reduction methods for background noise of SNR of 20dB



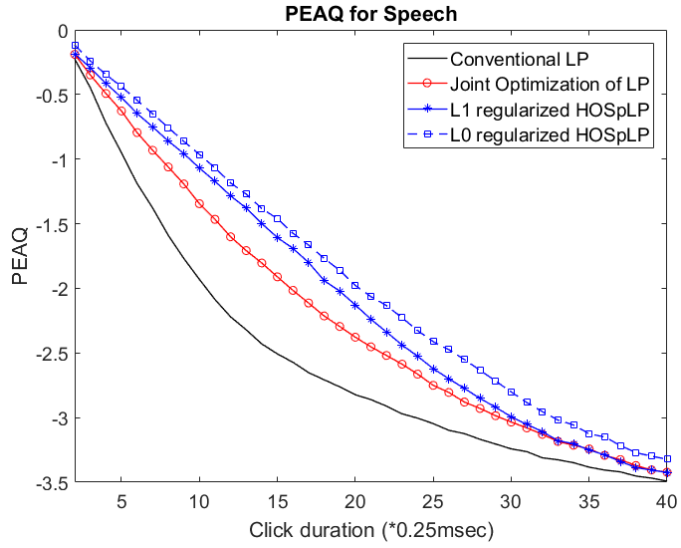
(c) Effect of noise reduction methods for background noise of SNR of 10dB

Figure 4.12: Effect of spectral and power subtraction on restoration performance.

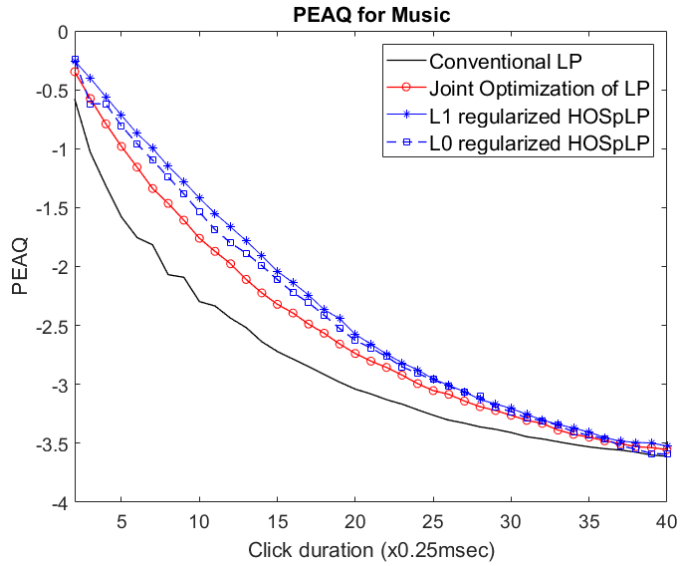
4.3.6 Perceptual evaluation of audio quality

Perceptual evaluation of audio quality (PEAQ) was used to estimate the subjective quality of the audio signal that is restored by using the proposed HOSpLP framework and the two benchmark LP models. The PEAQ was calculated for each audio fragment of speech and music as the original clean signal is available. The result of each fragment was then averaged for each type of audio. Figure 4.13(a) and 4.13(b) show the PEAQ results obtained for speech and music for each of the four approaches without background noise. It is seen that both $\ell 1$-norm regularized and $\ell 0$-norm regularized HOSpLP based restoration achieve better (i.e. higher) PEAQ as compared to conventional LP and the joint optimization approach for both speech and music. While the $\ell 0$-norm regularized HOSpLP based restoration achieves the highest PEAQ for speech, the $\ell 1$-norm regularized HOSpLP based restoration achieves the highest PEAQ for music. The improved PEAQ restoration using the $\ell 1$-norm regularized HOSpLP coefficient as compared to $\ell 0$-norm regularized HOSpLP coefficients for music is in line with the property of an audio signal. An audio signal may be composed of singing voice and multiple tones that may not be well modeled by AR filter with a fixed number of non-zero coefficients as is the case for $\ell 0$-norm regularized HOSpLP. Comparatively, the $\ell 1$-norm regularized HOSpLP gives more freedom for the optimization to allow for more number non-zero coefficients.

To assess the subjective quality of the four LP methods in the presence of background noise similar experiments were conducted. Background noise was added so that the audio signal has SNR from 30dB to 5dB. Figure 4.14 and 4.15 show that even in the presence of background noise the HOSpLP framework achieves noticeable improvement in PEAQ over both conventional LP and Joint optimization of LP coefficients. Furthermore, it is seen that for speech the $\ell 0$-norm regularized HOSpLP achieves noticeable PEAQ improvement over all the other methods over the entire range of click durations. For music on the other hand the performance of $\ell 0$-norm regularized HOSpLP is similar to the $\ell 1$-norm regularized HOSpLP even though the $\ell 1$-norm regularized HOSpLP was better for clean audio signal. This seems to be due to better noise robustness of the $\ell 0$-norm regularized HOSpLP.

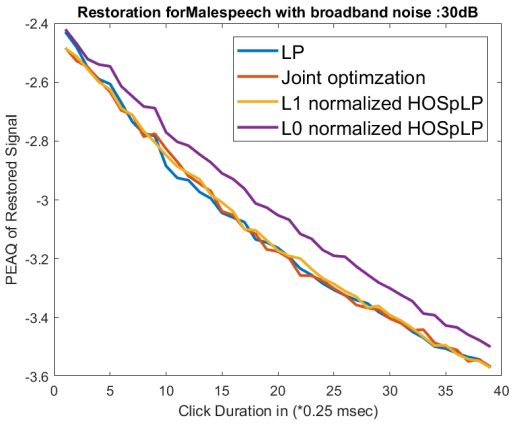


(a) PEAQ for speech.

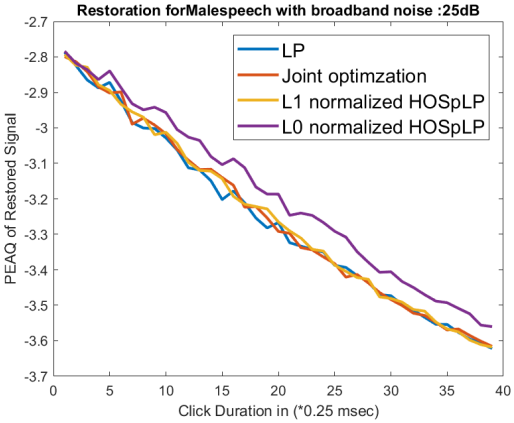


(b) PEAQ for music.

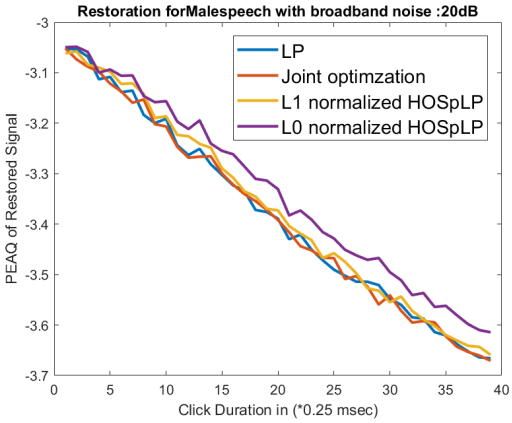
Figure 4.13: PEAQ of restored audio signal by using the four LP methods without any background noise.



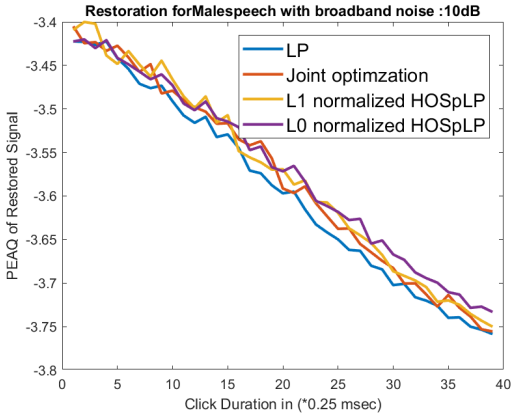
(a) PEAQ for speech for 30dB SNR background noise.



(b) PEAQ for speech for 25dB SNR background noise.



(c) PEAQ for speech for 20dB SNR background noise.

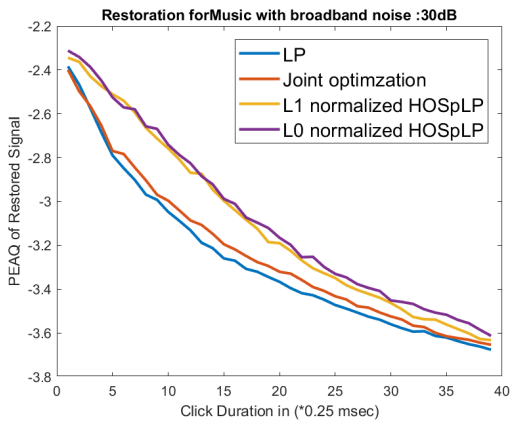


(d) PEAQ for speech for 10dB SNR background noise.

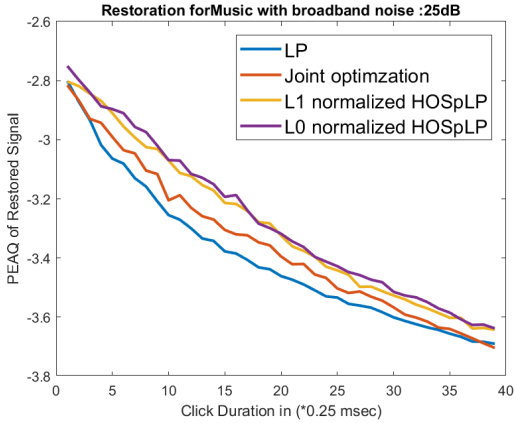
Figure 4.14: PEAQ of restored speech by using the four LP methods in the presence of background noise.

4.3.7 Computational complexity

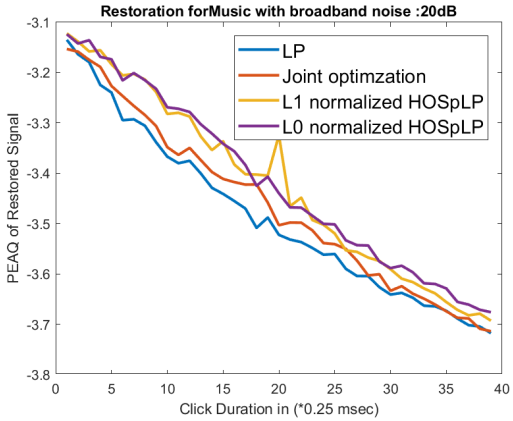
The four methods proposed except the conventional LP are iterative and it is quite difficult to analytically compare the complexity as the number of iterations is not fixed a priori. Therefore, their computational complexity is given in terms of actual time taken per given frame of length $32msec$. This



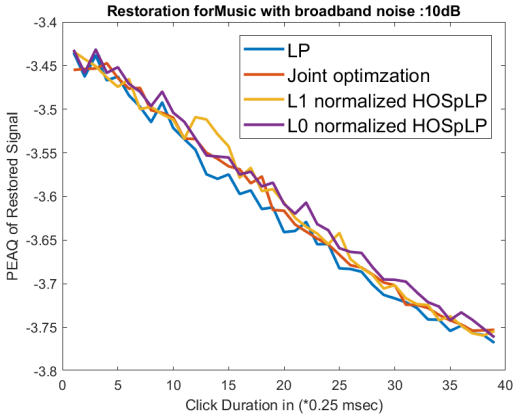
(a) PEAQ for music for 30dB SNR background noise.



(b) PEAQ for music for 25dB SNR background noise.



(c) PEAQ for music for 20dB SNR background noise.



(d) PEAQ for music for 10dB SNR background noise.

Figure 4.15: PEAQ of restored music signal by using the four LP methods in the presence of background noise.

value is averaged over all the datasets. Table 4.2 shows the actual time taken by each approach on a system with Core-i7-4510U dual core CPU, Windows 10 Professional operating system and running Julia Version 0.6.2.

The restoration by using the three LP coefficients except conventional LP is not real time. However, as the task at hand is the restoration of archived audio that has been stored for decades, the fact that they are not

Table 4.2: Computational Time Taken for a frame of length 32 msec

Method	**Time taken (in msec)**
Conventional LP coefficients	1.23
Joint optimization of LP coefficients	42.3
$\ell 1$-norm regularized HOSpLP	51.2
$\ell 0$-norm regularized HOSpLP	86.7

real time is not expected to be a significant limitation. As expected, the $\ell 0$-norm regularized HOSpLP approach takes the longest time. This is due to the inherent difficulty of cardinality problems. However, compared to the $\ell 1$-norm regularized HOSpLP and joint optimization of LP coefficients, the time taken by $\ell 0$-norm regularized HOSpLP is only about twice as long. This seems a reasonable compromise to pay given the improvement of restoration quality obtained by using the $\ell 0$-norm regularized HOSpLP approach.

