PRELIMINARY INVESTIGATION INTO MACHINE LEARNING AND SIGNAL PROCESSING METHODOLOGIES FOR BLIND RECOVERY AND RESTORATION OF OLD AUDIO RECORDINGS

Jiaxi You Department of EEE, The University of Manchester, Manchester, UK
Claire Mitchell Division of Neuroscience & Experimental Psychology, The University of Manchester, Manchester, UK
Keming Tan Department of EEE, The University of Manchester, Manchester, UK
Erdem Atbas Department of EEE, The University of Manchester, Manchester, UK
Patrick Gaydecki Department of EEE, The University of Manchester, Manchester, UK

1 INTRODUCTION

Vintage audio recordings often hold historical and cultural significance and can serve as educational tools. Restoring vintage audio allows people to experience nostalgia and connect with the past, enhancing their enjoyment and appreciation of cultural artifacts[1]. Additionally, recovering vintage audio provides insight into the development of record and playback technologies over time.

The recordings' degradation types can be various due to factors such as age, storage conditions, or technology limitations of the time[2]. In this paper, the degradations are grouped into two types: linear and nonlinear. Linear degradations typically encompass consistent alterations that extend across the entire audio signal, exerting influence over frequency bandwidth and overall fidelity. In contrast, addressing nonlinear degradations can be more challenging than linear degradations. Nonlinear degradations introduce irregularities and new elements, often resulting in more noticeable and disruptive distortions. The paper utilises multiple algorithms to address restoration tasks. It tackles linear problems like limited bandwidth and hiss noise, as well as nonlinear issues, the elimination of pops and crackles.

At the end of this paper, the drawbacks of relying solely on traditional Digital Signal Processing (DSP) methods for audio restoration are briefly discussed. Additionally, the rationale for integrating Machine Learning (ML) and neural networks in this field is highlighted.

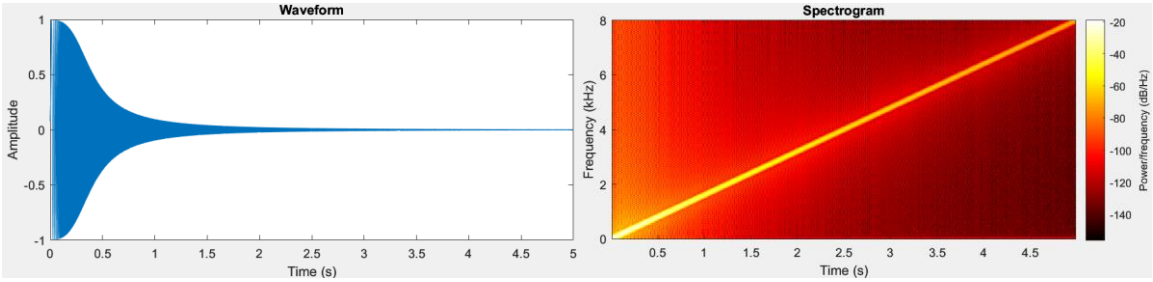
2 LINEAR DEGRADATIONS

2.1 Limited Bandwidth

Vintage audio recordings often have limited bandwidth due to the technological limitations of the equipment and recording techniques used when those recordings were made. The recording equipment, such as tape recorders and preamplifiers, restricted the ability to capture and reproduce the full spectrum of audio frequencies. Also, the playback equipment during the time of the vintage recordings, for instance, early radios and phonographs, had barriers in terms of frequency response. To ensure optimal playback with these systems, engineers might have tailored the recordings, resulting in limited bandwidth[1]. Improper storage or deterioration of the recording medium can lead to audio quality issues, such as limited bandwidth as well[2]. Physical deterioration or oxidation and rust of analogue recordings and magnetic tapes can cause the loss of high-frequency details. In this case, the quality of sound might be similar to a high-fidelity signal that has been processed by a bandpass filter.

To simulate these conditions, a test audio signal was artificially degraded by low-pass filters and blind deconvolution was employed to restore the audio to its original quality. To better understand the

recovered signal, a sinusoidal sweep signal ranging from 0 Hz to 8K Hz was used for testing. The signal was degraded by a second-order low-pass Butterworth filter with a -3 dB point of 1k Hz.



(a) (b)

Figure 1. The degraded signal and its spectrogram.

