

# Noise Reduction and Increased VAD Accuracy Using Spectral Subtraction

Jeffery J. Faneuff  
Bose Corporation  
The Mountain

Framingham, MA 01701  
(508) 766-6048

Jeff\_Faneuff@bose.com

D. Richard Brown III  
Worcester Polytechnic Institute  
100 Institute Rd  
Worcester, MA 01609  
(508) 831-5000

drb@ece.wpi.edu

## ABSTRACT

This paper shows that performing voice activity detection (VAD) on the output of a spectral subtraction noised reduced signal increases the accuracy of the VAD and reduces the VAD sensitivity to fixed thresholds. An initial VAD decision is used to control the noise estimate update in the spectral subtraction algorithm. The more accurate VAD after the first spectral subtraction is then used to reprocess the original noisy speech again via spectral subtraction to reduce the noise while not attenuating the speech. Auditory masking thresholds were used to weight the spectral subtraction to avoid the introduction of musical noise artifacts.

Energy thresholds were used to detect voiced frames of speech recorded inside a car at an 8kHz sampling rate and combined with four different noise conditions. The received noise and the speech were combined to produce inputs to the algorithm at 0, 5, and 10 dB SNR where it was shown that the VAD accuracy consistently increased after spectral subtraction. However, if the VAD and spectral subtraction were iterated more than twice on the signal, then the VAD accuracy started to decrease. Visual inspection of the clean speech was used to determine which frames should be classified as voice and used to determine the accuracy of the VAD algorithm. The VAD accuracy was only increased by a few percent in each case, but this small improvement makes a big difference when using the resulting decisions to control the noise estimate of the spectral subtraction algorithm in order to avoid attenuating the speech. Modifications of the fixed offset for detecting voice had less of an effect when the VAD operated on the signal after spectral subtraction and compared to VAD processing on the original signal, which can be attributed to the reduced variance in the noise. Objective speech quality measures show that the algorithm removes a large amount of the stationary noise in a hands-free environment of an automobile with relatively minimal speech distortion.

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.

*ISPC '03*, March 31-April 3, 2003, Dallas, TX.

Copyright 2003 ACM 1-58113-000-0/00/0000...\$5.00.

## Categories and Subject Descriptors

[Speech/Audio]: Noise reduction – voice activity detection, spectral subtraction, perceptual weighting, speech enhancement.

## General Terms

Algorithms, Design, Theory

## Keywords

VAD, Spectral Subtraction, Noise cancellation

## 1. INTRODUCTION

Applications of VAD include speech recognition, voice compression, noise estimation/cancellation, and echo cancellation. [2] The accuracy of the VAD has a large impact on the performance of the algorithms that depend on the VAD decisions, thus many approaches have been developed including energy level detection, zero crossing rates, periodicity, LPC distance, spectral energy distribution, timing, pitch, zero crossings, cepstral features, and adaptive noise modeling. An important consideration for the VAD algorithm is the processing power required. This paper shows that an initial VAD decision and spectral subtraction can be used to produce a more accurate VAD for the purposes of noise reduction. The VAD and spectral subtraction algorithm are presented in sections 2 & 3 respectively, the measurements and simulation are described in section 4, and finally the iteration scheme to improve the VAD and results are covered in section 5.

## 2. Voice Activity Detection (VAD)

The purpose of Voice Activity Detection (VAD) is to determine whether a frame of the captured signal represents voiced, unvoiced, or silent data. Voiced sounds are periodic in nature and tend to contain more energy than unvoiced sounds, while unvoiced sounds are more noise-like and have more energy than silence. Silence has the least amount of energy and is a representation of the background noise of the environment. The VAD plays a central role in spectral subtraction algorithms because its accuracy dramatically affects the noise suppression level and amount of speech distortion that occurs. The noise estimate in spectral subtraction uses the VAD to decide when to update the noise reference in the absence of speech.

