Speech Enhancement of Single-Source Audio Signal Using Covariance Matrix Adaptation Evolution Strategy

Vincent Fasburg

Department of Computer Science  
Michigan State University  
East Lansing, MI. USA

fasburgv@msu.edu

Joshua Thomas

Department of Computer Science  
Michigan State University  
East Lansing, MI. USA

thom1212@msu.edu

**ABSTRACT**

A Covariance Matrix Adaptation Evolution Strategy (CMA-ES) is a method for efficiently searching through a landscape of multiple real-valued parameters to determine which parameter values give the best solution to a known problem. This paper explores the application of CMA-ES to the problem of speech enhancement, that is, recovering the speech signal from a single microphone signal composed of both speech and an unwanted background noise. This research makes use of some of the most popular techniques for speech enhancement, and seeks to find an optimal way to combine these techniques to best recover the original clean speech signal. The results show that the combination of these methods, with parameters tuned using CMA-ES, can outperform the component methods individually, with parameters tuned by experts as would commonly be done.

**Keywords**

Evolutionary Strategy; CMA-ES; Speech Enhancement; Audio;

# INTRODUCTION

The field of speech enhancement aims to apply transformations to a digital signal containing both spoken words as well as some type of unwanted noise, and recover the “clean” version of the signal containing speech alone. This is generally done to improve the intelligibility of the speech, either to be listened to by a human, or processed by a voice recognition system.

As mobile and hands-free technology progresses, cell phones are processing voice input not only for phone calls, but as an interface to many other device features. Voice commands are especially likely to be used while the user is driving a car, when it is often unsafe or illegal to use the traditional interface. For this reason a clip of speech corrupted by road noise was used for this research.

In order to produce the cleanest possible speech signal, 3 common methods of speech enhancements are combined using parameters determined using CMA-ES. These methods are Wiener Filtering, Spectral Subtraction, and Noise Gating, which have a combined total of 8 parameters to be tuned by the evolutionary strategy.

# EXPERIMENTAL SETUP

This section describes the component functions which make up this combined speech enhancement solution, as well as the parameters within these functions that were tuned using the CMA-ES to provide the best enhancement. Finally, the CMA-ES structure that is used for this optimization is described.

The audio used in this experiment follows the mathematical formula in equation 1. In this equation, x(t) represents the clean audio signal that we are attempting to recover, h(t) is the impulse response of the linear time-invariant system representing the environment of the speech (in this case the acoustic properties of a car), and n(t) represents the noise added to this clean signal. The output of this equation is the observed noisy signal s(t).

Equation

For the purposes of noise cancellation in this experiment, it is not important to model the impulse response of the acoustics of the car, so we can assume that h(t) ≈ 1.

## Wiener Filter

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

The Wiener Filter is a classical technique first proposed by Norbert Wiener in 1949[1]. The Wiener Filter is used to estimate a signal of interest through the application of a linear time-invariant filter to a signal consisting of the signal of interest plus additive noise. Although the Wiener filter can be implemented in different ways, with both causal and non-causal forms, the version used in this project is a causal FIR filter, allowing processing to be done in real time, such as during a telephone call. The gains for this filter are determined by taking a noise-only sample of the signal and deconvolving this with the noisy signal. Feeding the original noisy signal through this filter results in an estimate of the clean speech signal.

Equation 2 shows the calculation of a Wiener filter in the frequency domain G(f), where H(f), N(f), and S(f), are the frequency spectra of h(t), n(t), and s(t) from equation 1 respectively.

Equation

The Wiener Filter has been expanded and improved many times since its original introduction. The specific version that was used in this project includes a two-step noise reduction technique by [3] that uses a decision directed approach to minimize reverberation noise artifacts left over by the original Wiener Filter. In order to further reduce such artifacts, smoothing is used to make the signal to noise ratio (SNR) estimate more constant over time.

## Spectral Subtraction

Spectral subtraction is a method of noise cancellation which uses the power spectrum of the underlying signal noise to estimate that of the clean audio signal. This method relies on the assumption that the frequency content of the additive noise present in a signal is roughly constant over time. For our main application of road noise in a car, this is a fairly accurate assumption over short periods of time (after which the noise spectrum can be recalculated).

The simplest form of the spectral subtraction technique requires converting a section of audio that is known to contain only noise into the frequency domain using the Fourier transform. Since the assumed constant noise will have the same frequency content over time, these values can be subtracted from the frequency domain version of the observed signal. This is performed for a few hundred time samples at a time, and then combined using the overlap-add method to recombine the processed time samples into a continuous string of samples.