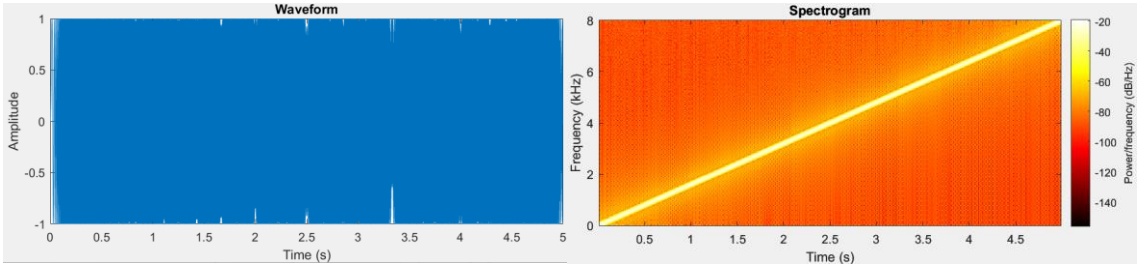
The degraded signal $y[n]$ has a roll-off of 12 dB per octave and any frequency above 1k Hz is attenuated, as shown in Figure 1 (a). In real restoration cases, it is not easy to estimate the accurate degradation filter process To ensure the effectiveness of the following deconvolution algorithm, it was assumed that the degradation frequency response was known. With the filter's impulse response $h[n]$ generated by the filter design package *Signal Wizard*[3], a Toeplitz matrix A was created, whose size was determined by the lengths of the degraded signal and impulse response, and convolution shape. The example here shows the size of the degraded signal segment to be 1000 with 127 impulse response coefficients (taps), with full convolution. Thus the convolution pattern matrix A has a size of (1000, 1126) and output of full convolution[4] $y[n]$:

$$y[n] = (x * h)_{full}[n] = \sum_{k=0}^{\infty} x[k] \cdot h[n-k] \tag{1}$$

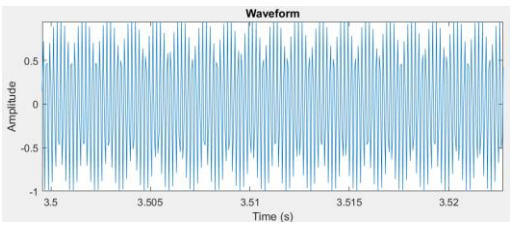
With the convolution pattern A, the degraded output $y[n]$:

$$y[n] = A \cdot x[n] \tag{2}$$

To restore the original signal $x[n]$ by deconvolving the degraded signal $y[n]$, a least-square solution was empoyed. This solution minimizes the error $|y[n] - A \cdot x[n]|$ to derive the vector $x[n]$ representing the original signal. The recovered result is shown in Figure 2.



(a) (b)



(c)

Figure 2. Recovered signal and its spectrogram.

The attenuated signal in Figure 1 is fully recovered to the original amplitude, but the resulting spectrogram shows the introduction of some high-frequency noise, which is barely audible. This could result from the digitization, since the floats data type has 32-bit intermediate precision. This algorithm requires a significant amount of time for the restoration of extended audio segments, owing to its two-dimensional matrix computation. Nevertheless, the obtained results yield an excellent level of accuracy in the bandwidth-limited restoration process, confirming the efficacy of the presented method.

2.2 Hiss Noise

An adaptive filter is an adjustable digital filter whose parameters or coefficients can change in response to the input or the desired output. Least Mean Squares (LMS) algorithm and Recursive Least Squares (RLS) algorithm adaptive filters are discussed here for hiss removal, which is a type of broadband noise that is most associated with analogue recording and playback systems[2]. Ideally, adaptive filters require two inputs, the signal degraded with noise and the noise. In most cases of historical restoration, the noise source is unknown, so the delayed input adaptive filter is used, as shown in Figure 3.

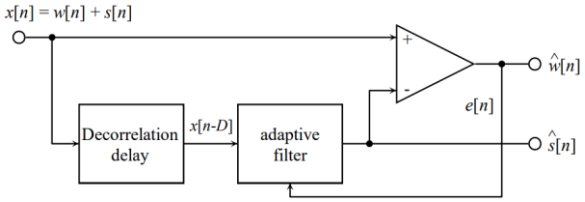


Figure 3. A delayed input adaptive filter[4].