4.4 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 degradation is 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 [64].

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,
$\mathbf{i}$	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 a click degradation is accurately estimated. 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 of 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.

4.4.1 Iterative detection and restoration

An iterative click detection method can be used by taking advantage of the very accurate detection of the start of the click degradation and the restoration approaches proposed in this research. The iterative detection assumes at each iteration that the click degradation is very short in duration, therefore only the starting point of the click degradation needs to be estimated. Once the start of the click degradation is estimated, a restoration is done on the assumed very short duration click degradation. After restoration, a detection is then redone on the restored click degraded signal at the next iteration. If the click degradation is longer than the assumed short duration, then the next sample points in the click degradation will be detected as the start of click for the next iteration. The approach is as follows.

1. Check if there is a click degradation in the signal by using energy based thresholding techniques that have been shown to provide accurate estimation of the start [1]. As such, such simple methods can be used in this step.
   - If click detected, go to step 2.
   - If no click is detected, go to step 4.
2. Take the estimated start of the degradation as a start point and restore the click degraded signal by assuming that the click degradation is very short (2 samples) in duration.
3. Take the restored signal as the click degraded signal and go back to step 1.
4. END as either there is no click degradation in the signal or the click degradation has been detected and removed iteratively.

The assumption made here is that by assuming the click degradation is very short in duration, the samples near the start of the click degradation are restored by using HOSpLP based restoration techniques. If the click degradation is longer than the assumed very short duration, then in the next click detection iteration the next click degraded sample will be detected as a click degradation.

Experiments were conducted to assess the performance of the proposed approach in detection of click degraded samples without a priori knowledge of the location of the click degradation. The different audio signals were artificially degraded with a click degradation having a variance twice the variance of the audio signals. The impact of the click degradation variance on the detection as well as restoration performance of the proposed HOSpLP framework is investigated in section 4.5.1.

The restoration in each iteration was done by using conventional LP and HOSpLP coefficients to see if there is any advantage in using the HOSpLP coefficients. It should be noted that the use of the HOSpLP coefficients in each iteration introduces a significant computational cost as compared to conventional LP, due to the iterative nature of the HOSpLP framework. The Normalized Mean Square Error (NMSE) error in click duration estimation, equation (4.4), is taken as a measure of the accuracy of click duration estimation.

$$NMSE = \sum_{h=1}^{H} \frac{|T_{click}(h) - \hat{T}_{click}(h)|^2}{|T_{click}(h)|^2} \tag{4.4}$$

where

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

The results in Figure 4.16 show the performance of the proposed click detection method for music. It is seen that for short click durations the detection performance is good. However, as the click duration increases, the iterative detection and restoration performance degrades quickly. For click duration of more than $4msec$, i.e. more than 32 samples for audio sampled at $8kHz$, the performance was noticed to be quite poor. It is therefore clear that a better estimation methodology is needed if the click duration is expected to be quite long. Similar results are observed for speech.

It is also noted that the use of the HOSpLP framework did not lead to noticeable improvement in click duration estimation performance. This can be attributed to the fact that at each iteration, even though the HOSpLP framework leads to better restoration SNR, the iterative approach only requires the current assumed very short click duration be restored to a degree where the residual will be lower than the threshold. For the conventional LP approach even though its restoration SNR is lower than the HOSpLP framework, its restoration performance is acceptable for the iterative approach. Therefore, the use of the HOSpLP framework does not seem relevant in the iterative detection approach. This is also expected to be the case for $\ell 0$-norm regularized HOSpLP.

It was also observed that, the procedure outlined above does not rely on clever selection of a thresholding value. Different thresholding values were tested and it was observed that the estimated click duration does not change much as long as the threshold value is large enough to discard periodic peaks in voiced periods.

4.4.2 Backward prediction

Another alternative is the use of the backward prediction error [1]. The use of the backward prediction error for the detection of clicks has been proposed

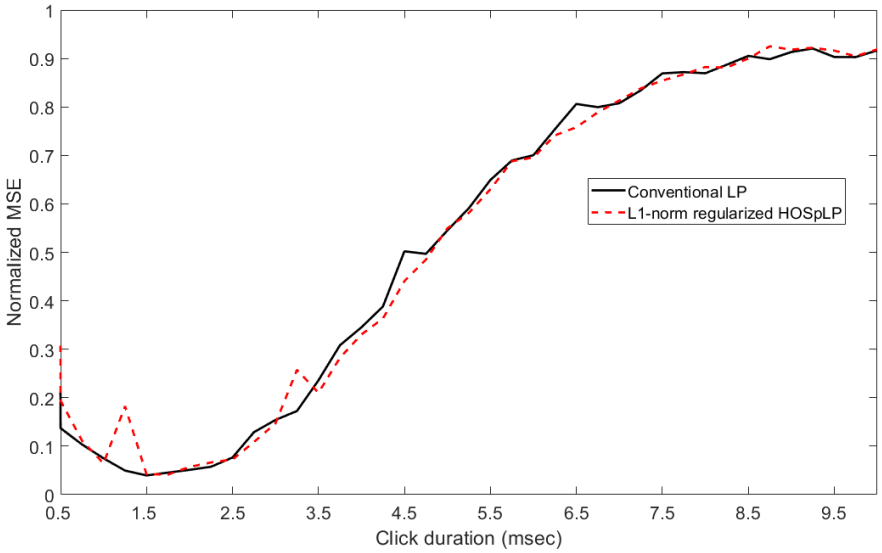


Figure 4.16: Click duration estimation using iterative detection for music.

in [1], [65]. 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

$$\epsilon_n^b = x_n - \sum_{i=1}^{P} b_i x_{n+i} \tag{4.5}$$

where, b_i, $i = 1...P$, are the backward LP coefficients.

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.

The HOSpLP framework is incorporated in click detection, see Algorithm 5, by exploiting the fact that the short-term and the long-term predictors can be jointly estimated using HOSpLP leading to a residual that is less spiky due

to the quasi-periodic excitation [16]. 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 undegraded samples between two click degradations that are close together, the backward prediction error is checked to be greater than the threshold in local a window around the detected click start.

The function COEFFICIENTS($\mathbf{x}, P, R, \gamma$) obtains the LP coefficients as follows.

- **ℓ1-norm regularized HOSpLP**: the ADMM algorithm for solving the ℓ1-norm regularized problem [23] is used to obtain the HOSpLP coefficients [49].
- **ℓ0-norm regularized HOSpLP**: the ℓ0-norm regularized problem (3.5) is solved via the StructuredOptimization Julia package to obtain the HOSpLP coefficients [52].

Algorithm 5 Backward prediction using HOSpLP model

procedure BACKWARD_PRED_HOSPLP
  Input: $\mathbf{x}, P, \gamma, R, K, N$
  Output: $\mathbf{c}$
  $\hat{\mathbf{a}}$ =COEFFICIENTS($\mathbf{x}, P, R, \gamma, \zeta$);
  for $n = 1$ to N **do**
    $\epsilon_n = \hat{\mathbf{x}}_n - \sum_{j=1}^{P} \hat{a}_j \hat{\mathbf{x}}_{n-j}$;
    σ_e =standard_deviation(ϵ_n);
    if ($|\epsilon_n| \leq K\sigma_e$), $n = n + 1$ and go to 6;
    else $\mathbf{c}_n = 1$;
    $\hat{\mathbf{b}}$ =COEFFICIENTS($\mathbf{x}^B, P, R, \gamma, \zeta$),
    $\exists r \in \{n, ..., n + k_{max}\}$:$\epsilon_r^b = x_r - \sum_{j=1}^{P} \hat{b}_j x_{r+j}$
    for $l = n$ to $n + K_{max}$ **do**
      if ($|\epsilon|_l^b \geq K\sigma_e$) $\mathbf{c}_l = 1$; $n = n + 1$; go to 12;
      for $j = l \leq l + W$ **do**
        if ($|\epsilon|_j^b \geq K\sigma_e$) $\mathbf{c}_{l:j} = 1$; $l = j$; goto 12;
      end
      if ($j \geq l + W$) $n = l$; go to 6;
    end
  end
  Return

Where:

$\mathbf{x}$	is the click degraded signal vector;
$\mathbf{x}^B$	is the time-reversed click degraded signal vector;
K	is the threshold value;
N	is the number of samples in each frame;
R	is the maximum number of ADMM iterations for $\ell 1$-norm HOSpLP;
W	is a local window size;
γ	is the regularization parameter for $\ell 1$-norm HOSpLP;
ζ	is the residual stopping criterion for the ADMM algorithm in $\ell 1$-norm HOSpLP;
$\mathbf{c}$	is a vector with 1 representing the presence of clicks and 0 representing the absence of clicks at each sample.

As a comparison, a recently proposed method by Ciołek et. al. [15] 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 a selected threshold and a click end is detected if the residual at k_0^{th} iteration is smaller than a threshold and consecutive residuals are smaller than the same threshold. In this approach, the LP coefficients are estimated by the Levinson-Durbin recursion and restoration is done by LS interpolation [24]. 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 incorporate the HOSpLP framework in this method by using the HOSpLP coefficients instead of using the conventional LP coefficients solved via the Levinson-Durbin recursion. Algorithm 6 shows a simplified algorithm to illustrate where the HOSpLP coefficients are to be used. The code for the original implementation is available in [66]. It should be mentioned that the use of HOSpLP coefficients in this method leads to a significant computational cost as it yields to a solution to an iterative problem nested in another iterative problem, i.e., the branch to line 9 from line 15 and reestimating the HOSpLP coefficients at line 12.

The function COEFFICIENTS($\hat{\mathbf{x}}, P, M, \gamma, \zeta$) obtains the LP coefficients using Levinson-Durbin in the original method [15] and using Algorithm 2 in

Algorithm 6 Iterative detection and restoration via leave-one-out interpolation [15] by incorporating HOSpLP model

procedure CIOLEK_HOSPLP
  Input: $\mathbf{x}, P, M, \gamma, K, N$
  Output: $\mathbf{y}, I$
  $\hat{\mathbf{x}} = \mathbf{x}$;
  for $n = 1$ to N **do**
    $\hat{\mathbf{a}}$ =COEFFICIENTS($\hat{\mathbf{x}}, P, M, \gamma, \zeta$);
    $\epsilon_n = \hat{\mathbf{x}}_n - \sum_{j=1}^{P} \hat{a}_j \hat{\mathbf{x}}_{n-j}$;
    if $(|\epsilon_n| \leq K\sigma_e)$ $n = n + 1$ and go to 6;
    $\mathbf{i}_n = 1$;
    $\hat{\mathbf{x}}$=Leave_One_Out_Interpolation($\hat{\mathbf{x}}$,n,$\hat{\mathbf{a}}$);
    $\hat{\mathbf{a}}$ =COEFFICIENTS($\hat{\mathbf{x}}, P, M, \gamma, \zeta$);
    $\epsilon_n = \hat{\mathbf{x}}_n - \sum_{j=1}^{P} \hat{a}_j \hat{\mathbf{x}}_{n-j}$;
    $n = n + 1$;
    if $\exists l \in \{0, ..., k_0\}: |\epsilon_{n-l}| \geq K\sigma_e$ go to 9;
  end
  Return

our proposed $\ell 1$-norm regularized HOSpLP variation of [15].

Figure 4.17 and 4.18 show the normalized MSE in click duration estimation for speech and music by using backward prediction based on conventional LP and HOSpLP and Ciołek's method. It is observed that for long click durations (longer than $4msec$), all methods lead to similar detection performance. However, as the click duration decreases, the conventional LP and $\ell 1$-norm regularized HOSpLP accuracy decreases significantly. The use of backward prediction error by using the $\ell 0$-norm regularized HOSpLP leads to the best click duration estimation results for all click durations, except for very short click durations (less than $1msec$). For music, it is seen that the $\ell 1$-norm regularized HOSpLP performs best for long click durations. It is also noted for very short click durations, that the backward prediction approach fails entirely. Ciołek's method and backward prediction based on conventional LP and $\ell 1$-norm regularized HOSpLP have similar performance.

