Denoising Speech Using APA Family

Shantanu Kapoor,Electrical and Computer Engineering, University of Florida, Gainesville, U.S.A

***Abstract*—This paper provides a comparison between two similar linear filtering methods Normalized Least Mean Square (NLMS) and Affine Projection Algorithm (APA) for de-noising speech. Both the algorithms were evaluated on their rate of convergence, computational complexity and echo return loss enhancement loss. The data set used for evaluation is based on speech recorded on a dual microphone setup, deteriorated by broadband noise. During the simulation, it was shown that the APA algorithm performs better than NLMS in a non-stationary environment with a faster rate of convergence but at a cost of higher computational complexity**

***Index Terms*—speech de-noising, linear filtering, adaptive filters, normalized least mean square, adaptive projection algorithms, speech enhancement.**

I. Introduction

Noise is any unwanted signal that interferes with the desired signal and the real world environment is full of it. In speech processing, noise can be any wanted sound like background noise, music, or other people talking . Such noise hinders the performance and removing it from speech signals is an imperative task in many speech processing applications, such as speech recognition, language identification, or speaker diarization. The goal of noise reduction is to remove as much noise as possible while preserving the speech signal.

Moreover, speech de-noising approaches are often divided into two main categories: single microphone and multi-microphone methods. Single microphone methods have limitations in real environments and introduce speech distortion. On the other hand, multi-microphone approaches have better spatial noise reduction, but require higher computational costs. Therefore, dual microphone approaches is preferred as a trade-off between multi-channel and single-channel methods.

In dual microphone setup as explained in [2], the first microphone measures the noisy signal (noise+speech), while the second microphone measures the environment noise. The second microphone signal is used as a noise reference to eliminate the noise in the first microphone by means of adaptive noise cancellation. In this case, noise reduction techniques rely on assumptions about the speech and noise signals.Here, the noise is additive and slowly varying, so that the noise characteristics estimated in the absence of speech can be used subsequently in the presence of speech.

The above case can be solved using various noise reduction techniques such as spectral subtraction, Wiener filtering, and adaptive filtering algorithms, or the powerful Deep Neural Networks. In recent years, Deep Neural Networks (DNN)[1] have been used often to tackle the problem of noise removal. However, these techniques are data-hungry and have large memory requirements for training. Therefore, in this paper, we will discuss the less complex linear adaptive filtering, namely Normalized Least Mean Square (NLMS) and Affine Projection Algorithm-2 (APA-2), belonging to the APA family of algorithms for dual-microphone setup.

These algorithms are explained and compared in the following subsections. Section IIdiscusses adaptive filtering and the basics of the APA algorithms. Section III discusses about the dataset. Section IV evaluates their performance based on Echo Return Loss Enhancement (ERLE), computational complexity, and misadjustment. Moreover, it also discusses the effect of step size and window length on their respective Signal-to-Noise Ratios (SNRs).

II. Adaptive Filters

An adaptive filter is a system that adjusts its parameters according to an optimization algorithm. It contains a digital filter with adjustable weight coefficients that are modified with respect to a loss function usually Mean Square Error (MSE). The adaptive filter then produces an estimate of noise y(n), which will be subtracted from the corrupted signal x(n) . The adaptive algorithm updates the filter coefficients until the loss function is minimized. The adaptive filters are simple and can adjust to the non-stationary signals.

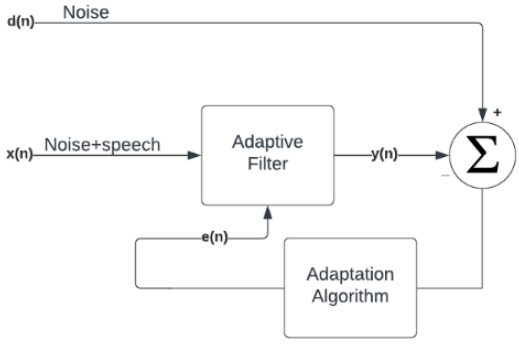

Fig. 1. adaptive filter

For speech de-nosing, the input to the adaptive filter x(n) is speech + environment noise and the reference signal d(n) is the environment noise. The adaptive filters adjust to the error signal e(n) such that it converges to desired speech signal s(n). To solve aforementioned using APA family of algorithms [4] , here are the required equations:

$$\mathbf{w}(0) = \mathbf{w}(i-1) + \eta\mathbf{X}(i)(\mathbf{X}(i)^T\mathbf{X} + \epsilon\mathbf{I})^{-1}[\mathbf{d}(i) - \mathbf{X}^T\mathbf{w}(i-1)] \quad (1)$$

$$\mathbf{X}(i) = [\mathbf{x}(i-K+1), \cdots, \mathbf{x}(i)]_{LxK} \quad (2)$$

$$d(i) = [d(i-K+1), \cdots, d(i)]_{Kx1} \quad (3)$$

Here the **x** denotes the input vector x(n) with past values and the vector **d** denotes our desired response up to the filter projection K. The adaptive weights **W** of filter length N in 1 are updated with respect to the APA - 2 method. The η denotes the learning rate while the ϵ is a hyperparameter to avoid singular matrices. If the projection order K = 1 the following equation becomes normalized least mean square (NLMS), which shows that it is a subset of APA algorithm shown below.

$$\mathbf{W}(i+1) = \mathbf{W}(i) + \eta \frac{e(n)x(n)}{x(n)^T x(n)} \tag{4}$$

III. Dataset Used

To simulate a dual microphone environment, a microphone was placed close to a working vacuum cleaner, while the other microphone was placed at a distance of 20-30cm from the speaker. The speech utterance was only muttered by a single female speaker into the microphone for the speech source. The noise of the vacuum was such that the speaker's voice was barely audible. The vacuum noise was chosen for its high frequency bandwidth which approximately simulates the white Gaussian noise. The signals were sampled at 21Khz which was adequate as the speech usually lies between 300Hz - 5000kHz. Both of the recordings were almost 4 seconds in length. The speech signals were initially peak normalized to remove any kind of unwanted gain from the signals shown below

$$x(n) = \frac{x(n) - x_{mean}}{x_{max}} \tag{5}$$

IV. Experiment and Results

First of all, both NLMS and APA-2 were analysed based on their step size and filter order [5]. To compare the performance of both the filters Echo Return Loss Enhancement (ERLE) metric was used which is value of the ratio between the input and the error component.

$$ERLE = 10log(\frac{E[X^2]}{E[e^2]} \tag{6}$$

In our experiment, ERLE proved to be an efficient method for choosing and comparing filters as it focused on the major changes in speech quality that arose during simulation. Moreover, the filters were also compared based on their misadjustment and computational complexity.

A. Analysing 2 Tap filters

Initially both the algorithms were first implemented using their 2 tap filters.For NLMS, it took more number of samples to model the noise. This is prevalent as the $2^n d$ filter weight continued to converge until the last second as show in figure 2. Also, due to non-stationary property of speech we can see the changes present in the learning curve as well as weight curve. There is also a sudden change in the last second shown in figure 3, observed in all other cases , which might show sudden change in either speech signal or noise source. The nlms with 2 tap filter produced an ERLE of around 8db. This 2 tap filter was trained with $\eta = 0.003$ which was later found in the subsectionIV-C