The energy level detection VAD algorithm used in this paper is described below. The initial noise spectrum, mean, and variance

are calculated assuming the first 10 frames are noise only. Thresholds are calculated for speech and noise decisions and all statistics are gradually updated when a noise frame is detected. The update factors  $\alpha$  and  $\beta$  can be tuned and have been set to 0.95 as in work done by Virag [8]. Other research has extended the energy calculation VAD to dual and multiple spectral sub-bands within each frame.

The first step of the algorithm is to buffer the data into the  $k^{th}$  frame,  $x(n, k)$ , and transform it into the frequency domain.

$$X(w, k) = FFT(x(n, k)) \quad (2.1)$$

Next, the noise spectrum and noise mean for  $k=1$  are initialized.

$$N(w) = X(w, k) \quad (2.2)$$

$$\mu_N = \frac{1}{L} \sum_{w=0}^{L-1} N(w) \quad (2.3)$$

If  $VAD = 0$ , then the noise spectrum, mean, and standard deviation for frame are all updated. Frames 2 through 10 are assumed to be noise in order to get a good initial average of the stationary noise in the environment.

$$N(w) = \alpha N(w) + (1 - \alpha) X(w, k) \quad (2.4)$$

$$\mu_N(k) = \frac{1}{L} \sum_{w=0}^{L-1} N(w) \quad (2.5)$$

$$\mu_N = \beta \mu_N + (1 - \beta) \mu_{N(k)} \quad (2.6)$$

$$\sigma_N = (\beta \sigma_N^2 + (1 - \beta) \mu_{N(k)}^2)^{1/2} \quad (2.7)$$

The mean of the noise estimate is  $\mu_N$ , the standard deviation of the noise estimate is  $\sigma_N$ , and the noise estimate variance is represented by  $\sigma_N^2$ .

Thresholds are updated if a frame does not contain speech, using the mean and variance of the noise estimate, where threshold settings are adjusted using the multipliers  $\alpha_S$  and  $\alpha_N$ , which can be adapted and set experimentally. Optimally adapting these VAD thresholds has been the subject of recent research, but was not attempted in this paper because sensitivity to the thresholds was reduced by the iteration of the algorithm as mentioned in section 5.1.

$$Thresh_S = \mu_N + \alpha_S \sigma_N \quad (2.8)$$

$$Thresh_N = \mu_N + \alpha_N \sigma_N \quad (2.9)$$

VAD decisions can be made using a speech threshold determination, where if the signal energy exceeds twice the standard deviation above the mean of the noise, then the frame is classified as speech. If the signal energy falls within some fraction of the noise standard deviation, then it is classified as noise and modifies the reference accordingly. If neither speech

nor noise is chosen, then the last frame's decision is repeated for the current frame.

$$\begin{aligned} &\text{if}(\text{Energy}(k) > \text{Thresh}_S), \text{VAD}(k) = 1 \\ &\text{if}(\text{Energy}(k) < \text{Thresh}_N), \text{VAD}(k) = 0 \\ &\text{else } \text{VAD}(k) = \text{VAD}(k-1) \end{aligned}$$

### 3. Spectral Subtraction (SS)

This section describes the spectral subtraction algorithm used and the steps required to calculate the perceptual mask threshold.

#### 3.1 SS Algorithm

Spectral subtraction uses the short-term spectral magnitude of the noisy speech and an estimate or reference of the noise signal. [1] Most single channel spectral subtraction methods use a voice activity detector (VAD) to determine when there is silence in order to get an accurate noise estimate and the noise is assumed to be short-term stationary so that noise from silent frames can be used to remove noise from speech frames. In order to estimate the clean speech frame a phase estimate is also required, but Wang and Lim [9] have determined that it is sufficient to use the noisy phase spectrum as an estimate of the clean speech phase spectrum. Figure 3-1 shows the signal flow for spectral subtraction where  $m(k)$  is a frame of unprocessed noisy data,  $k$  is the frame index,  $\omega$  is the frequency index,  $M(\omega, k)$  is the spectrum of the frame,  $N(\omega, k)$  is the spectrum of the noise estimate,  $T(\omega, k)$  is the perceptual mask threshold,  $a()$  and  $b()$  are the weighting functions,  $\hat{S}(\omega, k)$  is the spectrum of the speech estimate, and  $\hat{s}(k)$  is the speech estimate frame in the time domain.