Similar to the iterative detection approach, the backward prediction based click detection by using HOSpLP does not rely on clever selection of a thresholding value. Different thresholding values were tested and it was observed that the estimated click duration does not change much as long as the thresh-

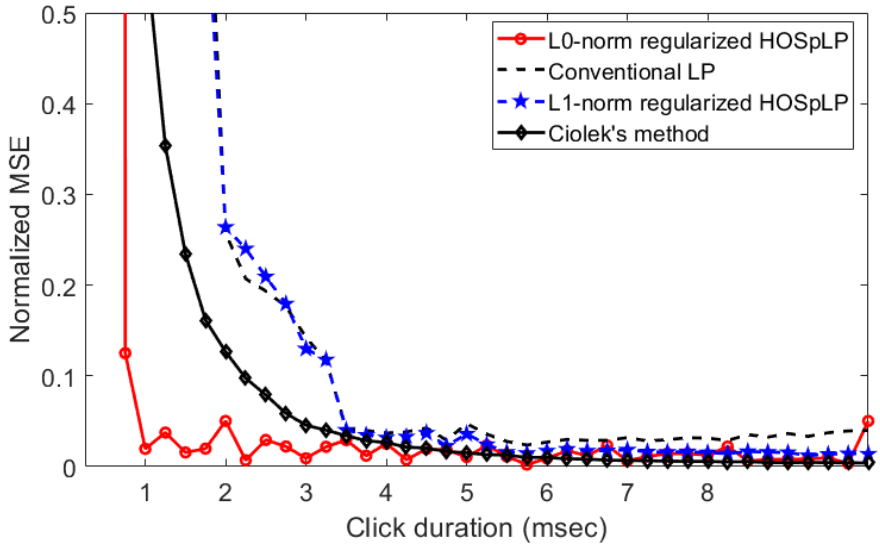


Figure 4.17: Performance of click detection by using backward prediction speech.

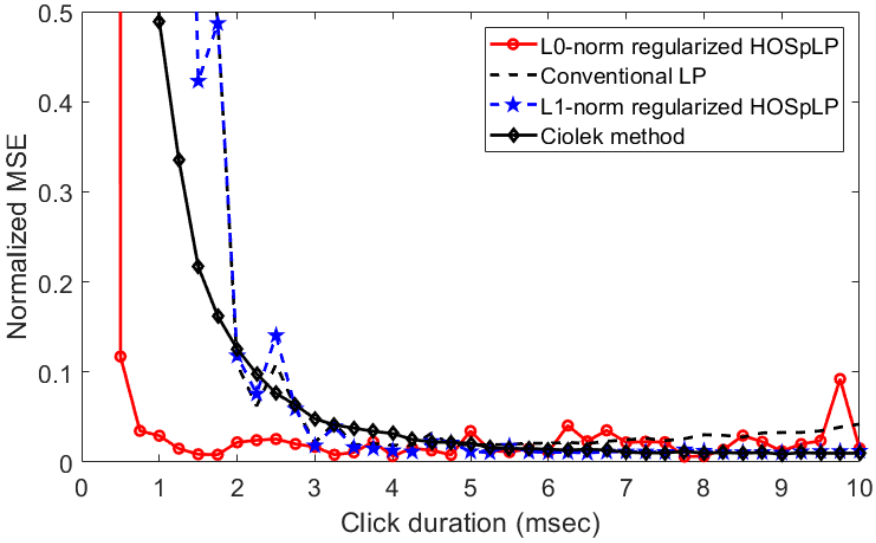


Figure 4.18: Performance of click detection by using backward prediction for music.

old value is large enough to discard periodic peaks in voiced periods. In the results shown, the threshold value was selected by using the well-known "3-sigma" rule [15]. However, it was observed that if the threshold value selected is unnecessarily large, the detection performance starts to degrade.

A comparison between the iterative detection and the backward prediction based detection is seen in Figure 4.19. It shows that the use of backward prediction by using $\ell 0$-norm regularized HOSpLP leads to the best click duration estimation results for all click durations, except for very short click durations (less than $1msec$).

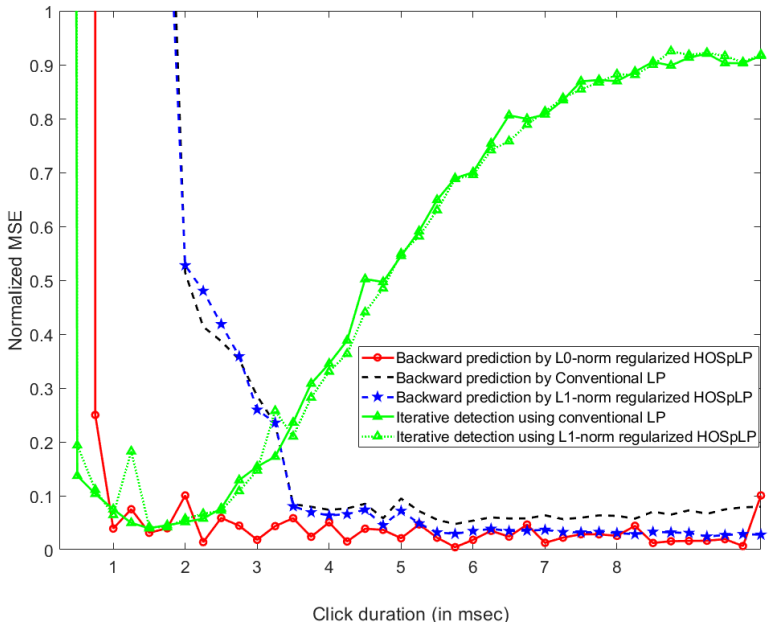


Figure 4.19: Comparison between backward prediction and iterative based detection for speech.

4.5 A unified approach for detection and restoration of click-degraded audio

In this section, a unified approach is proposed that detects the location of click degraded samples and restores these samples by using the HOSpLP

framework without a priori knowledge on the type of audio and the location and duration of click degradation. Initially, the backward prediction based on $\ell 0$-norm regularized HOSpLP is used to detect samples degraded by click degradation as this is shown to provide the best click detection performance (see section 4.4.2.) Then these samples are restored by the Janssen iteration by using $\ell 0$-norm regularized HOSpLP for the restoration as this is shown to provide the best signal restoration performance (see section 4.1.1.) The different steps of this unified approach are summarized in Algorithm 7.

Algorithm 7 Detection and Restoration using backward prediction and Janssen restoration based on $\ell 1$-norm regularized HOSpLP

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
  $\mathbf{\Theta} = \left|\mathbf{v}\mathbf{1}_{1\times N} - \mathbf{1}_{M\times 1}[1, 2, \cdots N]\right|;$
  $\hat{\mathbf{x}}_{\mathbf{u}} = \mathbf{x}_{\mathbf{u}}$; $\hat{\mathbf{x}}_{\mathbf{v}} = 0$; $\mathbf{\Phi} = \mathbf{0}_{M\times N}$; $l = 0$;
  for $l \leq L - 1$ **do**
    $\hat{\mathbf{a}}$ =COEFFICIENT($\hat{\mathbf{x}}$,P,γ,R,ζ);
    $\mathbf{b} = \begin{bmatrix}1 & -\hat{\mathbf{a}}^T\end{bmatrix}\mathbf{A}$;
    $\mathbf{\Phi}_{i,j} = \mathbf{b}_{\mathbf{\Theta}_{i,j}+1}$; $\forall i, j : \mathbf{\Theta}_{i,j} > P$
    $\hat{\mathbf{x}} = -\mathbf{\Phi}^{-1}_{(1:M,\mathbf{c})}\mathbf{\Phi}_{(1:M,\mathbf{v})}\mathbf{s}_{\mathbf{v}}$;
    $l = l + 1;$
  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.

As a comparison, the joint detection and restoration of click degraded archived audio proposed by Ciołek et. al. [15] that uses a joint evaluation of signal prediction errors and leave-one-out signal interpolation errors is used. In addition, we also propose the use of the HOSpLP coefficients in their work to obtain the LP coefficients in each iteration. The two methods are based on thresholding the energy of the residual and the threshold value in both methods is not signal dependent and does not require clever selection. In both cases, different threshold values were tested and a value of 3 led to the best result in agreement with the "3-sigma" rule [15].

To measure the unified detection and restoration performance of the two approaches, the artificially click degraded audio was restored by using the two approaches and then the SNR over the entire signal duration was computed and averaged for each dataset. No information regarding the location and duration of the click degradation was given to the two methods. Figure 4.20 and Figure 4.21 show the results of the detection and restoration for music and speech respectively.

For click durations in the range of $0.5msec$ to $4msec$, the backward prediction based HOSpLP framework is observed to perform better than Ciołek's method. It is seen that for very long click duration HOSpLP based Ciołek's method outperforms the proposed backward prediction based HOSpLP framework and conventional LP based Ciołek's method. The improved performance of the proposed backward prediction based HOSpLP for short durations can be attributed to its better click detection performance for short durations as seen in 4.19. For longer click durations, even though the detection performance of the two methods is similar, the improved restoration performance of Ciołek's method may be attributed to a better restoration approach than our proposed Janssen iteration based restoration. For click durations beyond 9 ms conventional LP based Ciołek's method offers the highest SNR, especially for speech.

Figure 4.22 and 4.23 show the SNR improvement obtained by using the backward prediction based HOSpLP coefficients, Ciołek's method with HOSpLP coefficients and Ciołek's method with conventional LP coefficients for the detection and restoration of click degraded speech and music. This is the difference between the SNR of the restored audio and SNR of click degraded audio. It is seen that both restoration methods achieve significant SNR improvement over the click degraded audio. The proposed HOSpLP based click detection and restoration approach is observed to lead to SNR improvement up to $4.5dB$ over Ciołek's method using both conventional LP

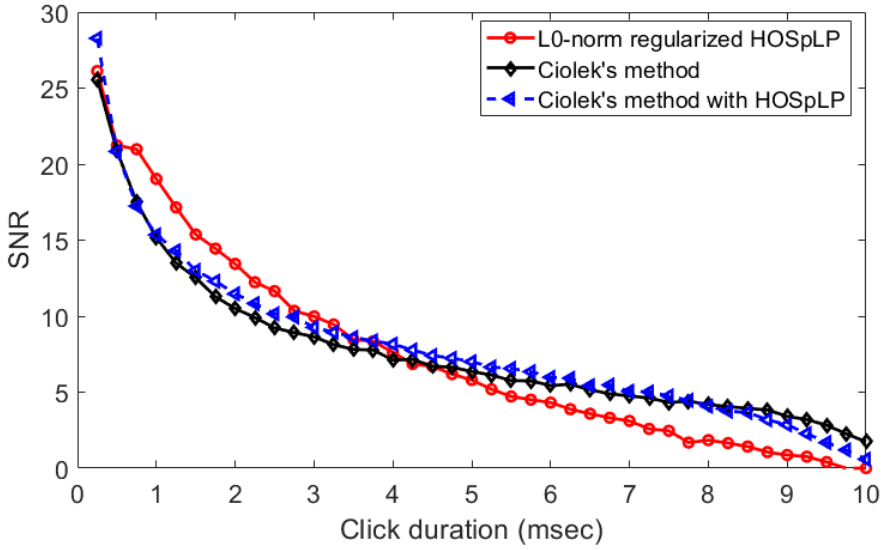


Figure 4.20: SNR of restored audio by using detection and restoration without any a priori knowledge on location and duration of click degradation for speech.

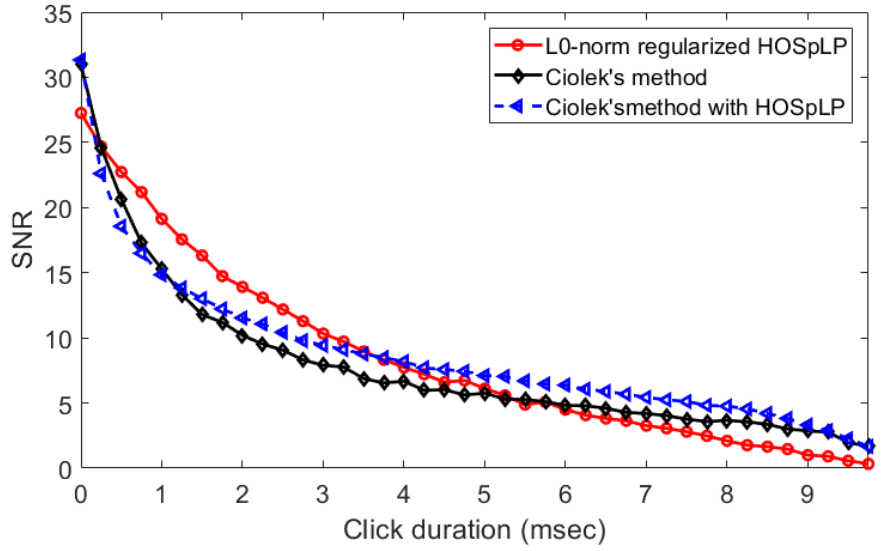


Figure 4.21: SNR of restored audio by using detection and restoration without any a priori knowledge on location and duration of click degradation for Music.

