Reduction of Additive Noise in the Digital Processing of Speech

Mid Year Progress Report

Avner Halevy

Department of Mathematics

University of Maryland, College Park

ahalevy at math.umd.edu

Professor Radu Balan

Department of Mathematics

Center for Scientific Computation and Mathematical Modeling (CSCAMM)

University of Maryland, College Park

rvbalan at math.umd.edu

Abstract

This project implements a few standard algorithms for reducing additive white noise in the processing of speech signals. Among these are spectral subtraction and iterative Wiener filtering. The performance of the algorithms is evaluated using TIMIT, a database of phonetically rich sentences, widely used in the industry in the development of speech processing algorithms. Initially only objective measures are used to evaluate the quality of processed speech, but if time allows, evaluation will be extended to include subjective listening tests as well.

Background

The need to enhance speech signals arises in many situations in which the speech signal originates from a noisy location or is degraded by noise over a communication channel. Speech enhancement algorithms can be used to enhance both quality and intelligibility of speech signals, thus making communication more effective and reducing listener fatigue. The precise goals of speech enhancement algorithms depend on the specific application, and the specific type of noise involved, as well as its statistical relation to the clean signal. The main challenge in designing effective speech enhancement algorithms is reducing noise without introducing perceptible distortion to the speech signal.

This project focuses on the reduction of additive white Gaussian noise which is uncorrelated with the clean speech signal. We are assuming that y(n), the noisy signal, is composed of the clean speech signal x(n), and the noise d(n), i.e. $y(n) = x(n) + d(n)$.

Progress

In accordance with the timeline set forth in the project proposal, so far I have focused on implementing spectral subtraction (including some variations) as well as some initial testing. A description of the essential parts of the algorithm follows.

First, since speech signals are highly non stationary, a short time Fourier transform is used to analyze the signal, with a Hamming window spanning 20 msec (or 320 samples at a sampling rate of 16kHz). A times step of ¼ the window length (80 samples) is used to advance from one frame to the next. The overlap and add method is used to synthesize the enhanced signal once processing has been completed.

The first five frames of the noisy signal are assumed to be noise only. The noise power spectrum is averaged over these frames to obtain an estimate which is used for the rest of the algorithm.

The heart of the algorithm consists of subtracting the estimate of the noise magnitude from the magnitude of the noisy signal spectrum, to recover (an estimate) of the magnitude of the clean signal spectrum. Due to fluctuations in the noise spectrum, this subtraction may result in negative values, in which case the most basic approach is to set these values to zero. The magnitude estimate is then combined with the noisy phase, an approximation (of the clean phase) which has been shown to be good enough for practical purposes. Symbolically, the algorithm is described as follows:

$$y(n) = x(n) + d(n)$$

$$Y(\omega) = X(\omega) + D(\omega)$$

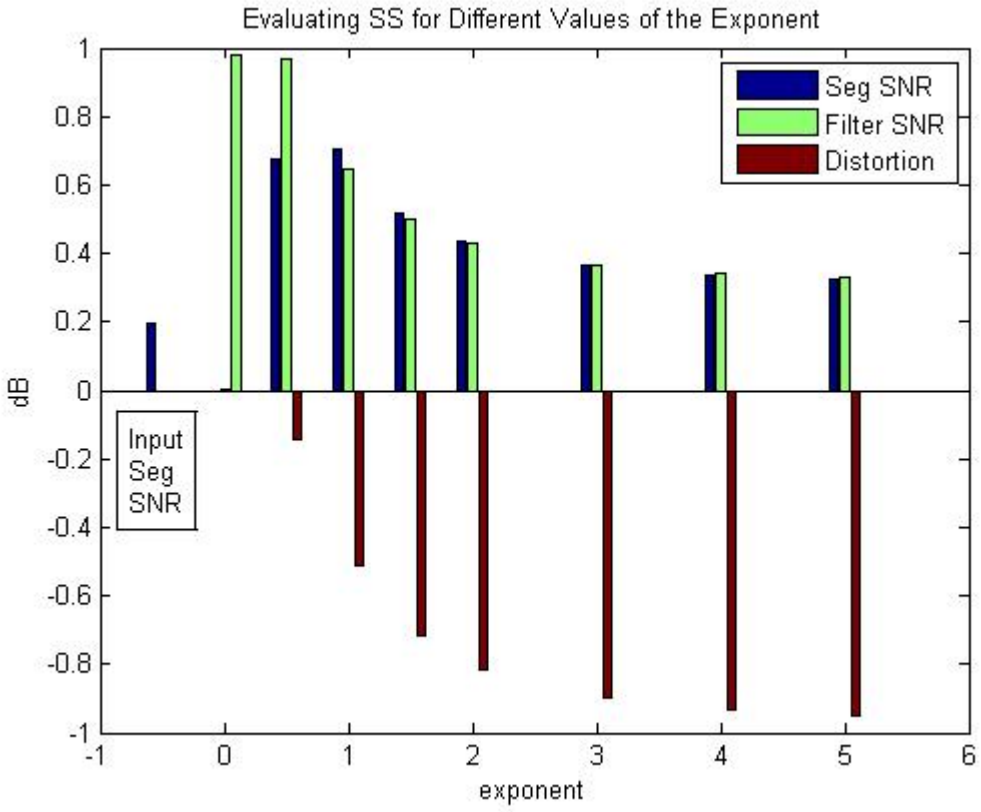
$$Y(\omega) = \left|Y(\omega)\right| e^{i\phi_y(\omega)}$$