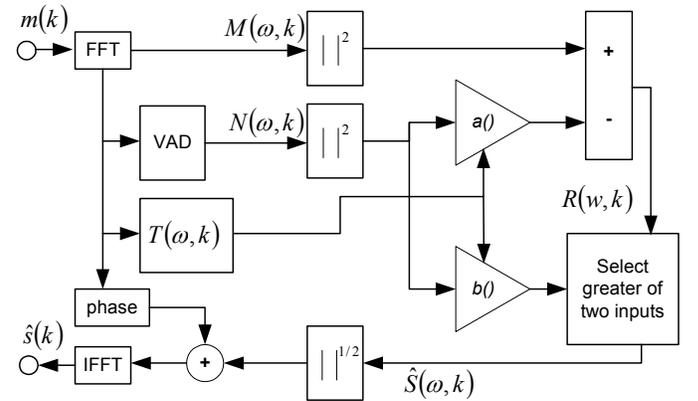


Figure 3-1: Spectral Subtraction with perceptual weighting

After subtraction, the spectral magnitude is not guaranteed to be positive and some possibilities to remove the negative components are by half-wave rectification (setting the negative portions to zero), full wave rectification (absolute value), or weighted difference coefficients. Half-wave rectification is commonly used but introduces “musical” tone artifacts in the

processed signal. Full wave rectification avoids the creation of musical tones, but is less effective at reducing noise. Much of the spectral subtraction research has focused on how to remove or reduce the creation of musical tones while maximizing the suppression of noise. [3] The SS algorithm prevents the negative spectral components from accruing by weighting the spectral gain function according to the masking threshold and a lower limit of zero.

### 3.2 Perceptual mask threshold

The algorithm in this paper uses perceptual nonlinear weighting of the gain function with spectral subtraction, which enables it to aggressively attenuate the noise while avoiding the introduction of annoying artifacts to the speech signal. SNR, signal to noise ratio, is the most broadly used criteria for reducing noise in a received speech signal and has been very successful, but it is limited because inaccuracy of the noise estimate can cause either excess residual noise or distortion of the signal. Taking advantage of the human auditory system's characteristics can help mitigate the effects of residual noise and render the speech to be more perceptually pleasing to the ear because the distortion of the signal is minimized by not processing noise that is effectively inaudible. The short-time spectral amplitude, STSA, enhancement methods can take advantage of how people perceive the frequencies instead of just working with SNR.

Perceptual speech enhancement techniques have the challenge that there is no clean speech reference or accurate spectral noise estimate in order to determine exact auditory masking thresholds. If the clean-speech masking threshold is too high then more noise will be left in the signal, but if the clean-speech masking threshold is calculated too low, then information about the desired signal will be lost. Spectral subtraction is used to obtain an estimate of the clean speech from which the masking thresholds are calculated. The steps required to calculate the masking threshold are taken directly from the paper by Johnson [6] and are listed below and shown in Figure 3-2.