Table 4.3: PEAQ evaluation for speech

Method	**Click duration in msec**						
	0.5	1	1.5	2	3	5	8
Backward prediction with HOSpLP	-0.40	-1.09	-1.79	-2.26	-2.63	-3.07	-3.48
Ciolek's method	-0.78	-1.35	-1.10	-1.96	-2.34	-2.61	-2.84
Ciolek's method with HOSpLP	-0.41	-0.64	-0.69	-0.72	-1.19	-1.59	-1.99

Table 4.4: PEAQ evaluation for music

Method	**Click duration in msec**						
	0.5	1	1.5	2	3	5	8
Backward prediction with HOSpLP	-0.85	-0.49	-1.2	-1.45	-2.43	-3.04	-3.78
Ciolek's method	-1.2	-1.12	-1.90	-2.02	-2.1	-2.58	-2.8
Ciolek's method with HOSpLP	-0.68	-0.91	-0.39	-0.52	-1.01	-1.12	-1.45

and HOSpLP coefficients for click durations less than $4msec$. However for longer durations, Ciołek's method with HOSpLP coefficients is observed to lead to SNR improvement up to $2dB$ over the proposed backward predication based HOSpLP based click detection and restoration. The use of HOSpLP coefficients in the Ciołek's method achieves significant SNR improvement for music as compared to speech. This can be attributed to the fact that the conventional LP is not well suited to represent music.

4.5.1 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 based on HOSpLP framework 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 4.3 and 4.4 show the PEAQ evaluation obtained for speech and music respectively by using the backward prediction method with HOSpLP, Ciołek's method and Ciołek's method with HOSpLP.

It is seen that, for short click durations the backward prediction method

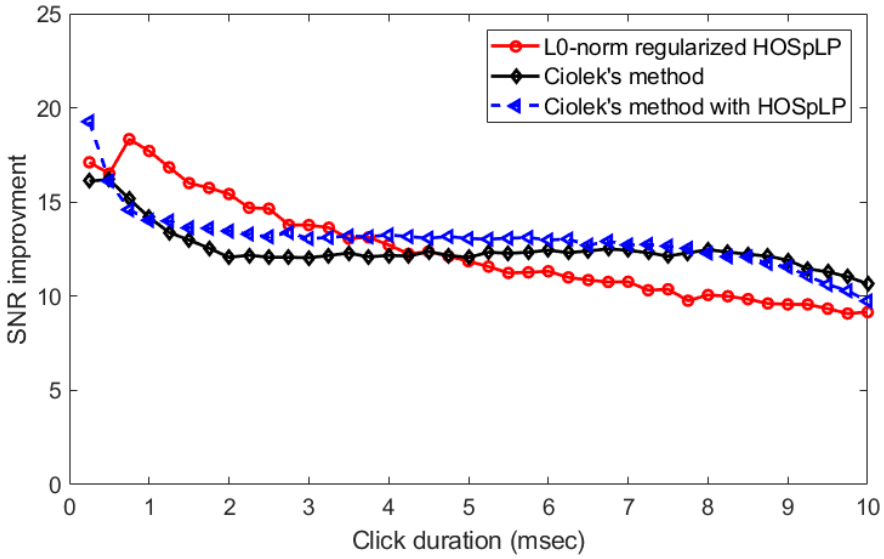


Figure 4.22: SNR improvement by detection and restoration without any a priori knowledge on location and duration of click degradation for speech.

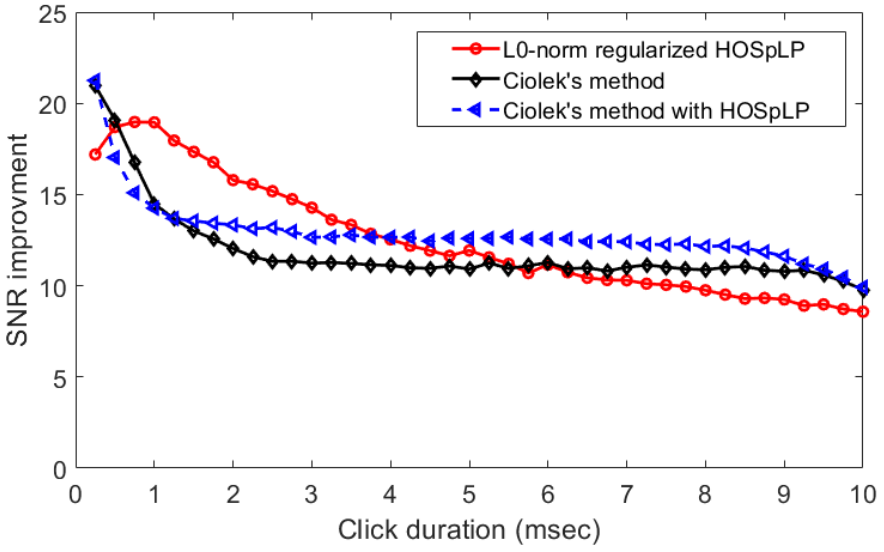


Figure 4.23: SNR improvement by detection and restoration without any a priori knowledge on location and duration of click degradation for music.

with HOSpLP leads to better PEAQ results as compared Ciołek's method. However for longer durations, Ciołek's method becomes better. Ciołek's method with HOSpLP is observed to result in the best performance overall.

4.5.2 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 framework 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 ($\sigma_c^2 = \sigma_s^2$) and quarter of the audio signal ($\sigma_c^2 = \frac{\sigma_s^2}{4}$).

Figure 4.24 and 4.25 show 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. It is seen that the three methods achieve significant SNR improvement. For click durations more than 0.5 msec, 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.

4.6 Application Program for Detection and Restoration of Click Degraded Audio

A MATLAB application program is developed for the detection and restoration of click degraded archived audio. The application program does not need information regarding the location and duration of the click degrada-

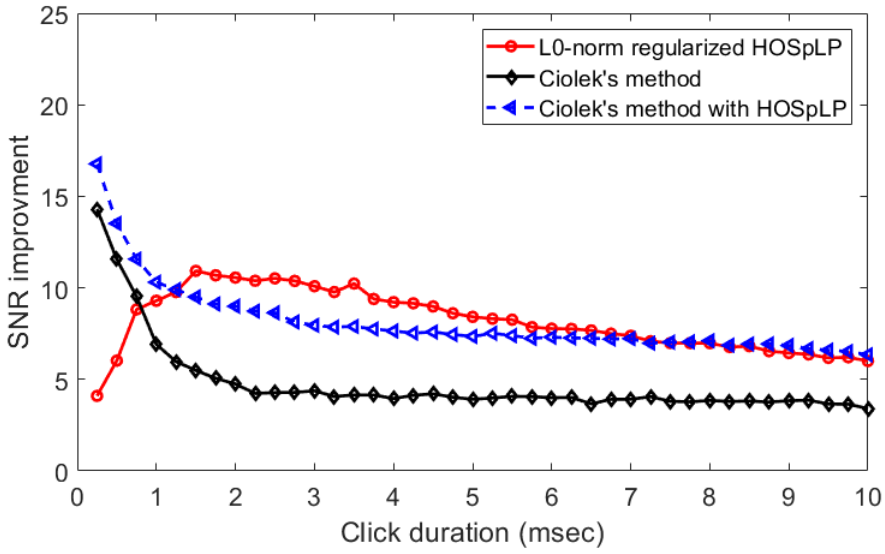


Figure 4.24: SNR improvment by detection and restoration without any a priori knowledge on location and duration of click degradation for music with click variance: $\sigma_c^2 = \sigma_s^2$

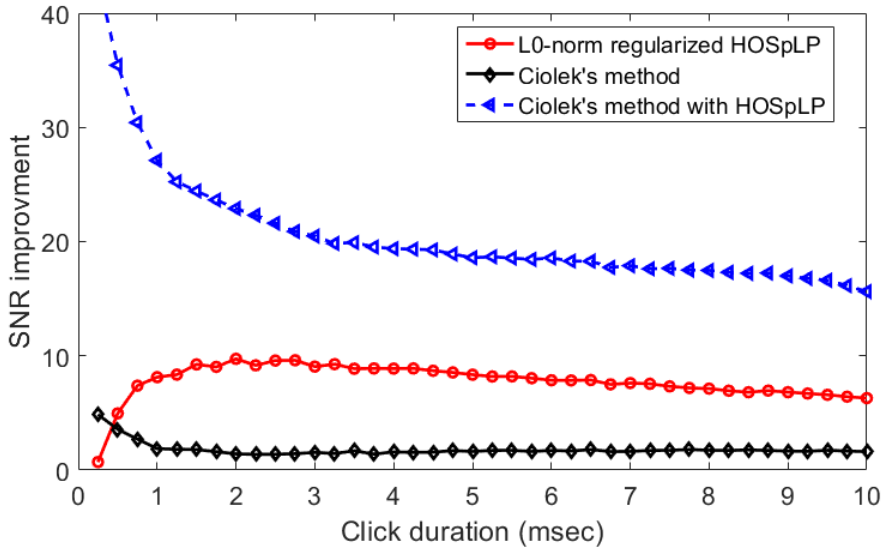


Figure 4.25: SNR improvment by detection and restoration without any a priori knowledge on location and duration of click degradation for music with click variance: $\sigma_c^2 = \frac{\sigma_s^2}{4}$

tion. Furthermore, the user does not need expert knowledge on the different algorithms used in this research to use this application program. It takes the path of the audio file to be restored, the type of degradation (click-in this case or wow), the algorithm to be used, the window size and the coefficient type. Figure 4.26 shows a screen shot of the application program.

The following four approaches are selected for the application program based on their performance in terms of best result, noise robustness and computational time consideration. The application program was tested on artificially click degraded audio signals and it was noted to perform the same as the algorithms.

- Backward prediction based on HOSpLP
- Backward prediction based on conventional LP
- Ciołek's method using HOSpLP
- Ciołek's method using conventional LP

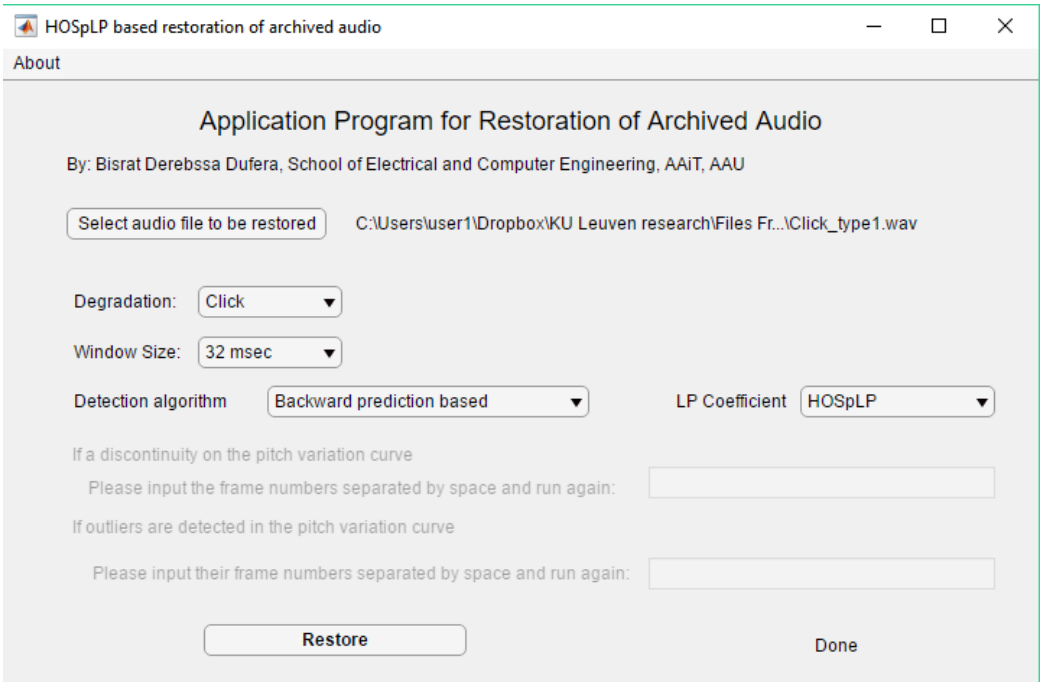


Figure 4.26: Application program for the restoration of Archived Audio.

4.7 Conclusion

In this chapter, the use of high-order sparse linear prediction is proposed for the detection of clicks and restoration of audio corrupted by click degradation that works for both speech and tonal audio without prior knowledge about the type of signal or pitch period. Several experiments were conducted to assess the performance of the proposed HOSpLP based restoration methods in terms of comparison with conventional approaches, noise robustness, perceptual audio quality and computational complexity. The proposed method achieved an improvement in SNR and PEAQ over conventional LP and joint optimization based LP coefficients for all considered speech and audio data types in clean and noisy conditions. Even though both $\ell 1$-norm and $\ell 0$-norm regularized HOSpLP based restoration methods are not real time they only take 2-3 times the duration of the frame in consideration. Considering the application at hand is for the restoration of archived audio media, this computational time is not expected to be a significant limitation.