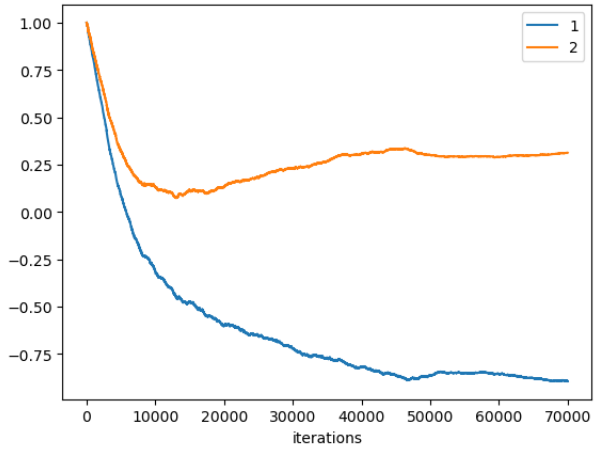

Fig. 2. change in weights with time step for nlms

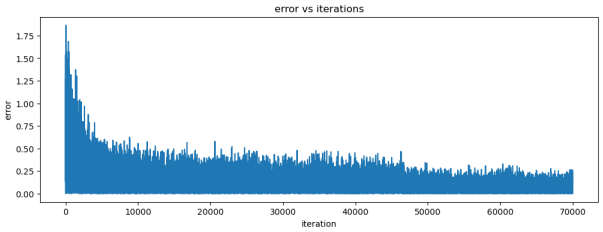

Fig. 3. error for nlms with 2 taps

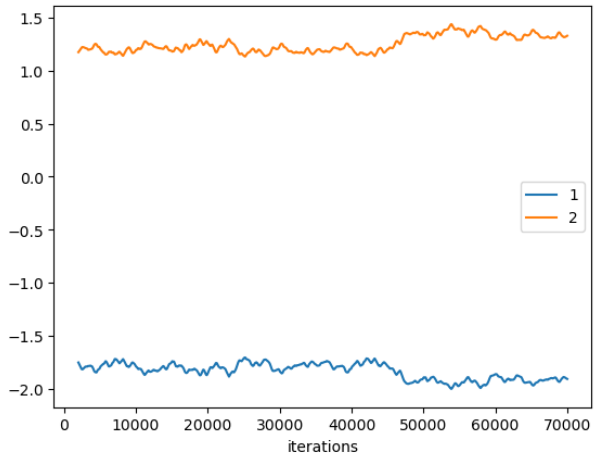

Fig. 4. wts for 2 tap delay apa filter

For the APA2, the weights converged quickly as expected due to their projection length. Here the projection length was taken 511 as each phoneme lies under 20-40ms and during this we assume the statistics of the noise and speeech both stationary. Here also we can see the sudden change in weights that occur around 2 sec prevelant in both weights Figure 4

and error 5. The $\eta = 0.001$ was chosen as acquired in the subsection IV-C. The respective ERLE loss was 9.12dB.

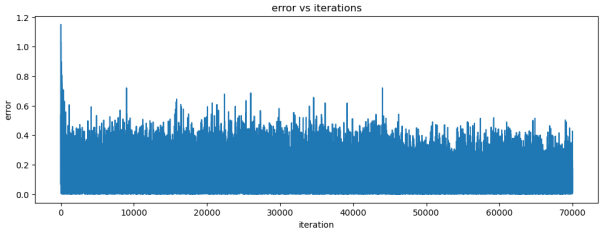

Fig. 5. error for apa-2 with 2 taps

B. Effect of Fitler Length

The filter played an important role for both NLMS and APA-2 algorithm. Both showed improvement in their performance with increment in the filter length. However, the increase in the filter length and higher ELRE loss does not always mean good speech intelligibility. This is because with higher filter length the model start to model speech signal along with the noise. Thus , the error signal e(n) contains more distorted speech with better speech-denoising. With this trade off, we choose filter order of length 10 and length 7 respectively for nlms and APA-2.

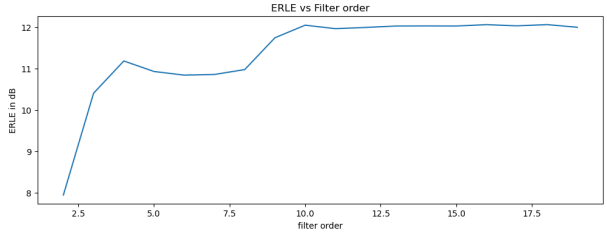

Fig. 6. ERLE vs fitler order for nlms

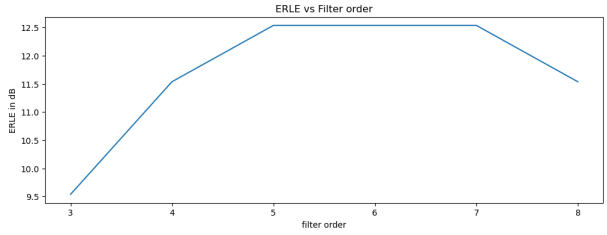

Fig. 7. ERLE vs filter order for apa

C. Effect of learning rate

The learning rate plays a very important role for convergence. If too large the model converges faster with higher misadjustment. If too low, it doesn't fully converge due to not sufficient number of samples. For both the models, NLMS with filter order 5 and APA with filter order 2 were used to evaluate the performance with change in step size. For the nlms the best $\eta = 0.003$ and for apa $\eta = 0.001$, both plots are shown below.