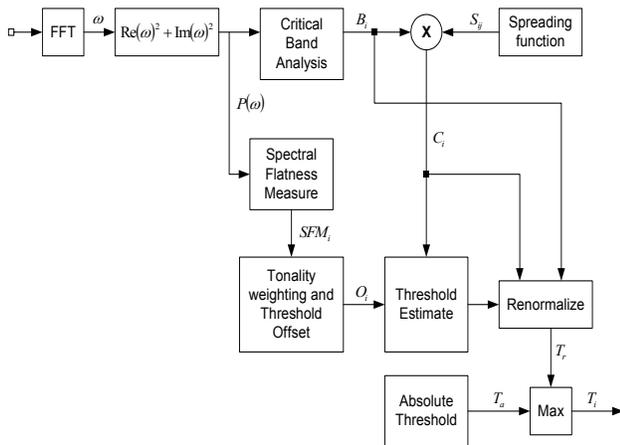


Figure 3-2: Masking Threshold Calculation

1. Obtain initial estimate of the speech
2. Apply the spreading function to the critical band spectrum

3. Calculate the spread spectrum masking threshold accounting for spectral flatness
4. Convert the spread spectrum back to the Bark domain via renormalization
5. Adjust for absolute thresholds
6. Relate the spread masking threshold to the critical band masking threshold

Critical Band Analysis partitions the power spectrum into critical bands according to the Bark scale as in Zwicker & Fastl [10]. The power spectrum is calculated from the frequency data as in equation 3.1.

$$P(\omega) = \text{Re}^2(\omega) + \text{Im}^2(\omega) \quad (3.1)$$

The energy in each critical band is summed in equation 3.2, where  $B_i$  is the energy for critical band  $i$ ,  $b_{li}$  is the lower frequency for the band, and  $b_{hi}$  is the upper frequency for the band.

$$B_i = \sum_{\omega=b_{li}}^{b_{hi}} P(\omega) \quad (3.2)$$

The number of critical bands used will depend on the bandwidth of the signal in question. Humans can only perceive frequencies between 20 Hz and 20kHz, so that places a bound on the range of frequencies to consider. There are 22 critical bands for an 8kHz signal that is sampled at the Nyquist rate of 16kHz.

The spreading function is used to estimate the effects of masking across critical bands. The spreading function is calculated as  $\lfloor (j-i) \rfloor \leq 25$ , where  $i$  is the Bark frequency of the masked signal and  $j$  is the Bark frequency of the masking signal. The term Bark is used to indicate the frequencies of one critical band. The spreading function is put into matrix form,  $S_{ij}$ , and convolved with critical band energies  $B_i$ . The spread critical band spectrum,  $C_i$ , is given in equation 3.3, where  $*$  is the convolution operator.

$$C_i = S_{ij} * B_i \quad (3.3)$$

There are different masking thresholds based on spectral flatness of the signals.

1. Tone masking noise is estimated as  $(14.5 + i)$  dB below  $C_i$ , where  $i$  is the Bark frequency.
2. Noise masking a tone is estimated as 5.5 dB below  $C_i$  uniformly across the critical band.

Spectral Flatness Measure, SFM, is defined in equation 3.6 as the ratio of the geometric mean,  $G_m$ , of the power spectrum to the arithmetic mean,  $A_m$ , of the power spectrum. Arithmetic mean is given in equation 3.4 and geometric mean is given in equation 3.5.

$$A_m = \frac{P(\omega_1) + P(\omega_2) + P(\omega_3) + \dots + P(\omega_n)}{n} \quad (3.4)$$

$$G_m = \sqrt[n]{P(\omega_1) \cdot P(\omega_2) \cdot P(\omega_3) \cdot \dots \cdot P(\omega_n)} \quad (3.5)$$

$$SFM_{dB} = 10 * \log_{10} \frac{G_m}{A_m} \quad (3.6)$$

The coefficient of tonality in equation 3.7,  $\alpha$ , is calculated where an SFM = SFM<sub>dbmax</sub> = -60 dB indicates the signal is very tone-like and an SFM = 0 indicates the signal is more noise-like. For example an SFM = -30dB would result in  $\alpha = 0.5$ .

$$\alpha = \min\left(\frac{SFM_{dB}}{SFM_{dB\max}}, 1\right) \quad (3.7)$$

The offset in equation 3.8,  $O_i$ , for the masking energy in each band, is determined by using the tonality to weight the masking thresholds for tones and noise.