A method for the detection of click degraded samples that incorporates HOSpLP was also proposed and demonstrated to have superior performance over conventional LP based click detection. A unified approach for the detection of click degraded samples and the restoration of these samples was also proposed and compared with a recent method that also jointly detects and restores click degraded audio. Results show that for click durations less than $10msec$, the proposed HOSpLP framework achieves superior joint detection and restoration performance.

Only artificial click degradation was considered in this experiment. Therefore, the performance of the methods should also be investigated for real click degradation. However, note that the samples in the click duration are first discarded before restoration. Therefore, the results obtained in this research are expected to hold also for real click degraded signals.

Chapter 5

Restoration of Wow Degraded Audio

Wow and Flutter refer to overall pitch variation in an audio recording which was not present in the original recording. "It is perceived as an undesired frequency modulation in the range of approximately 0.5 to 6 Hz" [26]. This variation can be due to the following [1].

- Variation of rotational speed of recording medium,
- Physical deformation of storage medium,
- Uneven stretching of magnetic tape and others.

'Wow' describes smooth variation of pitch while 'flutter' refers to a pitch variation that varies rapidly with time. In the context of this research only smooth variation of pitch, wow, is considered as flutter introduces difficulty in selecting appropriate window size where the frame can be considered to be stationary with sufficient number of samples to infer statistical information.

Preventive measures can be taken to minimize the effect of these degradations in the analog domain, on the physical storage media itself. For example, in principle it is possible to correct a poorly punched disc center hole. However, these measures can put the physical media in danger of damage and cannot completely remove the degradation [26]. The distortion can be found on wax cylinders, disks, magnetic tapes. Digital signal processing techniques provide an additional means for the detection and removal of these degradations once the analog signal is converted to digital format. Some research has been done to model and remove these degradations [27] [28], [26], [29].

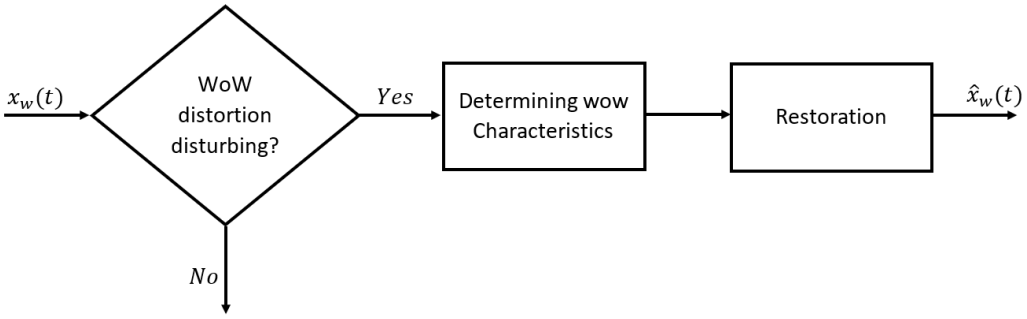


Figure 5.1: Typical wow restoration approaches.

5.1 Modeling

Wow and flutter degradation is typically modeled as a non-uniform warping of the time axis [1],

$$x_w(t) = x(f_w(t)) \tag{5.1}$$

where
$x(t)$ is the undegraded signal,
$x_w(t)$ is the degraded signal,
$f_w(t)$ is the time axis warping function.

It can be represented by the time-warping function, $f_w(t)$ or by the pitch variation function, $p_w(t)$.

$$p_w(t) = \frac{d(f_w(t))}{dt} \tag{5.2}$$

The main task of a system used for the restoration of audio signals degraded by wow degradation is therefore the estimation of the time warping function or the pitch variation function. If the time-warping function, $f_w(t)$, is known then the original signal can be obtained using time-domain non-uniform re-sampling approaches.

$$\hat{x}_w(t) = x_w(f_w^{-1}(t)) \tag{5.3}$$

Such systems generally follow the block diagram approach shown in Figure 5.1.

The decision block that determines whether a disturbing wow degradation is present or not can be done manually by an expert that listens to

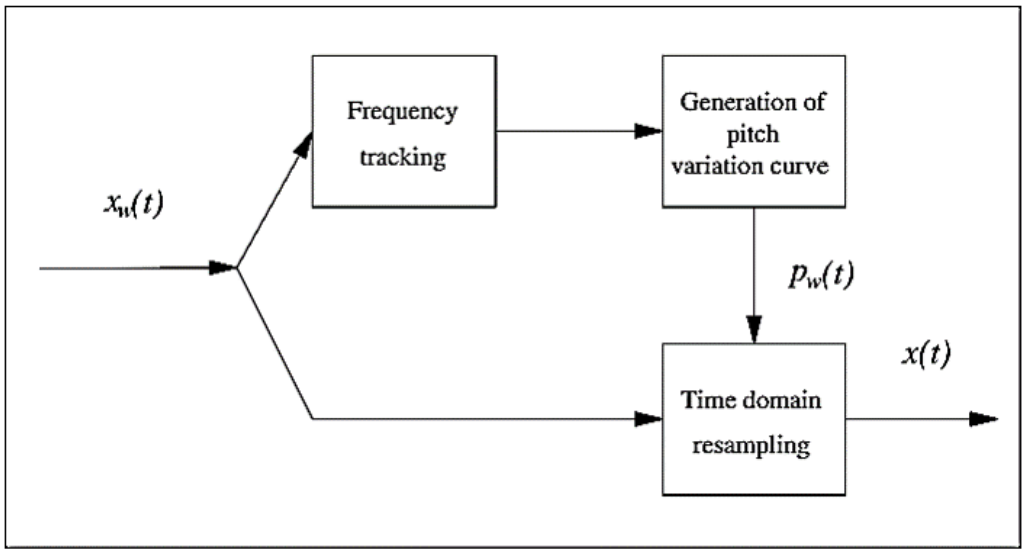


Figure 5.2: Typical wow detection and restoration approach [1].

each archived recording and determines if restoration is needed or not or a perceptual model can be used to determine whether it is disturbing or not. Alternatively, the block can be ignored altogether.

The determination of wow characteristics is conventionally done via frequency tracking algorithms. By using frequency tracking and by using probabilistic models that incorporate prior information about the overall wow degradation, the pitch variation curve is estimated as shown in Figure 5.2 [1].

The frequency tracking, generation of pitch variation curve and the incorporation of the HOSpLP framework are discussed in the following sections.

5.2 Frequency Tracking

The goal of frequency tracking is to transform the raw data into a time-frequency map that identifies the main frequency components in the data. This is based on the knowledge that an audio/speech signal consists of a few number of spectral peaks (tones or formants) that vary slowly in time. This is attributed to the property of the speech production system, vocal tract, and the property of musical instruments. Tracking these spectral peaks in time allows one to observe if there are frequency variations common to all of

the spectral peaks or only specific to one spectral peak. The assumption is that frequency variations that are common to all the spectral peaks may be attributed to wow degradation while frequency variation to a single or few spectral peaks is due to genuine frequency variation inherent to the audio signal [1].

To identify and track these spectral peaks the following approach is followed.

- **Windowing:** Raw data is divided into overlapping frames by using a proper windowing function. The length of the window is selected to be as short as possible so that the spectral peaks correspond to stationary frequency components in the frame while at the same time be as long as possible to provide a finer frequency resolution. This trade-off is usually determined by the statistical properties of the signal. It has been shown that a window length of $25msec$ to $40msec$ is sufficient for audio/speech [67]. In this research a window length of 32.5 msec is taken as we only deal with Wow degradation.

- **Spectral peak identification:** Each frame of data is then analyzed to obtain an estimate of the spectral peaks. Different approaches can be used to estimate these spectral peaks. The STDFT is commonly used to transform the signal to the frequency domain and then magnitude peak picking algorithms are used to estimate these peaks [1], [27], [28], [26]. However, it is straightforward to point out that the use of the STDFT does not incorporate any prior information we have regarding the signal generation process. As such, it is not robust to other forms of degradation. Model based signal analysis can lead to better results.

- **Tracking:** Once the spectral peaks in each frame are identified, a tracking algorithm is then used to estimate the time evolution of these peaks. If the number of spectral peaks does not change from one frame to the next, the peaks would simply be tracked by ordering the peaks in ascending frequency order. However, the number of spectral peaks changes, the location of spectral peaks changes as pitch changes and there will be rapid changes in both at speech phoneme or musical note transitions. By comparing the spectral peaks in successive frames, the following assessment can be made [68].

  - Smooth change - Frequency and magnitude of spectral peaks in one frame can change by some amount in the next frame.

- Death - A spectral peak in one frame may cease to exit in the next frame.
- Birth - A new spectral peak may exist in a frame that was not present in the previous frame.

A method for tracking these spectral peaks has been used successfully in sinusoidal coding of speech and audio [68].

5.3 Generation of Pitch Variation Curve

The pitch variation curve is estimated from the tracked spectral peaks. The change in frequency, magnitude, birth and death of spectral peaks from frame to frame can be attributed to either of the following.

- Pitch variation due to wow degradation or
- Noise due to inaccuracies in the frequency estimation or genuine signal-related change in spectral peak.

The noise component can be assumed to be additive or multiplicative to the pitch variation due to wow degradation [1]. The additive model is appropriate if the noise is assumed to be due to frequency estimation errors. The multiplicative noise is on the other hand appropriate if the noise is assumed to result from genuine signal-related change in spectral peaks. The two noise models are given as follows [1],

$$F_n^i = F_0^i P_n^i V_n^i \tag{5.4}$$

or

$$F_n^i = F_0^i P_n^i + V_n^i \tag{5.5}$$

where:

F_n^i : frequency of tracked spectral peaks of n^{th} frame and i^{th} tonal component
F_0^i : center frequency of genuine i^{th} tonal component
V_n^i : noise component at i^{th} tonal component frequency in n^{th} frame
P_n^i : pitch variation curve of i^{th} tonal component frequency in n^{th} frame.

By taking the natural logarithm of the multiplicative model leads to a linear estimation task.

$$f_n^i = f_0^i + p_n^i + v_n^i \tag{5.6}$$

where lower case (f, p, v) represent the natural logarithm of the upper case (F, P, V). This linear representation leads to significant computational savings over the additive model.

Using the multiplicative noise model and assuming the noise components are zero-mean i.i.d. Gaussian random variables with variance σ_v^2, the likelihood function of spectral peaks given the pitch variation curve and the center frequency of the genuine tonal components was derived in [1] to be,

$$p(\mathbf{F}|\mathbf{p}, \mathbf{f}_0) = \frac{1}{\left(2\pi\sigma_v^2\right)^{\frac{NR}{2}}} \exp\left(\frac{1}{2\sigma_v^2} Q\right) \tag{5.7}$$

where:

$\mathbf{F}$ $= [\mathbf{f}_1 \quad \mathbf{f}_2 \quad \cdots \quad \mathbf{f}_N]$,
$\mathbf{p}$ $= [p_1 \quad p_2 \quad \cdots \quad p_N]$,
$\mathbf{f}_0$ $= [f_0^1 \quad f_0^2 \quad \cdots \quad f_0^N]$,
R : number of spectral peaks tracked. Assumption is made here that number of tracked peaks is same for all frames.
N : number of frames being processed.

and

$$Q = \sum_{n=1}^{N} \sum_{i=1}^{R} \left(v_n^i\right)^2 \tag{5.8}$$

Maximization of the likelihood function is achieved by minimizing Q, which effectively results in a least-squares approach. This can be achieved by taking the gradient of Q in terms of $\mathbf{f}_0$ and $\mathbf{p}$ and setting them to zero. However, this leads to a singular system of equations [1]. This problem can be alleviated by incorporating a priori regularizing information.

5.3.1 Bayesian Estimator

Instead of starting from the raw data, the Bayesian estimator uses the frequency tracks $\mathbf{F}$ obtained from the raw data by using frequency tracking methods [1].

$$p(\mathbf{p}, \mathbf{f}_0|\mathbf{F}) \propto p(\mathbf{F}|\mathbf{p}, \mathbf{f}_0) p(\mathbf{p}, \mathbf{f}_0) \tag{5.9}$$

This is advantageous as the frequency tracks are easier to track by using spectral peak peaking and tracking methods than to work with the raw data itself. However, this leads to loss of some information in the frequency tracking step. By taking a uniform prior for $\mathbf{f}_0$ as the location of the frequency of tonal components is not known and a zero-mean Gaussian prior for the pitch variation curve $p(\mathbf{p}) = N(\mathbf{0}, \mathbf{C_p})$, maximizing the above probability leads to the maximum a posteriori (MAP) [1],

$$\mathbf{p}^{MAP} = \left[\left(R\mathbf{I} - \frac{R}{N}\mathbf{1}_{N\times N}\right) + {\sigma_v}^2\mathbf{C}_p^{-1}\right]^{-1}\left[\mathbf{I} - \frac{1}{N}\mathbf{1}_{N\times N}\right]\mathbf{F}^T\mathbf{1}_R \tag{5.10}$$

However, note that the solution in (5.10) requires knowledge of the prior covariance matrix of the pitch variation curve. This is in practice not available and may be estimated from some a priori information regarding the pitch generation process.