The inputs of the filter are the signal $x[n]$ and its delayed version $x[n-D]$, where D is the decorrelation delay. For the broadband noise $w[n]$, $x[n]$ should have no correlation with $x[n-D]$ with respect to the noise while highly correlated with $x[n-D]$ with respect to the clean signal $s[n]$. The two outputs are the estimated noise $\hat{w}[n]$ and estimated clean signal $\hat{s}[n]$, so the error of the estimation is:

$$e[n] = x[n] - \hat{s}[n] = \hat{w}[n] \tag{3}$$

2.2.1 Least Mean Squares Adaptive Filter

In the LMS adaptive filter, the fraction of the error $e[n]$, which represents the discrepancy between the filter's response and the expected clean signal, is iteratively learned to update filter coefficients to minimize the error[4].

$$h_{new}[k] = h_{old}[k] + \Delta e[n]x[n-k-D] \tag{4}$$

To better understand the convergence of the LMS algorithm, a test signal was generated comprising a 4k Hz sine wave degraded by white noise, with a signal-to-noise ratio (SNR) of 8 dB.

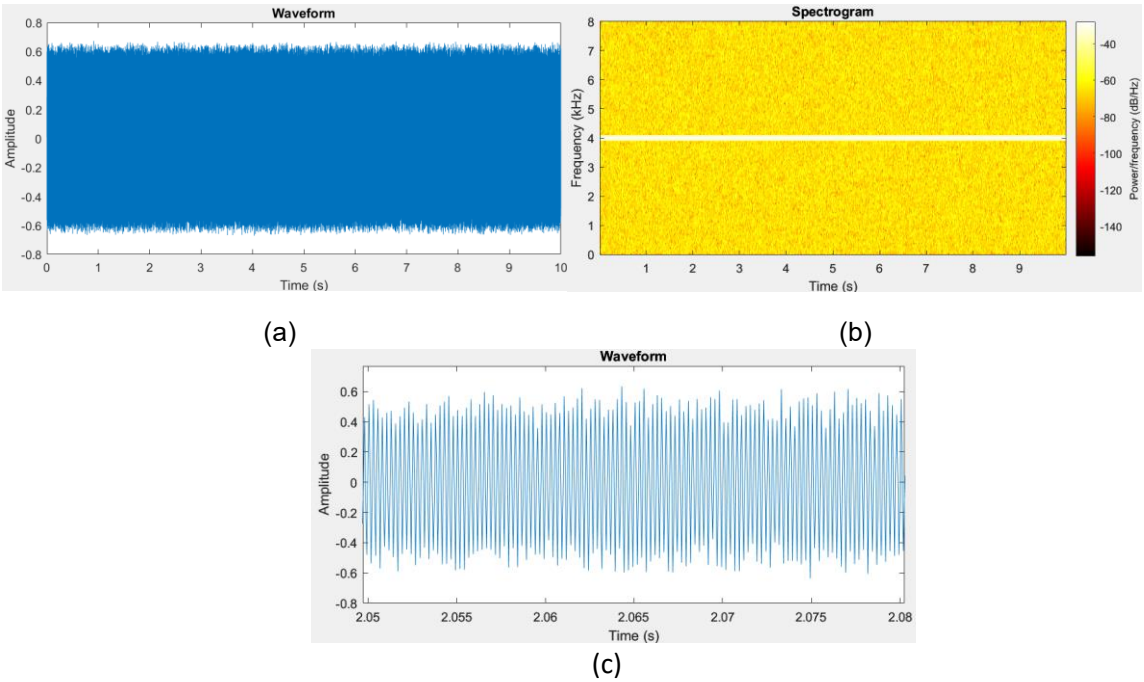


Figure 4. The degraded test audio.

Since broadband noise is completely random and there is no correlation between two intervals, the decorrelation delay *D* in this case is set to be 1. The LMS adaptive filter was designed to a size of 11 and a learning rate of 0.001, the audio was processed by the filter three times, a procedure termed cascade filtering. It can lead to a more stable update and help mitigate the effects of noise and outliers in the individual data points. The result is shown in Figure 5.