$$\left|\hat{X}(w)\right| = \begin{cases} \left(\left|Y(\omega)\right|^p - \left|\hat{D}(\omega)\right|^p\right)^{1/p} & \text{if } \left|Y(\omega)\right|^p - \left|\hat{D}(\omega)\right|^p \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{X}(\omega) = \left|\hat{X}(\omega)\right| e^{i\phi_y(\omega)}$$

$$\hat{x}(n) = \textit{Inverse Fourier}\,\{\hat{X}(\omega)\}$$

There is freedom in choosing the exponent p above, and various values in the range .1 – 5 have been experimented with, with the results summarized below:



These results suggest that as p increases, less noise is subtracted from the noisy signal, and thus less distortion is introduced. Conversely, as p decreases to zero, the signal is progressively annihilated. Preliminary subjective evaluation (listening) suggests that values in the range .5 - 2 offer the best compromise.

To gain further insight into the algorithm, as well as access to objective quality measures described below, we may view the subtraction as a “filtering” process:

$$\left|\hat{X}(\omega)\right| = (\left|Y(\omega)\right|^p - \left|\hat{D}(\omega)\right|^p)^{1/p}$$

$$\left|\hat{X}(\omega)\right| = \left[1 - \left(\frac{\left|\hat{D}(\omega)\right|}{\left|Y(\omega)\right|}\right)^p\right]^{1/p} \left|Y(\omega)\right|$$

$$\left|\hat{X}(\omega)\right| = H(\omega)\left|Y(\omega)\right| \quad \text{where}$$

$$H(\omega) = \left[1 - \left(\left(\frac{\left|Y(\omega)\right|}{\left|\hat{D}(\omega)\right|}\right)^{-1}\right)^p\right]^{1/p}$$

Using this view we see that the lower the SNR, the more the noisy magnitude is attenuated.

The main challenge in spectral subtraction is dealing with the so called “musical noise” artifacts which result from the flooring of negative components. One approach is to use an over subtraction coefficient for the noise spectrum. I have experimented with this approach (though only in a nonsystematic way), but the results have not been convincing.

Validation and Testing

A basic validation of the spectral subtraction algorithm implementation was done by setting the estimate of the noise spectrum magnitude equal to zero, in which case the output (enhanced) signal was identical to the input (noisy) signal, as desired.

Initial testing of the algorithm was performed using clean signals from TIMIT, a database of phonetically rich sentences spoken in 8 different dialects, widely used in the industry in the development of speech processing algorithms. White Gaussian noise was added using MATLAB’s **randn** function.

Evaluation of the algorithm has been mostly limited to the use of objective measures. The first measure, called *segmental SNR*, considers the energy of the clean signal as compared with the energy of the error in estimation, and is defined as

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}x^2(n)}{\sum_{n=Nm}^{Nm+N-1}(x(n)-\hat{x}(n))^2}$$

where *x(n)* is the clean signal, $\hat{x}(n)$ is the enhanced signal, *N* is the frame length, and *M* is the number of frames. In this, as well as in the SNR measure described next, lower and upper thresholds are used to prevent the final measure from being distorted by atypical frames: if the frame value computed is negative, it is set to 0, and if it is higher than 35, it is set to 35.

Two other measures use the "filter" *H* discussed above. Once again an average over all frames is computed, but the quantity computed in each frame depends on *H*. If we denote by $\tilde{x}$ the time domain result of applying *H* to x (the clean signal) and by $\tilde{d}$ the analogous quantity for the noise signal d, then the first measure, which we call *Filter SNR*, considers the energy of $\tilde{x}$ compared with the energy of $\tilde{d}$ and is defined as

$$Filter\ SNR = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}\tilde{x}^2(n)}{\sum_{n=Nm}^{Nm+N-1}\tilde{d}^2(n)}$$

The second measure, which we call *Distortion*, considers the energy of $x-\tilde{x}$ compared with the energy of x, and is defined as

$$Distortion = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}(x(n)-\hat{x}(n))^2}{\sum_{n=Nm}^{Nm+N-1}x^2(n)}$$

Future Timeline

January

Winter break

February

Implementation of iterative Wiener filtering

March

Testing, modification, finalization of code, comparison of algorithms

April

Preparation of final report and presentation

May

Delivery of final report and presentation

Bibliography

[1] Deller, J., Hansen, J., and Proakis, J. (2000) *Discrete Time Processing of Speech Signals*, New York, NY: Institute of Electrical and Electronics Engineers

[2] Quatieri, T. (2002) *Discrete Time Speech Signal Processing*, Upper Saddle River, NJ: Prentice Hall

[3] Loizou, P. (2007) *Speech Enhancement: Theory and Practice*, Boca Raton, FL: Taylor & Francis Group

[4] Rabiner, L., Schafer, R. (1978) *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice Hall