5.3.2 Autoregressive (AR) Model Based Estimator

By assuming that the pitch variation curve can be modeled as an AR process, the covariance matrix $\mathbf{C}_p$ in (5.10) can be represented as a function of the AR coefficients. AR modeling is the most general model as a wide range of pitch defect mechanisms from highly random to sinusoidal can be represented. Filtering a Gaussian white noise with variance σ_e^2 through an AR model generates a Gaussian random process with covariance matrix as shown in (5.11) [1]. This is then incorporated into (5.10) to obtain the solution.

$${\mathbf{C}_\mathbf{p}}^{-1} \approx \frac{\mathbf{A}^T\mathbf{A}}{\sigma_e^2} \tag{5.11}$$

Where $\mathbf{A}$ is a Hankel matrix constructed from the coefficient vector.

This is a general formulation and can incorporate different forms of pitch variation defects. However, it requires knowledge of the AR coefficients of the pitch variation process. This can be estimated from some prior knowledge of the cause of the pitch variation. For example, in the case of a 78rpm recording the pitch variation can be modeled by taking the poles of the AR model to be centered upon 78/60 Hz and close to the unit circle [28].

5.4 HOSpLP Framework for the Characterization of Wow Degradation in Audio Signals

Model based spectral peak estimation has been pointed out to be a better alternative than using the STDFT. In this research, the audio signal is divided into short frames and the HOSpLP framework is used to estimate an AR model of the audio signal in each frame. The wow degradation is then characterized by using spectral peaks of the estimated AR model.

The $\ell 1$-norm regularized and $\ell 0$-norm regularized HOSpLP coefficients are solved by using Algorithm 2 and Algorithm 4 respectively, while the Levinson-Durbin recursion is used to solve for the conventional LP coefficients as a comparison. The first 12 coefficients of the HOSpLP model are then taken to represent the short-term predictor assuming there are a maximum of six spectral peaks in each frame. Such an approach of using the first few coefficients to represent the short-term predictor has been proposed and successfully used for other applications in [16].

The poles of this short-term predictor are then used to obtain the spectral peaks, formants in the case of speech or tones in the case of music. The spectral peaks obtained in each frame are then tracked by using the method developed in [68] and then a pitch variation curve is generated by using any of the methods discussed in Section 5.3.

5.5 Experimental Results

To assess the performance of the proposed HOSpLP framework for the identification and characterization of wow degradation a clean audio signal was artificially degraded with different kinds of wow degradation. The proposed framework was then used to estimate the spectral peaks, these peaks are then tracked and a pitch variation curve is estimated. The conventional STDFT based approach is compared with the conventional LP and HOSpLP framework and the results are reported in the following sections.

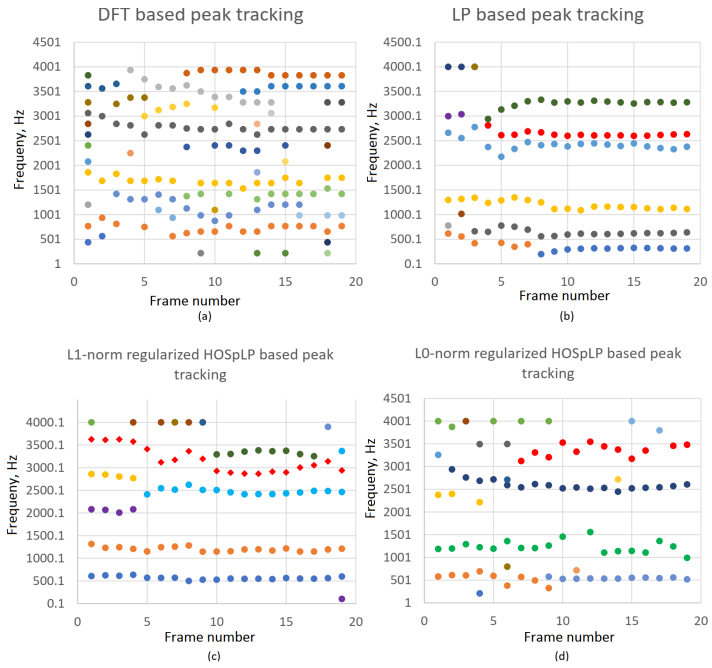


Figure 5.3: Spectral peak frequency tracking example for natural vowel without degradation. Horizontal axis is frame number while vertical axis is frequency in Hz.

5.5.1 Estimation of Spectral Peaks

To observe the performance of the different methods for the identification of spectral peaks in the absence of degradation, the spectral peaks are estimated by using STDFT, LP, ℓ1-norm regularized HOSpLP and ℓ0-norm regularized HOSpLP in each frame for a vowel without wow degradation. The identified peaks are shown in Figure 5.3. For artificial vowel, the formant frequencies are expected to be few in number usually 4 - 6 and to be constant in the entire duration. This property is observed for spectral peaks identified by the use of a model based approach, both conventional LP and HOSpLP, as compared to STDFT. This indicates that model based approaches leads to a better spectral peak identification in the absence of wow degradation.

To see the impact of using model based approaches for the estimation and tracking of spectral peaks in the presence of wow degradation, a vowel was artificially degraded by a sinusoidal speed fluctuation to simulate wow degradation due to rotational speed variation. The artificial sinusoidal speed

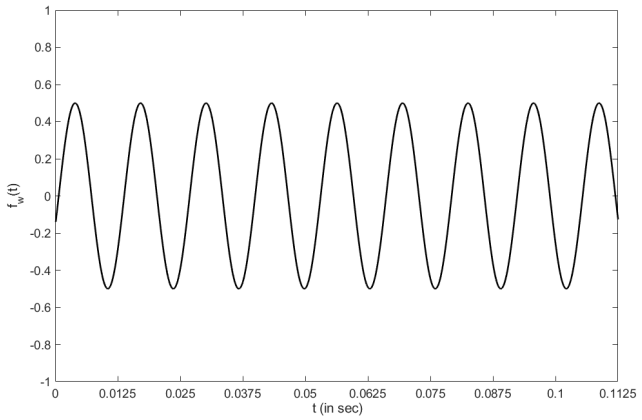


Figure 5.4: Artificial sinusoidal wow degradation due to sinusoidal speed fluctuation

fluctuation is shown in Figure 5.4.

Time domain non-uniform re-sampling as shown in (5.12) was done to introduce this artificial wow degradation into the clean signal.

$$x_\omega(t) = x\big(f_\omega(t)\big)$$

Suppose the wow degradation is periodic speed variation due to disc motor, the actual speed of rotation is then given as follows.

$$\omega_w(t) = \big(1 + \alpha\cos(\omega_0 t)\big)\omega_0$$

As the disc is supposed to be replayed at constant speed, the actual speed of rotation effectively introduces a time warping. The time warped signal is then given as follows.

$$x_w(t) = x\big(\frac{1}{\omega_0}\int_0^t \omega_w(\tau)d\tau\big)$$

$$x_w(t) = x\big(\int_0^t (1 + \alpha\cos(\omega_0\tau))d\tau\big)$$

$$x_w(t) = x\big(t + \frac{1}{\omega_0}\alpha\sin(\omega_0 t)\big)\big) \tag{5.12}$$

Where ω_o is the rate of change of the sinusoidal wow degradation and α is the amplitude of wow degradation.

The spectral peaks identified by using STDFT, LP, $\ell 1$-norm regularized HOSpLP and $\ell 0$-norm regularized HOSpLP in each frame for a vowel with wow degradation is shown in Figure 5.5. In the presence of wow degradation, it is seen that the use of the STDFT results in a lot more spectral peaks to be identified than the spectral peaks in the original signal making it quite difficult to track changes in subsequent frames. On the other hand, a model based approach leads to a better spectral peak identification in the presence of wow degradation. The spectral peaks obtained via $\ell 1$-norm regularized HOSpLP and $\ell 0$-norm regularized HOSpLP show that all the spectral peaks are degraded by a certain periodic degradation.

5.5.2 Performance of the Proposed Model Based Pitch Variation Curve Estimation

To see the performance of the proposed model based frequency tracking and pitch variation curve estimation, the frequency of the spectral peaks identified by using the three model based approaches were tracked by the algorithm developed in [68]. The generated frequency tracks were then used to estimate a pitch variation curve by using the Bayesian estimator and AR model based estimator. The estimated pitch variation curves are shown in Figure 5.6 and Figure 5.7.

It is clearly seen from Figure 5.6 and 5.7 that the three model based spectral peak identification methods lead to a pitch variation estimate that is much better than the conventional STDFT based approach. The LP, $\ell 1$-norm regularized HOSpLP and $\ell 0$-norm regularized HOSpLP based pitch variation estimation approaches lead to similar accuracy. MSE in pitch variation curve estimation is used to compare the accuracy of the pitch variation curve estimation in the entire duration as shown in Table 5.1. As can be seen in Table 5.1, in the absence of noise the three model based spectral peak tracking followed by pitch variation curve estimation lead to similar MSE, with LP based spectral peak tracking leading to a very small improvement over the $\ell 1$-norm regularized HOSpLP and $\ell 0$-norm regularized HOSpLP.

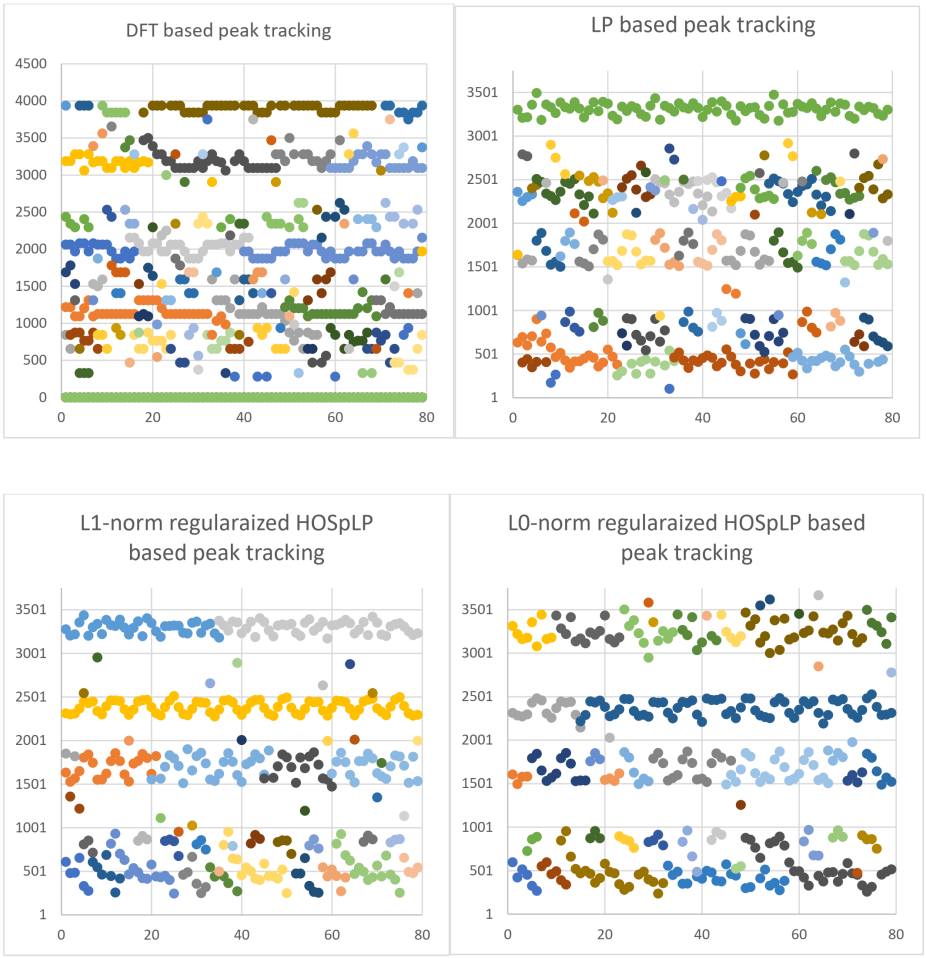


Figure 5.5: Spectral peak frequency tracks for natural vowel with wow degradation. Horizontal axis is frame number while vertical axis is frequency in Hz.