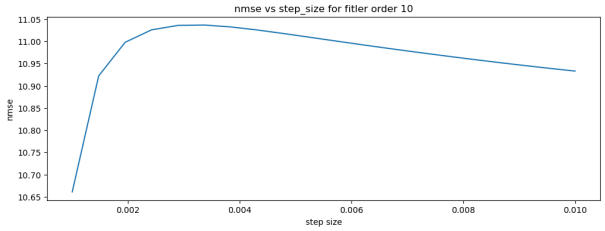

Fig. 8. ERLE vs step size for NLMS

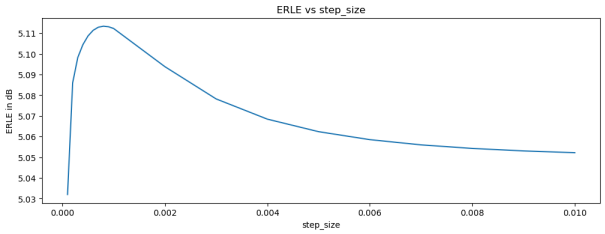

Fig. 9. ERLE vs step size for APA

D. Computational Complexity

The NLMS is not memory intensive. The computational complexity requried by the NLMS is O(K) where K is the filter order. Hence, it is easy to implement. However, it comes with a trade off. It works better mostly for signals that are stationary as it requires more number of samples for convergence. This was seen in the error figure 3 where after sudden change in signal it redirects weight but takes time due to slow convergence rate.

Meanwhile, the computational complexity of the APA 2 algorithm is $O(L^2)$, where L is the projection length of the filter. This memory helps consider the signal as stationary inside the projection length and compute the desired response. This also helps APA-2 to readily change and immediately converge with lower misadjustment [6].

V. Conclusion

According to the experiment, we can conclude that both NLMS and APA algorithms can work for denoising speech signals in a dual microphone setup. The primary loss function, ERLE, is useful in determining the approximate filter length for the algorithm. However, a too big value of ERLE can be counterproductive as the denoised speech comes at the cost of speech intelligibility. Moreover, one should also be considerate of the filter length and filter step size as they both affect the adaptive model significantly. Furthermore, the filter step size plays a crucial role in faster convergence rate, which is important for the lower misadjustment value. The APA algorithm has faster convergence rate and better misadjustment compared to NLMS. However, it comes at an exponential computational cost. Hence, the tradeoff between algorithms must be decided carefully for real-time applications

REFERENCES

[1] Jingdong Chen, J. Benesty, Yiteng Huang and S. Doclo, "New insights into the noise reduction Wiener filter," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 1218-1234, July 2006, doi: 10.1109/TSA.2005.860851.

[2] S. Boll and D. Pulsipher, "Suppression of acoustic noise in speech using two microphone adaptive noise cancellation," in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 6, pp. 752-753, December 1980, doi: 10.1109/TASSP.1980.1163472.

[3] arXiv:1706.07162

[4] S. L. Gay and S. Tavathia, "The fast affine projection algorithm," 1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, 1995, pp. 3023-3026 vol.5, doi: 10.1109/ICASSP.1995.479482.

[5] arXiv:1106.0846 [cs.SD]

[6] T. Shao, Y. R. Zheng and J. Benesty, "An Affine Projection Sign Algorithm Robust Against Impulsive Interferences," in IEEE Signal Processing Letters, vol. 17, no. 4, pp. 327-330, April 2010, doi: 10.1109/LSP.2010.2040203.Science, 1989.

