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 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.

**CCS Concepts**

• **Information systems➝Database management system engines**   • **Computing methodologies➝Massively parallel and high-performance simulations.**This is just an example, please use the correct category and subject descriptors for your submission*.* The ACM Computing Classification Scheme:

<http://www.acm.org/about/class/class/2012>. Please read the [HOW TO CLASSIFY WORKS USING ACM'S COMPUTING CLASSIFICATION SYSTEM](http://www.acm.org/publications/article-templates/CCS-HOWTO-v6-12Jan2015.docx) for instructions on how to classify your document using the 2012 ACM Computing Classification System and insert the index terms into your Microsoft Word source file.

**Keywords**

Keywords are your own designated keywords separated by semicolons (“;”).

# 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 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.

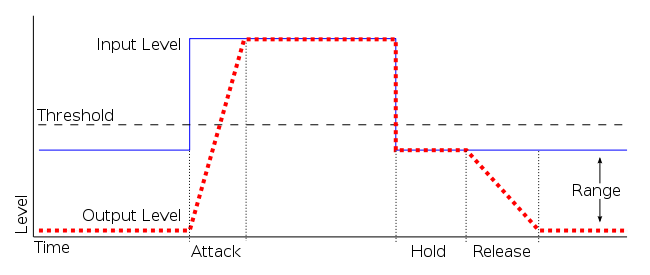


Figure : Noise Gate [5]

## 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 the 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |

Table 1: Parameter List

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.

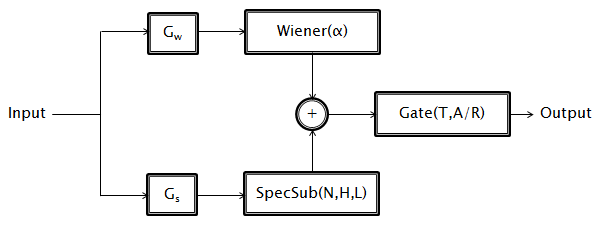
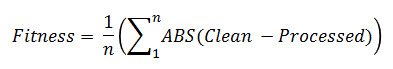


Figure 2: Audio Processing

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 3

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.

The

## Page Numbering, Headers and Footers

Do not include headers, footers or page numbers in your submission. These will be added when the publications are assembled.

# FIGURES/CAPTIONS

Place Tables/Figures/Images in text as close to the reference as possible (see Figure 1). It may extend across both columns to a maximum width of 17.78 cm (7”).

Captions should be Times New Roman 9-point bold. They should be numbered (e.g., “Table 1” or “Figure 2”), please note that the word for Table and Figure are spelled out. Figure’s captions should be centered beneath the image or picture, and Table captions should be centered above the table body.

# SECTIONS

The heading of a section should be in Times New Roman 12-point bold in all-capitals flush left with an additional 6-points of white space above the section head. Sections and subsequent sub- sections should be numbered and flush left. For a section head and a subsection head together (such as Section 3 and subsection 3.1), use no additional space above the subsection head.

## Subsections

The heading of subsections should be in Times New Roman 12-point bold with only the initial letters capitalized. (Note: For subsections and subsubsections, a word like *the* or *a* is not capitalized unless it is the first word of the header.)

### Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized and 6-points of white space above the subsubsection head.

#### Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

#### Subsubsections

The heading for subsubsections should be in Times New Roman 11-point italic with initial letters capitalized.

# 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