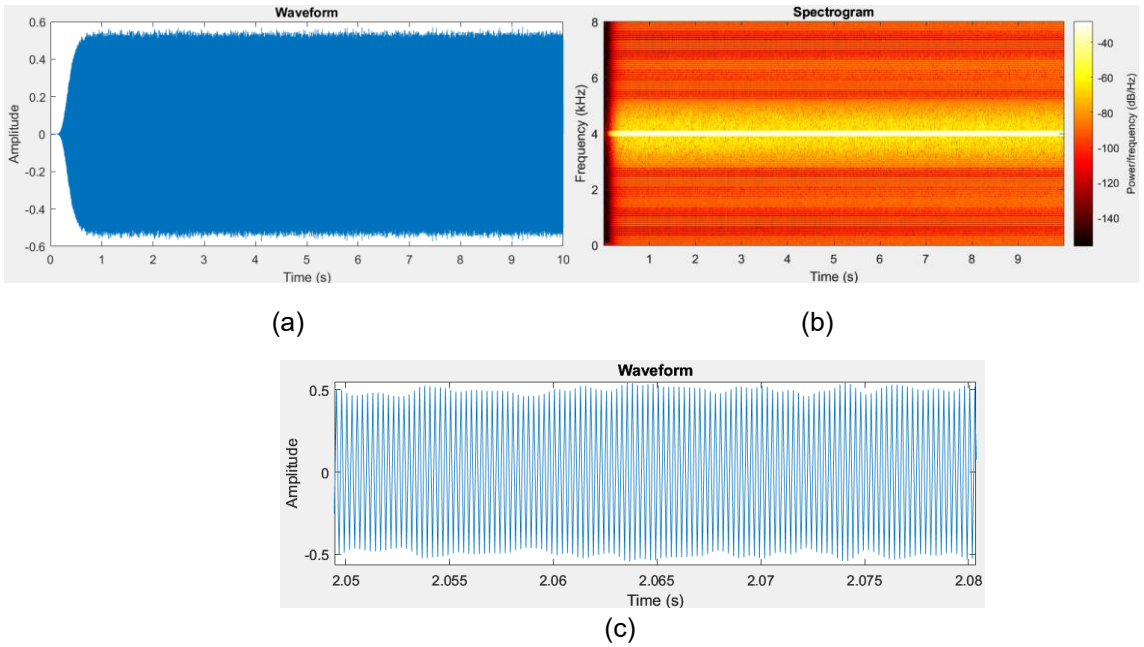


Figure 5. The result after LMS adaptive filter.

The result shows that cascade filtering provides stable and consistent convergence in this case but also exhibits imperfect recovery, confirmed by the spectrogram. This predominantly arises because the noise estimate is never perfectly accurate, and so the filter's performance is limited by the value

of *D*. The result might be improved by adjusting parameters in more detail. Generally, the LMS adaptive filter can give good results aurally if the correlation between noise and wanted signal can be analysed accurately.

2.2.2 Recursive Least Squares Adaptive Filter

The RLS algorithm[5] operates recursively, updating the filter coefficients as new data samples become available, which means it requires more memory and computational resources. The RLS filter updates coefficients using the Kalman gain $K[n]$ and the error signal.

$$K[n] = \frac{P[N-1] \cdot x[n]}{\lambda + x[n]^T \cdot P[N-1] \cdot x[n]} \tag{5}$$

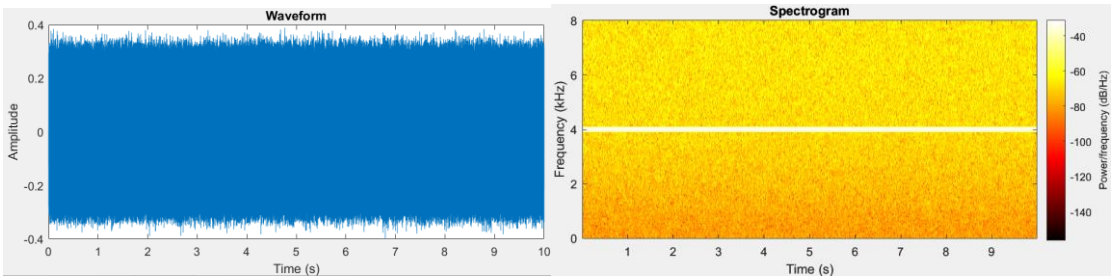
Where $P[N]$ is the inverse covariance matrix and λ is the forgetting rate, a small positive constant. $P[N]$ is updated with the following equation:

$$P[N] = \frac{P[N-1] - K[N] \cdot x[n]^T \cdot P[N-1]}{\lambda} \tag{6}$$

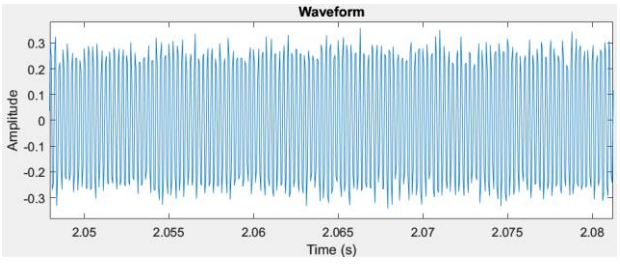
So filter coefficients are updated by:

$$h[n] = h[n-1] + K[N] \cdot e[n] \tag{7}$$

The test audio used above was sent to this RLS adaptive filter, whose forgetting rate is 0.75 with a filter size of 8000; the result is shown in Figure 6.



(a) (b)



(c)

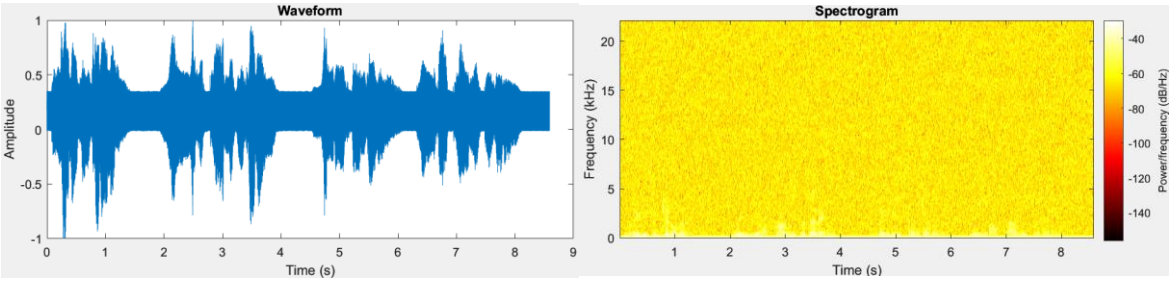
Figure 6. The result after RLS adaptive filter.

The result was still very noisy while it attenuated the overall audio amplitude. The recursive nature of the filter allows the filter to track changes in the input signal and noise characteristics quickly, which also means that the RLS adaptive filter can have fast convergence and should be more suitable for nonstationary audio, instead of this steady sine wave.

2.2.3 Spectral Subtraction

While both RLS and RMS adaptive filters work with signals in the time domain, some techniques operate in the frequency domain, such as spectral subtraction and the Wiener filter.

In the Fourier domain, signals are represented as a sum of sinusoidal components/ frequencies through the Fourier transform[2]. Spectral subtraction takes advantage of this representation by estimating the noise spectrum in the frequency domain and then subtracting the estimated noise spectrum from the observed spectrum to enhance the clean signal[6]. In audio recordings, broadband noise normally covers the entire audio while the actual desired signal includes periods of silence; thus the hiss noise can be estimated from these quiescent periods. Figure 7 shows speech audio degraded by the broadband noise, with an SNR of 15 dB.



(a) (b)

Figure 7. Degraded speech audio, SNR 15 dB.