This simple idea has been expanded many times to improve the quality of the resulting signal. The version of the spectral subtraction algorithm used in this experiment is based on an implementation by Esfandiar Zavarehei (2005) of spectral subtraction with residual noise reduction from [4]. This enhanced version of the spectral subtraction algorithm makes a binary decision if each frequency component primarily contains noise by comparing the spectral magnitude difference of the signal and noise to a pre-determined threshold known as the noise margin. Additional parameters called “hang-over” and “noise length” determine how long the signal should be attenuated when noise is detected.

## Noise Gate

Noise gating is a process by which, when the amplitude of an audio signal is below a pre-set threshold, the amplitude is further reduced to a very low level, often to 0 or to a level of inaudibility. Once the amplitude of the input audio signal exceeds the threshold, the output returns to match the input signal.

As a simplification of the behavior shown in Figure 1, the attack and release times are fixed to match each other, and the hold time is set to 0. The only parameters that were varied in this study are the attack/release and the gate threshold.

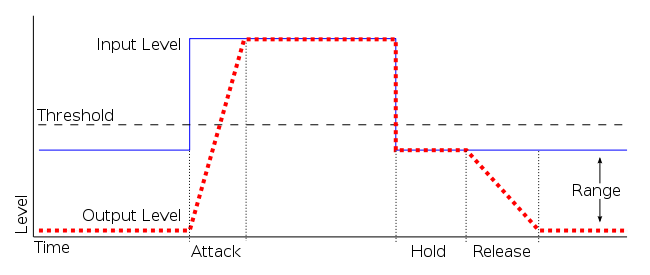


Figure 1: Noise Gate [5]

## CMA-ES

The use of Covariance Matrix Adaptation in Evolutionary Strategies has been shown to be a very effective means of optimizing real-valued parameters.[6] As described in Sections 2.1 – 2.3, there are a total of 8 parameters to be tuned using the CMA-ES, which are shown in Table 1. Each parameter in the CMA-ES is limited to values in [0,1] so that a constant starting learning step size value of σ = 0.2 can be used for all parameters, allowing successive mutations to easily cover the full range of possible values. These [0, 1] values are then scaled into the ranges shown in Table 1 for each parameter. These values are based on some prior knowledge of what values would be reasonable for the parameters, while still providing a large enough range for the evolutionary strategy to find any value that could conceivably have a good result.

Table 1: Parameter List

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Component** | **Parameter** | **Symbol** | **Min** | **Max** |
| Wiener Filter | Gain | GW | -2 | 2 |
| Wiener Filter | Smoothing Factor | α | 0 | 2 |
| Spectral Subtraction | Gain | GS | -2 | 2 |
| Spectral Subtraction | Noise Margin | N | 0 | 20 |
| Spectral Subtraction | Noise Length | L | 0 | 20 |
| Spectral Subtraction | Hang Over | H | 0 | 20 |
| Noise Gate | Threshold | T | 0 | 0.25 |
| Noise Gate | Attack/Release | A/R | 0 | 10 |

For this task, we used a (μ, λ) evolutionary strategy with weighted recombination, such that the mean for each successive generation is skewed toward the values of the most fit individuals in the current generation.

The final estimation of the clean speech signal is produced by applying the Wiener filter and spectral subtraction in parallel, and passing the summation of the two through the noise gate. This system is shown in Figure 2.

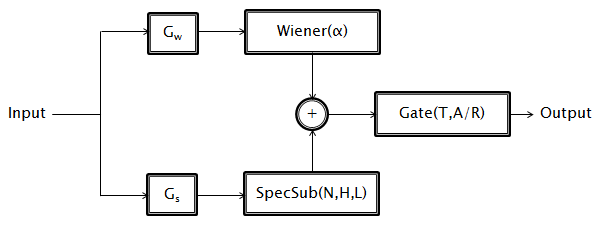
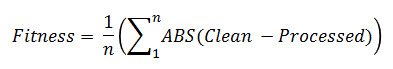


Figure : Audio Processing

The output audio resulting from this process is an estimate of the clean speech signal. This output will become closer to the reference clean signal as the evolutionary strategy progresses. The fitness function that was used to determine the similarity between the processed audio and the clean audio signal is shown in Equation 3. Note that the signal “clean” is known only because the noisy audio signal was manufactured, thus the fitness function in this form would not be suitable for a live application in which parameters were tuned during use.



Equation

From this function, it is clear that fitness is to be minimized by the CMA-ES. The ideal fitness is 0, as this would mean that the processed and clean signals were exactly the same.

# EXPERIMENTAL RESULTS

In this section, we look at the results of using the CMA-ES to determine the best values for the 8 parameters, described in Section 2, which will reduce the noise in an audio file.

## Experimental Setup

There were 3 different population types used with CMA-ES by varying the parent (μ) and offspring (λ) population. The parent populations were 6, 12, and 24 which correspond to the offspring populations of 12, 24, and 48 respectively. Each population type was run 50 times with 300 generations per run. The following sections will analyze the results of these runs.