$$O_i dB = \alpha * (14.5 + i) + (1 - \alpha) * 5.5 \quad (3.8)$$

The spread threshold estimate is calculated using equation 3.9.

$$T_i = 10^{\log_{10}(C_i) - \left(\frac{O_i}{10}\right)} \quad (3.9)$$

The spreading convolution must now be undone and the threshold converted back to the Bark domain. De-convolution is unstable due to the shape of the spreading function and would introduce undesired artifacts into the signal, so renormalization is used instead to remove the increased energy added to each band by the spreading function. Renormalization multiplies each  $T_i$  by the inverse of the energy gain, assuming a uniform energy of 1 in each band.

Critical band noise thresholds that are lower than the absolute threshold of hearing are changed to equal the mean of the absolute threshold of hearing for that band because it does not make sense to calculate a mask threshold for something that cannot be heard anyway. The absolute threshold of hearing has been measured with several experiments and is given as an estimated curve plotted versus frequency reported by Fletcher [5].

## 4. Experiments

Real data measurements and MATLAB<sup>®</sup> simulations were used to evaluate the algorithms, which are described below. Objective quality measures, used in the analysis, are described in section 4.3.

### 4.1 Measurements

Measurements were made in a 2001 Honda Odyssey minivan using a Larson-Davis BNK omni-directional microphone mounted between the visor and ceiling slightly above and in front of the driver, which was 38cm from the driver's mouth. The sampling rate was 16kHz with 16 bits of resolution. Adult male and female voices were recorded while the van was parked and off, with the windows up and down, in order get clean speech in a similar environment to where the noise would be measured. The clean speech also served as a reference for calculating speech quality metrics. The noise measurements in the van were made separately without the driver speaking and mixed with the speech later, so the SNR could be set to known values.

The frequencies below 1 kHz contain most the energy for the signals involved and most of the speech energy is in the lower frequencies with peaks around the pitch of the desired talker just below 200 Hz. About 90% of the road noise is contained below 120 Hz. The fan noise has significantly more energy around 200 Hz, which caused more problems than the road noise when mixed with the speech. The interfering talker noise has strong harmonics at 300 and 600 Hz, which corresponds well to the expected pitch of the children talking.

### 4.2 Simulations

The simulations were done in MATLAB<sup>®</sup> using the data recorded in the van down-sampled to 8kHz, where the various combinations of noise and speech were mixed at 0, 5, and 10 dB SNR. The voice activity detector (VAD) and frames-size were kept constant through all the simulation runs because they have such a large impact on the results. Using the same VAD enables fair comparison between regular spectral subtraction and the enhanced algorithm. MATLAB<sup>®</sup> M-files were also used to calculate the objective speech quality measures and create the plots to visualize the comparison of results.

### 4.3 Speech Quality Metrics

The person who listens to the speech is ultimately the one who decides its quality, thus subjective listening tests are the best way to judge the performance of an algorithm. Commonly used subjective tests are the Mean Opinion Score (MOS), Diagnostic Acceptability Measure (DAM), and Diagnostic Rhyme Test (DRT). The challenge with subjective measures is that a large number of people tested under consistent conditions are required to get valid results. Objective measures overcome this burden by allowing a computer to analyze the speech quality. The objective measures used in this paper were chosen for their good correlation to subjective tests and their use in related research.

**Table 4.1: Objective Speech Quality Measure Correlation to Subjective Tests**

Objective Speech Quality Measure	Correlation to Subjective Tests
Signal-to-Noise Ratio (SNR)	24%
Segmental SNR (SSNR)	77%
Articulation Index	67%
Itakura-Saito Distance	59%

The correlation measures in Table 4.1 were calculated against a database of subjective speech quality test data accumulated by Quackenbush [7], where the subjective quality test used was the Diagnostic Acceptability Measure (DAM). All the objective quality measures cited in Table 4.1 require the original speech for their calculations. The speech and the noise used in this paper were recorded separately in the same environment in order to have the required clean speech reference when computing objective quality measures.