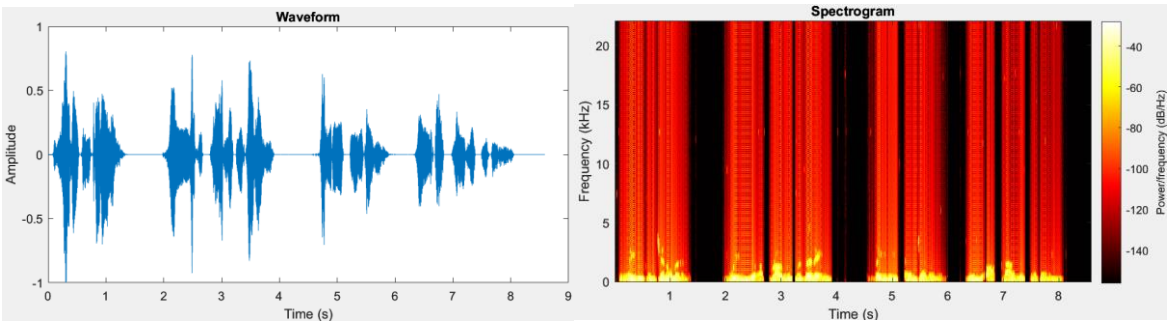
By listening to the audio and analysing the time domain waveform, the last 0.5 s of the audio is considered as the estimated noise. Both the original audio and the estimated noise are transformed to the frequency domain using the short-time Fourier transform (STFT)[2,4]. The window size is set to be 512, with an overlap of 75%, and the discrete Fourier transform of vector shows as follows, where $X[n]$ is the time domain signal and $Y[k]$ is the Fourier domain signal.

$$Y(k) = \sum_{j=1}^{n} X[j] e^{((-2\pi i)/n)(j-1)(k-1)} \tag{8}$$

After subtracting the estimated noise spectrum from the original audio spectrum, the inverse Fourier transform is applied to change the spectral domain result to the time domain waveform. The recovered result is shown in Figure 8.

$$S[k] = |X[k]| - N[k] \tag{9}$$

$$Z(j) = \frac{1}{n} \sum_{k=1}^{n} S[k] e^{-((-2\pi i)/n)(j-1)(k-1)} \tag{10}$$

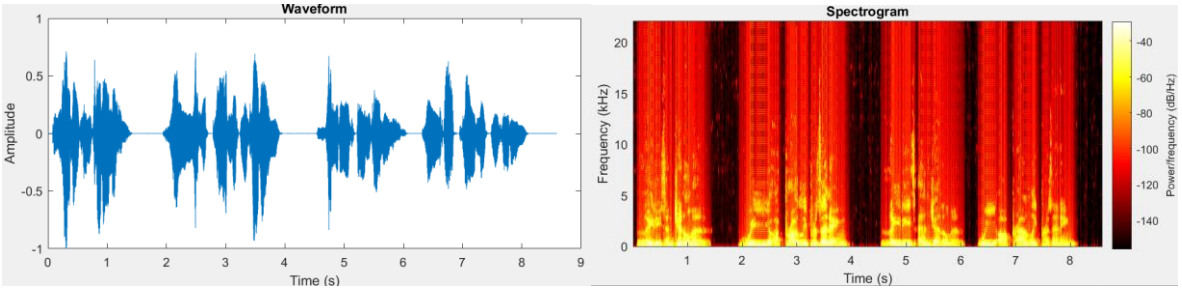


(a) (b)

Figure 8. The recovered speech audio of Figure 7.

In Equations (9) and (10), $S[k]$ is the result in the Fourier domain and $Z(j)$ is its form in the time domain. This yields an overall good recovery, with low-frequency harmonics restored; however, the audio exhibits a reduction in high-frequency harmonics, resulting in a perceptual characteristic commonly associated with diminished clarity and a somewhat attenuated acoustic profile. This

spectral subtraction algorithm was tested with audio signals with SNRs ranging from -40 dB to 40 dB. When the SNR was 15 dB and above, the results exhibited a substantial improvement in terms of restoration quality. In real degraded audio cases, most of the broadband noise is at a moderate or light level, which has an SNR range of around 20 dB to 40 dB[1]. The result of the degraded audio, with an SNR of 40 dB shown in Figure 9.



(a) (b)

Figure 9. The recovered speech audio degraded to have an SNR of 40 dB.