## Fitness

As described in Section 2.4, the fitness is a measure of the difference between the clean audio file and the processed audio file. As such, a lower fitness is more desirable since it represents less of a difference from the clean audio file. Figure 3 shows the average fitness of each generation for each population type (offspring size 12, 24, and 48). The fitness for all population types begins to level off at around generation 60. This trend continues through the final 300th generation.

Figure 3 also shows that as the offspring size is increased the converged upon fitness value gets better. This is expected since as the population increases a larger initial sample of the search space is available. Therefore, there is a high chance that something in that population will lead down a path that results in a better fitness. When the offspring size is increased from 12 to 24 there is a very noticeable improvement to the fitness. Then again as the offspring size is increased from 24 to 48 the fitness is again improved, but by a small factor than before.

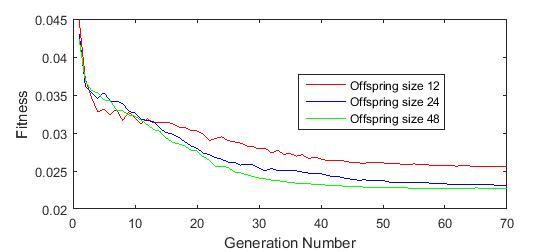


Figure 3: Average Fitness for Each Generation

Figure 4, Figure 5, and Figure 6 show the average fitness of each generation with 95% confidence intervals added in. Overall, this shows that the error bars in the confidence intervals are quite small. This indicates that the offspring currently in each generation generally exhibit similar fitness values. In addition, we can see that as the number of generations increases the error bars get smaller which indicates that population is becoming less diverse. This is expected as the CMA-ES begins to converge of the best parameters.

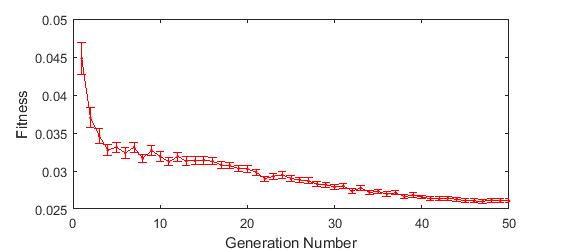
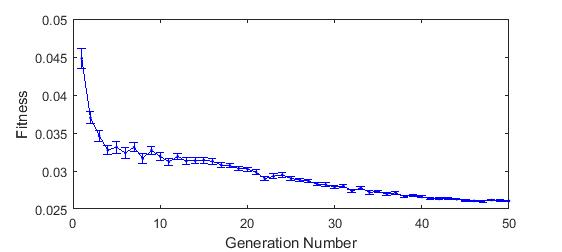


Figure 4: Confidence Interval of Fitness for Offspring Size 12

Figure 5: Confidence Interval of Fitness for Offspring Size 24

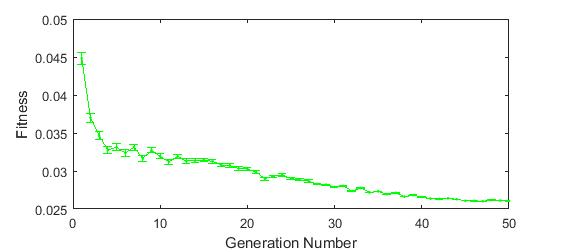


Figure 6: Confidence Interval of Fitness for Offspring Size 48

## Expert Results

## Best Parameter Values

## Weiner Filter with Best Values

## Spectral Subtraction with Best Values

# CONCLUSIONS

# ACKNOWLEDGMENTS

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

# REFERENCES

1. Wiener, Norbert. 1949. Extrapolation, Interpolation, and Smoothing of Stationary Time Series. New York: Wiley. ISBN 0-262-73005-7.
2. Huijun Ding, Ing Yann Soon, Soo Nee Koh, Chai Kiat Yeo, A spectral filtering method based on hybrid wiener filters for speech enhancement. In *Speech Communication,* Volume 51, Issue 3, March 2009, Pages 259-267, ISSN 0167-6393, http://dx.doi.org/10.1016/j.specom.2008.09.003.
3. Plapous, C.; Marro, C.; Scalart, P., "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, Issue 6, pp. 2098 - 2108, Nov. 2006Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.
4. Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing. ASSP-27 (2).
5. Public Domain, <https://commons.wikimedia.org/wiki/> File:Noise\_Gate\_Attack\_Hold\_Release.svg
6. Hansen, N. The CMA Evolution Strategy: A Tutorial. November 24, 2010. <http://citeseerx.ist.psu.edu/viewdoc/> download?doi=10.1.1.139.7369&rep=rep1&type=pdf

Columns on Last Page Should Be Made As Close As Possible to Equal Length