5.5.3 Noise Robustness of the Proposed Model Based Pitch Variation Curve Estimation

To test the robustness of the proposed model based frequency tracking and pitch variation curve estimation in the presence of background noise, white noise was added to the artificially wow degraded vowel. The spectral peaks were then identified by using the three model based approaches and then tracked by the algorithm developed in [68]. The generated frequency tracks

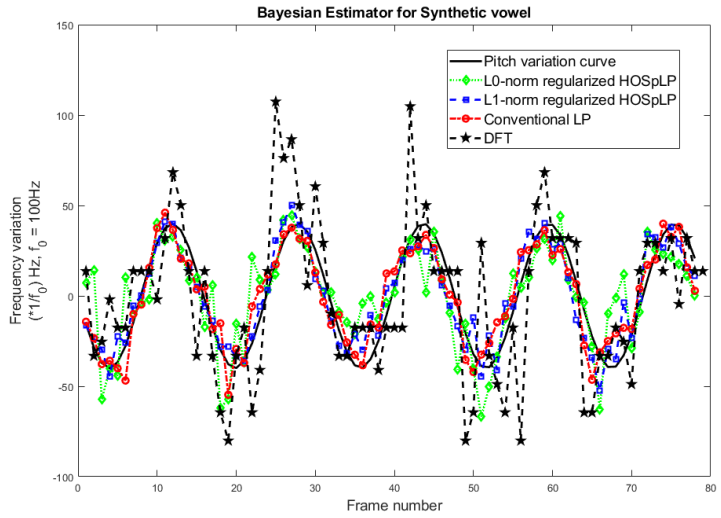


Figure 5.6: Pitch variation estimation sample with wow for vowel by using Bayesian pitch variation curve estimation.

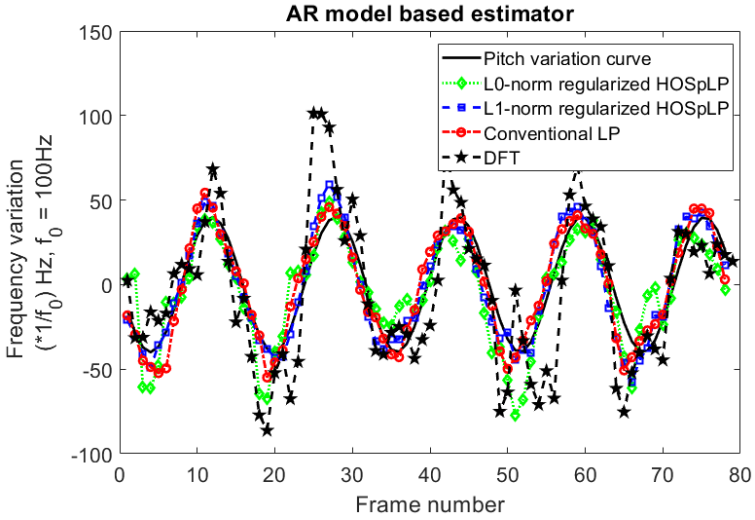


Figure 5.7: Pitch variation estimation sample with wow for vowel by using AR model based pitch variation curve estimation.

were then used to estimate a pitch variation curve by using the Bayesian estimator and AR model based estimator. The estimated pitch variation curves are shown in Figure 5.8 and 5.9 for SNR of 20dB and 5.10 and 5.11

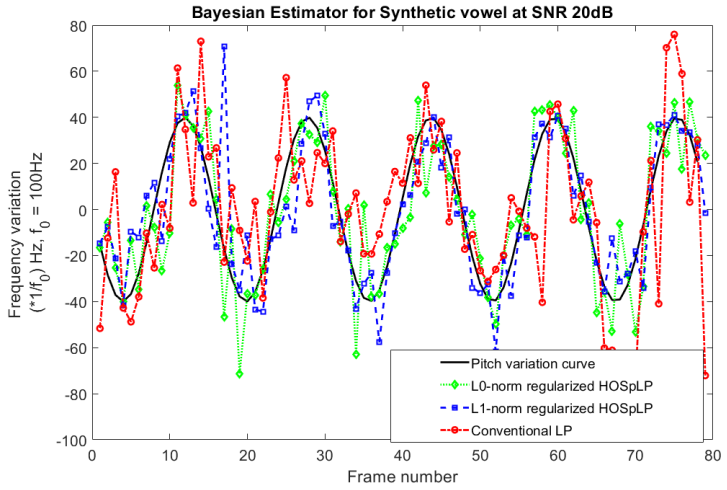


Figure 5.8: Pitch variation estimation example with wow for 20dB SNR noisy vowel by using Bayesian pitch variation curve estimation.

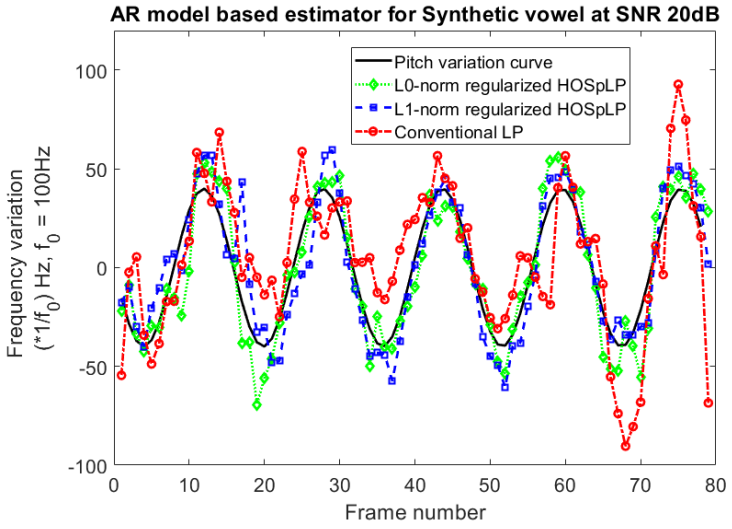


Figure 5.9: Pitch variation estimation sample with wow for 20dB SNR noisy vowel by using AR model based pitch variation curve estimation.

for SNR of 10dB.

It is seen in Figure 5.8 to Figure 5.11 that in the presence of background noise the use of $\ell 1$-norm regularized HOSpLP and $\ell 0$-norm regularized HOSpLP based approaches lead to pitch variation curve estimation that is much

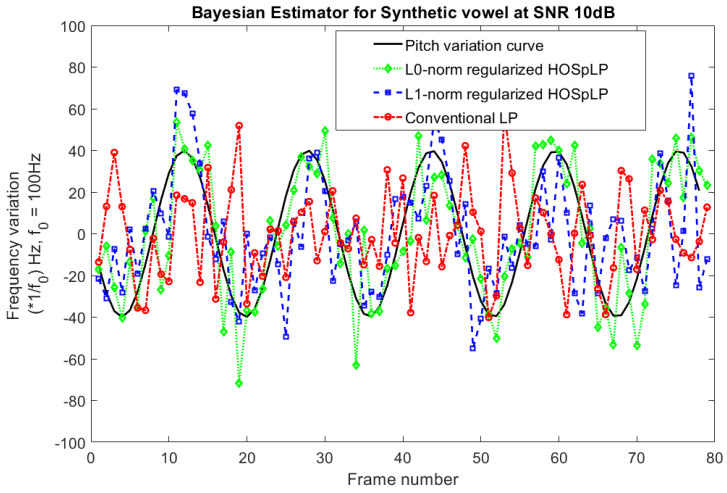


Figure 5.10: Pitch variation estimation sample with wow for 10dB SNR noisy vowel by using Bayesian pitch variation curve estimation.

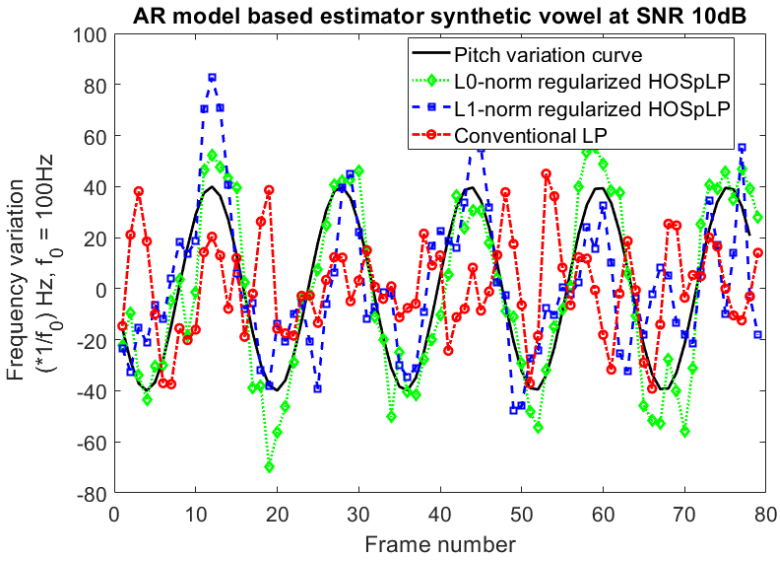


Figure 5.11: Pitch variation estimation sample with wow for 10dB SNR noisy vowel by using AR model based pitch variation curve estimation.

better than conventional LP based peak identification.

The mean-square error of the pitch variation curve estimation is computed to quantify the performance of the approaches in estimating the pitch vari-

Table 5.1: MSE in pitch variation curve estimation.

SNR		LP	ℓ1-norm HOSpLP	ℓ0-norm HOSpLP
∞dB	STDFT: 3972	122.23	126.98	170.78
20dB	Bayesian	682.96	274.85	304.29
	AR	560.53	193.90	194.81
10dB	Bayesian	1310	660.14	306.89
	AR	1191	481.11	186.27

ation curve. Table 5.1 shows the MSE of the four approaches in estimating the pitch variation curve. It is seen that even though the proposed HOSpLP framework achieves slightly higher MSE as compared to the conventional LP based approach for a clean signal, in the presence of background noise the HOSpLP framework leads to significantly lower MSE.

5.6 An Application Program for Characterization of Wow Degradation

A MATLAB application program is developed for the detection and characterization of wow degradation in archived audio. The application program does not need information regarding the wow degradation. Furthermore, the user does not need expert knowledge on the different algorithms used in this research to use this application program. It takes the path of the audio file to be restored, the type of degradation (click or wow-in this case), the algorithm to be used, the window size and the coefficient type. Figure 5.12 shows a screen shot of the application program.

The following four approaches are selected for the application program based on their performance in terms of best result, noise robustness and computational time consideration. The application program was tested on artificially wow degraded audio signals and it was noted to perform similar to the presented results.

- Bayesian estimator based on HOSpLP
- Bayesian estimator based on conventional LP
- AR model based estimator based on HOSpLP

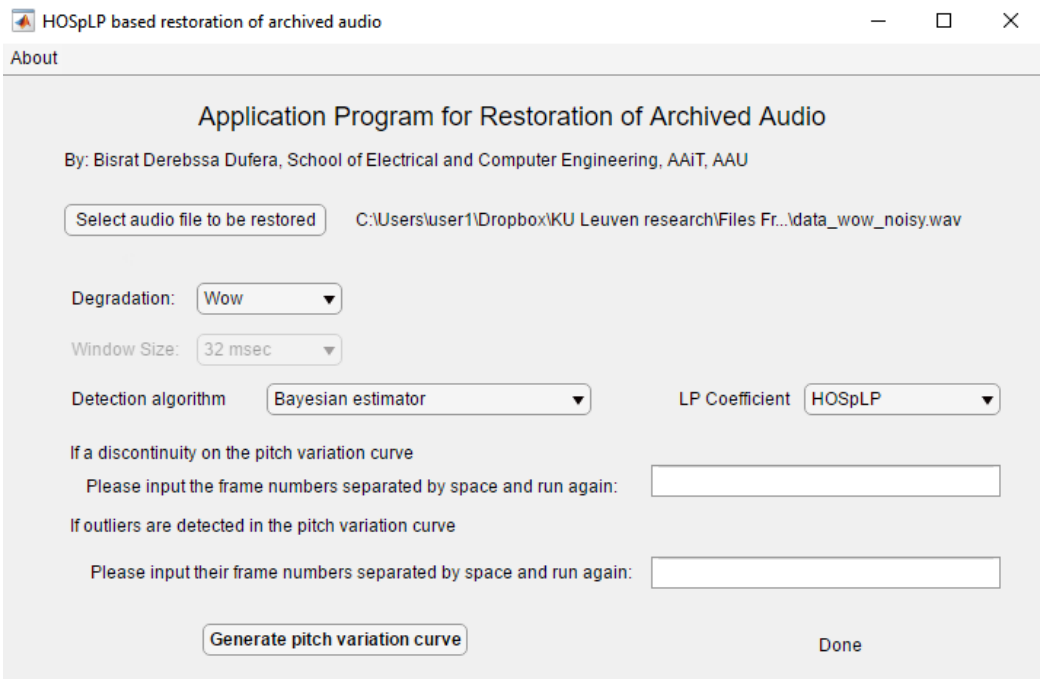


Figure 5.12: Application program for the restoration of Archived Audio

- AR model based estimator based conventional LP

5.7 Conclusion

In this chapter, the use of the high-order sparse linear prediction is proposed for the identification and characterization of wow degradation in audio. Several experiments were conducted to assess the performance of the proposed HOSpLP coefficients in identification and characterization of wow degradation for clean as well as noisy environments. The propose method achieved significant improvement in wow identification and characterization.