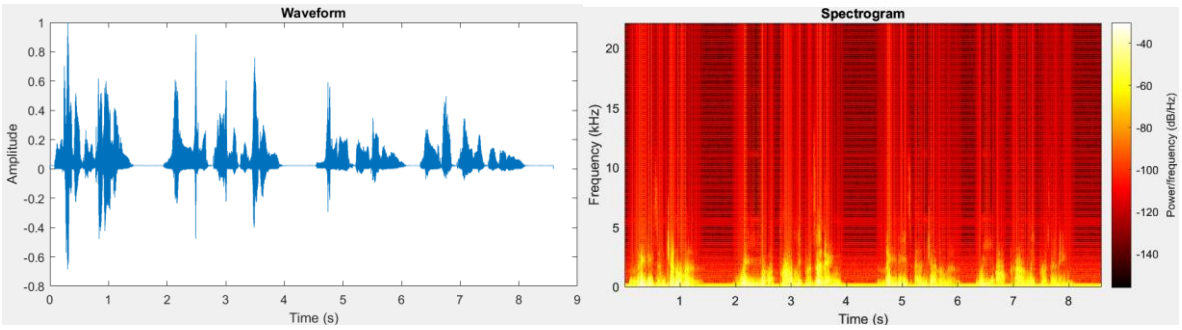
The result of spectral subtraction exhibits a pronounced restoration of critical signal elements, demonstrating a remarkable level of efficacy in mitigating the detrimental effects of degradation.

2.2.4 Wiener Filter

The Wiener Filter also operates in the frequency domain and aims to minimise the mean squared error (MSE) between the original clean signal and the filtered noisy signal[2]. The observed signal $x[n]$ is modelled as the sum of the clean signal ($s[n]$) and the noise ($v[n]$) and the frequency response $H(f)$ of the filter is computed to minimise the MSE.

$$H(f) = \frac{X(f)}{X(f)+V(f)} \tag{11}$$

In this equation, the power spectral density (PSD) of the observed signal and noise signal are used to establish the Wiener gain. The Wiener gain was applied to the segmented observed signal, and the result is shown in Figure 10.



(a) (b)

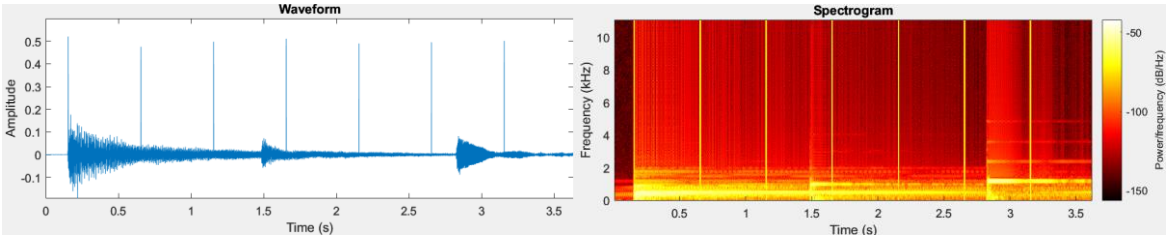
Figure 10. Recovered audio from the Wiener filter, which was degraded to have an SNR of 40 dB.

Compared to results from spectral subtraction, this recovered result exhibits a discernibly lesser degree of fidelity and restoration efficacy, but it retains more detailed low-frequency information than the spectral subtraction algorithm.

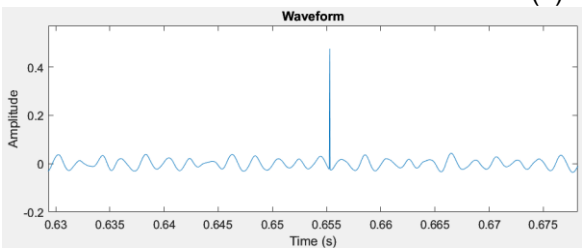
3 NONLINEAR DEGRADATIONS

3.1 Pops and Crackles

Unlike limited bandwidth or hiss noise, pops/ crackles are abrupt unwanted sounds. They are not a part of the original audio content and are introduced because of imperfections in the recording or playback process. Typically, the process of restoring audio compromised by crackles involves two fundamental stages: identification and localisation of the degradation's points of occurrence, followed by the subsequent interpolation of any absent or compromised samples. Figure 11 visually illustrates the degradation of a series of piano notes due to the presence of crackles.