VI. ADDITIONAL DIAGRAMS

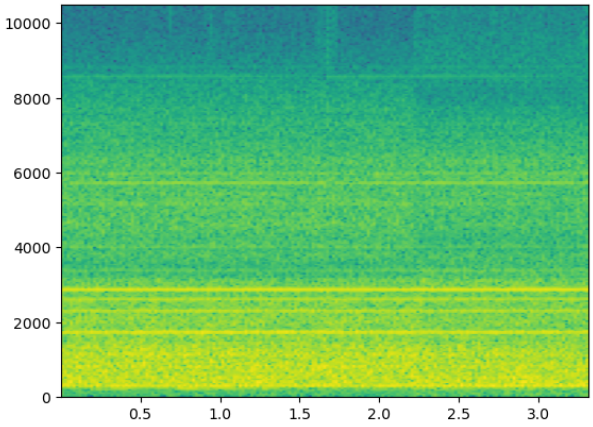

Fig. 10. Input speech + noise spectrogram

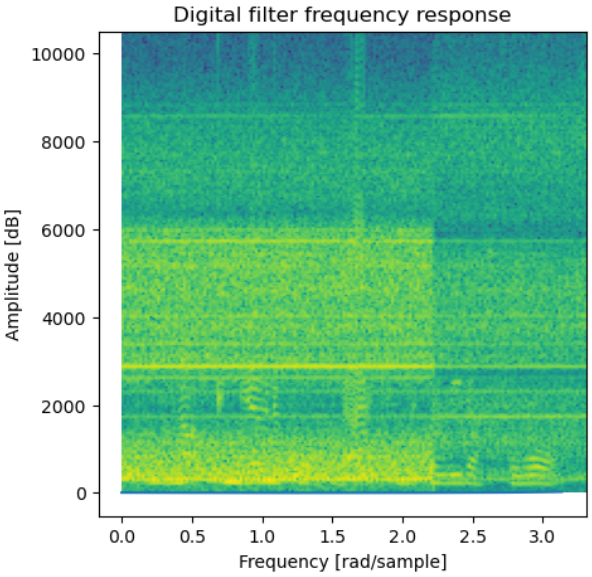

Fig. 11. speech spectrogram after filtering

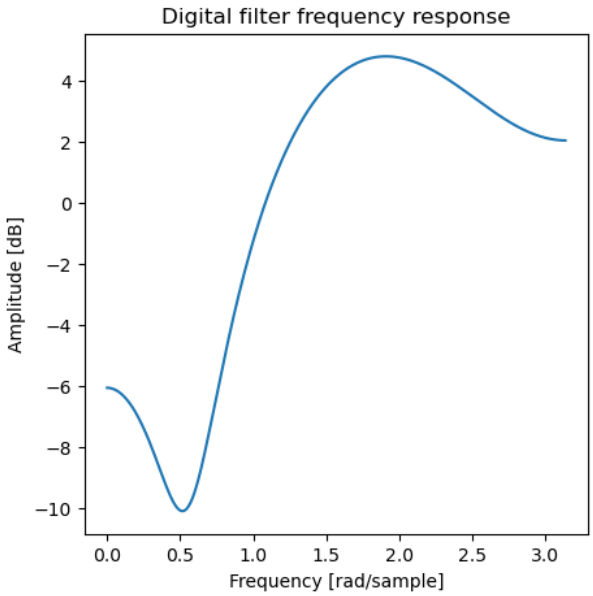

Fig. 12. best fitler response for apa

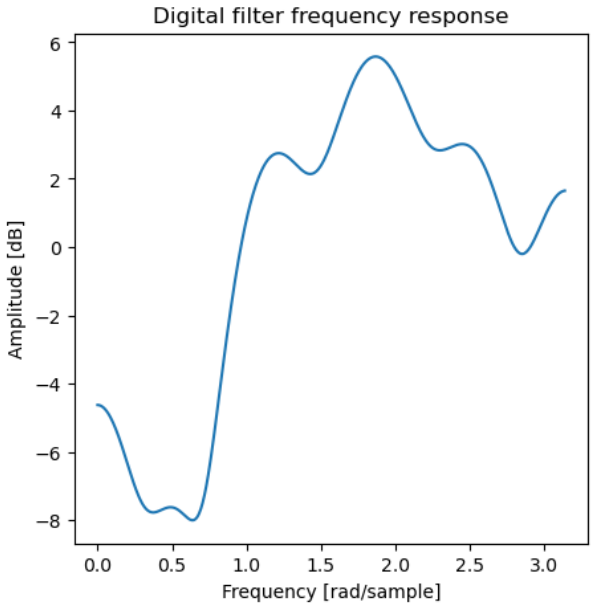

Fig. 13. best fitler response for NLMS