Chapter 6

Conclusion

This research proposed the use of a high-order sparse linear prediction coefficients for the restoration of an audio signal that has been degraded by click and wow degradation. The proposed methods rely on LP based outlier detection for the identification of samples degraded by clicks and LP based missing sample estimation for the restoration of samples degraded by clicks as well as spectral tracking by using LP model based spectral peak identification for audio degraded by wow degradation. The proposed methods used two high-order sparse linear prediction coefficients that do not require a priori knowledge on the type of audio, the location and duration of click degradation, and can identify and characterize wow degradation. Extensive analysis and simulation results using wide range of audio dataset has been conducted. From the results and discussion it is clear that the use of high-order sparse linear prediction coefficients lead to improved identification and characterization of degradations as well as improved restoration in terms of signal-to-noise ratio as well as perceptual evaluation of audio quality as compared to state of the art conventional LP based approaches.

One limitation of the experiments conducted in this research is the fact that clicks and wow degradations were treated separately. The modeling assumptions, the developed algorithms and the experiments conducted when dealing with each degradation ignored the other degradation. However, in practice, an audio may be degraded by these two degradations at the same time. An investigation is needed to evaluate how the modeling assumptions taken and each of the developed method for each degradation are affected when both degradations are present. From such investigation a way forward

on a joint restoration of an audio that is degraded by both clicks and wow may be proposed.

As the aim of any audio restoration technique is to increase the perceived audio quality as heard by a human listener, assessment of the proposed methods on real archived audio signals is needed for final remark on their performance. Evaluation of the mathematical models and algorithms developed in this research on real degraded audio signal is expected to introduce additional complications as the undegraded data is not available. As such, the different evaluation metrics used in this research cannot directly be applied. Subjective quality assessment methods are then needed for fair assessment of the proposed methods. However, it is anticipated that the results obtained in this research will still hold for real degraded audio data. This is due to the fact that the artificial click and wow degradations used in this research are obtained by using models that were developed by other researchers after analyzing actual degradations. Furthermore, in the case of click degradation we have chosen to discard the samples in the click degradation to make the assessment not dependent on the actual values inside the click degradation which may be different in the case of real degradations.

The restoration using HOSpLP coefficients is inherently time taking due to the complexity of the optimization problem involving $\ell 1$-norm and $\ell 0$-norm. As a result of this, non of the detection and restoration approaches are not expected to be real-time. Therefore, research on efficient solutions to the optimization problems is needed. Recent developments in efficient solutions to related optimization problems and advance in processing power are indications that such sparsity constrained solutions will be applied to several fields that have been hampered by complex and time taking optimization problems. One example is the case of the $\ell 0$-norm regularized HOSpLP which was solved using a very recent algorithm developed by researchers at KU Leuven on optimization problems for signal processing applications. However, as the task at hand is the restoration of archived audio that has been stored for decades, the fact that our proposed restoration methods are not real time is not expected to be a significant limitation.

The low restoration performance of all the investigated LP based restoration methods for long click durations can be attributed to the limitation of using least square autoregressive interpolator for long gaps in audio signals. This is due to the non-stationary nature of audio signals for longer durations that are needed to estimate longer LP coefficients for good performance for long click durations. Depending on the length of the click degradation and

whether the frame in consideration is near a phoneme transition, it may even not be possible to come up with a frame length and subsequently LP order large enough to estimate the missing samples. In such cases, the problem may be approached from a perspective of audio recognition based on natural language processing and similar works.

Bibliography

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

[2] "The state of recorded sound preservation in the united states," *National Recording Preservation Board, Library of Congress, Washington DC*, 2010.

[3] L. Koch, E. Gomez-Sanchez, S. Simon, M. Mengel, and A. Wiedmann, "Integrated solutions for preservation, archiving and conservation of endangered magnetic tapes and cylinders," *International Association of Sound and Audiovisual Archives*, vol. 1, pp. 44–49, 2009.

[4] "Save our sounds," *British Library*, October 2015.

[5] D. Bekele, "The legal framework for freedom of expression in ethiopia," *Global Campain for Free Expression, Sweden*, 2003.

[6] R. L. Hess, "Tape degradation factors and challenges in predicting tape life," *Association for Recorded Sound COllections*, pp. 240–247, 2008.

[7] P. Sprechmann, A. Bronstein, J.-M. Morel, and G. Sapiro, "Audio restoration from multiple copies," in *International Conference on Acoust., Speech and Signal Processing*, IEEE, 2013.

[8] S. J. Godsill and P. J. W. Rayner, "A bayesian approach to the restoration of degraded audio signals," *IEEE Trans. Speech Audio Processing*, vol. 3(4), pp. 267–278, July 1995.

[9] M. K. Mathai and J. Deepa, "Design and implementation of restoration techniques for audio denoising applications," in *IEEE Recent Advances*

in Intelligent Computational Systems, (Trivandrum, India), pp. 21–26, IEEE, 2015.

[10] M. J. Carey and I. Buckner, "A system for reducing impulsive noise on gramophone reproduction equipment," *The Radio and Electronic Engineer*, vol. 50(7), no. 7, pp. 331–336, 1980.

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

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

[13] M. Niedźwiecki and M. Ciołłek, "Elimination of clicks from archive speech signals using sparse autoregressive modeling," *J. Audio Eng. Soc.*, vol. 55(5), pp. 891–905, May 2005.

[14] M. Niedźwiecki, M. Cioek, and K. Cisowski, "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.

[15] M. Ciołek and M. Niedźwiecki, "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.

[16] D. Giacobello, M. G. Christensen, M. N. Murthi, 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.

[17] F. R. Avila and L. W. P. Biscainho, "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.

[18] M. Lagrange and S. Marchand, "Long interpolation of audio signals using linear prediction in sinusoidal modeling," *J. Audio Eng. Soc.*, vol. 55(5), pp. 891–905, May 2005.

[19] A. Janssen, R. Veldhuis, and L. Vries, "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.

[20] L. Shi, J. R. Jensen, and M. G. Christensen, "Least 1-norm polezero modeling with sparse deconvolution for speech analysis," in *Proc. 2009 IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 09)*, (New Orleans, LA, USA), June 2009.

[21] D. Giacobello, M. G. Christensen, J. Dahl, S. H. Jensen, and M. Moonen, "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.

[22] I. Erer and K. B. Sarikaya, "Enhanced radar imaging via sparsity regularized 2d linear prediction," in *Proc. 22nd European Signal Process. Conf. (EUSIPCO '14)*, vol. 20(3), (Taipei, Taiwan), pp. 1751–1755, March 2014.

[23] T. L. Jensen, D. Giacobello, T. van Waterschoot, 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.

[24] M. Niedźwiecki and K. Cisowski, "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.

[25] S. Canazza, G. D. Poli, and G. A. Mian, "Restoration of audio documents by means of extended kalman filter," *IEEE Transaction on Audio, Speech, and Lang. Processing*, vol. 18(6), pp. 1107–1115, August 2010.

[26] A. Czyzewski, J. Kotus, M. Kulesza, and P. Maziewski, "DSP Techniques for Determining Wow Distortion," *J. Audio Eng. Soc.*, vol. 55, no. 4, p. 19, 2007.

[27] S. J. Godsill, "Recursive restoration of pitch variation defects in musical recordings," in *International Conference on Acoust., Speech and Signal Processing*, vol. 2, IEEE, 1994.

[28] S. J. Godsill and P. J. W. Rayner, "The restoration of pitch variation defects in gramophone recordings," in *Proc. 1993 IEEE Workshop Appls. Signal Process. Audio Acoust. (WASPAA '93)*, IEEE, 1993.

[29] A. Czyzewski, "Wow detection and compensation employing spectral processing of audio," in *117th Audio Engineering Society Convention*, p. 19, 2004.

[30] C. J. Shahani, M. H. Youket, and N. Weberg, "Compact disk service life: An investigation of the estimated service life of prerecorded compact discs," tech. rep., Library of Congress, 2009.

[31] B. D. Dufera, "Review of early storage media degradation factors, preservation techniques and trends in ethiopia," *ZEDE*, vol. 37, pp. 27–38, May 2019.

[32] D. Schulle, "Audio and video carriers: Recording principles, storage and handling, maintenance," *Training for Audiovisual Preservation in Europe*, February 2008.

[33] P. Shambarger, "Cylinder records: An overview," *Association for Recorded Sound Collections*, vol. XXVI, pp. 133–161, 1995.

[34] S. Brylawski, M. Lerman, R. Pike, and K. Smith, "Arsc guide to audio preservation," *Association for Recorded Sound Collections, Council on Library and Information Resources and The Library of Congress*, 2015.

[35] , "Wax cylinders," in *Available: http://www.cylinder.de/guide_black-wax-cylinders.html.[Accessed 12 December 2016].*

[36] A. G. Lemcoe and P. M. M, "Preservation and storage of sound recordings," tech. rep., Library of Congress, 1959.

[37] G. S. Laurent, "The preservation of recorded sound materials," *Association for Recorded Sound Collections*, vol. 22, pp. 144–156, 1992.

[38] D. Schuller, "The safeguarding of the audio heritage: Ethics, principles and preservation strategy," *International Association of Sound and Audiovisual Archives*, vol. 3, 2005.

[39] H. K.Yamamoto, "A kinetic study of hydrolysis of polyester elastomer in magnetic tape," tech. rep., SONY Corporation Sendai Technology Center, Japan, 1998.

[40] M. Casey, *Format Characteristics and Preservation Problems*. PhD thesis, Indiana University, USA, 2007.

[41] D. F. Daniel and E. D. E, "Causes of failure in magnetic tape," in *Magnetic Recording in Science and Industry*, 1967.

[42] I. T. Committe, "Guidelines on the production and preservation of digital audio objects," tech. rep., IASA, 2009.

[43] L. of Congress, "The state of recorded sound preservation in the united states," tech. rep., Library of Congress, 2010.

[44] D. Giacobello, T. van Waterschoot, M. G. Christensen, S. H. Jensen, and M. Moonen, "High-order sparse linear predictors for audio processing," in *Proc. 20th European Signal Process. Conf. (EUSIPCO '10)*, (Aalborg, Denmark), pp. 234–238, August 2010.

[45] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley-IEEE Press, October 1999.

[46] M. Ciołłek and M. Niedźwiecki, "Detection of impulsive distrubances in archive audio signals," *in Advances in Speech and Language Technologies for Iberian Languages. (Torre Toledano D. et al., Eds.)*, vol. 328(5), pp. 247–256, May 2012.

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

[48] Y. Qian, G. Chahine, and P. Kabal, "Pseudo-multi-tap pitch filters in low bit-rate celp speech coder," *Speech Communication*, vol. 14, pp. 339–358, March 1994.

[49] B. D. Dufera, K. Eneman, and T. van Waterschoot, "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.

[50] P. Stoica and T. Sderstrm, "High order yule-walker equations for estimating sinusoidal frequencies: The complete set of solutions," *Signal Process*, vol. 20, pp. 257–263, July 1990.

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

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

[53] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20(3), pp. 922–932, March 2012.

[54] J. Durbin, "The fitting of time series models," *Review of the International Statistical Institute*, vol. 28, p. 11, 1960.

[55] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 3(67), pp. 971–995, March 1980.

[56] J. M. Hillenbrand, "Acoustic characteristics of american english vowels." https://homepages.wmich.edu/ hillenbr/voweldata.html. accessed Dec. 14, 2017.

[57] Voxforge.org, "Free speech ... recognition (linux, windows and mac) - voxforge.org." http://www.voxforge.org/. accessed Dec. 14, 2017.

[58] SMALL, "Sparse models, algorithms and learning for large-scale data." http://www.small-project.eu/. accessed Oct. 13, 2017.

[59] M. Brandt, S. Doclo, T. Gerkmann, and J. Bitzer, "Impulsive disturbances in audio archives: signal classification for automatic restoration," *J. Audio Eng. Soc.*, vol. 65(10), pp. 826–840, Oct. 2017.

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

[61] B. D. Dufera, E. Adugna, K. Eneman, and T. van Waterschoot, "Restoration of click degraded speech and music based on high order sparse linear prediction," in *IEEE AFRICON 2019*, (Accra, Ghana), September 25-27, 2019.

[62] Z. Ben-Haim and Y. C. Eldar, "The cramr-rao bound for estimating a sparse parameter vector," *IEEE Trans. Signal Process.*, vol. 58(6), pp. 3384–3389, June 2010.

[63] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*. Wiley, January 2009.

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

[65] M. Niedźwiecki and M. Ciołek, "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.

[66] M. Ciołek and M. Niedźwiecki. http://eti.pg.edu.pl/katedra-systemow-automatyki/ICASSP201. [Accessed 01 February 2020].

[67] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Pearson, January 1978.

[68] R. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-34, p. 11, August 1986.