(a) (b)



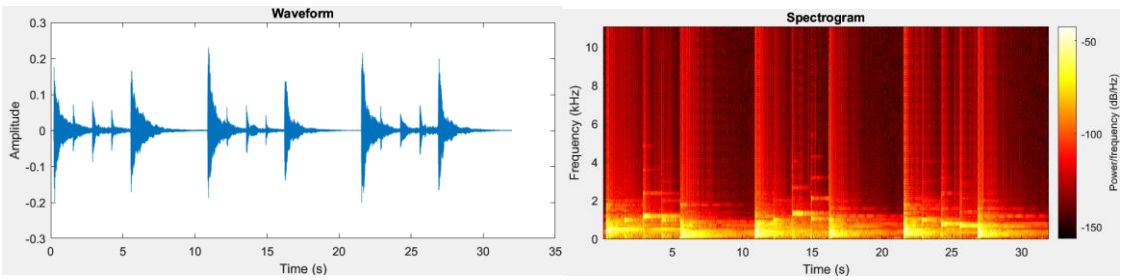
(c)

Figure 11: Piano audio degraded by crackles.

Utilising the abrupt nature of crackles, their locations can be pinpointed by analysing the derivative of amplitude changes between consecutive samples, the equation given by

$$\frac{\partial y}{\partial x} = \frac{|input[n] - input[n-1]|}{1/sampling\_rate} \tag{12}$$

A sharp change indicates a crackle's presence and central to the majority of click removal methodologies lies an interpolation scheme, wherein incomplete or impaired samples are substituted with estimations of their authentic values. Since this piano audio is consistent without abrupt alterations, the missing samples are replaced by the mean of the adjacent samples. The result of the recovered audio is shown in Figure 12.



(a) (b)

Figure 12: Recovered piano audio.

The outcome of the audio restoration process is marked by an efficient reduction in crackle degradation, yielding a discernibly enhanced listening experience. Some crackles persist during abrupt changes in musical notes, necessitating the application of advanced statistical methodologies, such as autoregressive models or moving average techniques, to address artefacts.

4 MACHINE LEARNING AND FUTURE WORK

Traditional DSP techniques often rely on pre-established rules or assumptions, utilising fixed and unchanging processing methods[7]. However, this approach may not be optimal for the dynamic nature of real-world audio signals. In contrast, ML can directly discern temporal patterns and interconnections from extensive datasets. It can adapt to fluctuations and varying situations over time, autonomously deriving pertinent features and uncovering intricate associations[8]. Essentially, ML imparts the model or algorithm with a perceptive grasp of audio recordings[9]. A noteworthy distinction that sets audio restoration apart from other tasks is its focus solely on addressing audible signals. ML also demonstrates its aptitude in intelligently selecting and processing this pertinent information. Both speech and musical audio can be represented as compositions of frequency harmonics, as depicted in the aforementioned spectrograms. Through a comparative analysis of audio recordings from contemporary devices and vintage sources, neural networks have an inherent capacity to meticulously comprehend the intricacies of the frequency domain. Consequently, the neural networks exhibit an enhanced capability to deduce more accurate estimations of imperceptible or severely compromised signals. This estimation is predicated upon an underpinning of prior ML research endeavours. Subsequent investigations will be directed towards an exploration of hybrid methodologies, integrating both DSP and ML. This synergistic approach is envisaged to perform better enhancement for vintage audio restoration.

5 REFERENCES

1. J. Sterne. The Audible Past: Cultural Origins of Sound Reproduction. New York: Duke University Press, 2003.

2. S. J.Godsill, P. J. W. Rayner. *Digital Audio Restoration*. Cambridge: Springer, 1998.

3. Signal Wizard Systems. https://www.signalwizardsystems.com (accessed Jul. 21, 2023).

4. P. Gaydecki. (2022). Digital Signal Processing [Lecture notes].

5. H. Simon. “Chapter 10 The Recursive Least-Squares (RLS) Algorithm,” in *Adaptive Filter Theory*, 5th ed., Essex, England: Pearson Education Limited, 2013, pp.449-473.

6. K. Furuya, A. Kataoka. “Robust Speech Dereverberation Using Multichannel Blind Deconvolution With Spectral Subtraction”, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.15, No.5, July 2007.

7. J. A. Moorer. “DSP Restoration Techniques for Audio”, *IEEE International Conference on Image Processing*, Vol.4, p.IV-5-IV-8, 2007.

8. H. Purwins, B. Li, et al., “Deep Learning for Audio Signal Processing”, *IEEE Journal of Selected Topics on Signal Processing*, Vol.13, No.2, May 2019.

9. J. Hou, S. Wang, et al. “Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks”, *IEEE Transactions on Emerging Topics in Computational Intelligence*, Vol.2, No.2, April 2018.