## 5. Results

The results subjective listening tests agreed with the objective speech quality measures reported in section 5.1. Improved VAD accuracy is demonstrated and explained in section 5.2.

### 5.1 Iteration of Spectral Subtraction

Voice activity detection is generally more accurate for higher SNR, thus VAD accuracy should improve when using a noise reduced signal versus the original noisy speech. Experiments iterating the SS algorithm were motivated by the desire to calculate the VAD on a noised reduced signal. [4] Voice activity detection was improved by iterating the algorithm twice, as shown in Figure 5-1, and also made the VAD less sensitive to the fixed energy thresholds for detection of speech vs. noise.

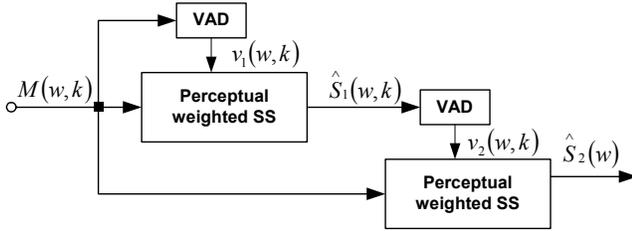


Figure 5-1: SS Iteration

Table 5.1 shows the speech quality results of iterating the algorithm and the corresponding percentage of voice frames detected. The example in Table 5.1 is for road noise at 5 dB SNR, but similar results were found for all noise types and SNR levels. It was seen that all the speech quality measures improve on the second iteration of the algorithm and the improvement is directly related to the performance of the VAD. The results also show that iterating more than two times starts to degrade the quality of the speech and provide less noise suppression.

Table 5.1: Speech Quality and SS Iteration

Signal	Road / Engin e	SS iter #1	SS iter #2	SS iter #3
%VAD	-	52	60	62
SNR $\eta$	5	15.5	16.8	15.4
SSNR $\eta$	1.23	11.17	12.30	11.74
AI $\eta$	0.49	0.54	0.62	0.61
IS $\iota$	0.51	0.58	0.39	0.42

Another interesting effect of iterating the algorithm is that the VAD is less sensitive to the fixed energy thresholds used to determine if speech or noise is present. The threshold for speech in the simulations was fixed to 0.7 of the variance of the noise estimate. This threshold could be moved up or down by as much as 0.3 with little change in % VAD reported in the second iterations. In contrast the first iteration would change the % VAD reported directly corresponding to any variation of the threshold.

The behavior of the VAD using the enhanced speech estimate is very logical when the algorithm for voice detection is examined. The speech threshold is a fixed constant,  $\alpha_S$ , multiplied by the standard deviation of the noise estimate,  $\sigma_N$ , and added to the mean,  $\mu_N$ , of the noise estimate as shown in equation 5.1.

$$Thresh_S = \mu_N + \alpha_S * \sigma_N \quad (5.1)$$

Noise variance is significantly smaller when the signal has passed through the first iteration of the perceptually weighted non-linear spectral subtraction. The smaller variance causes the overall value of the speech detection threshold,  $Thresh_S$ , to be lower, which naturally detects more speech frames. Lower noise variance also makes the VAD less sensitive to the choice of the fixed threshold constant because the value of the standard deviation,  $\sigma_N$ , that the fixed constant multiplies is less, so the impact of the constant,  $\alpha_S$ , on the VAD performance is also less.

The lower threshold is less likely to classify a speech frame as noise, thus avoiding attenuation of the speech. If this is taken too far by iterating the algorithm many times, then not enough frames contribute to the noise estimate and the noise is less effectively removed from the signal.

### 5.2 Improved VAD Accuracy

A closer analysis of the VAD is warranted because it plays such a critical role in the SS algorithm. Correct VAD decisions for each frame are determined by visually inspecting the clean speech signal and used as a reference,  $VAD_{ref}$ , for comparison to the VAD decisions calculated by the algorithm.

$$VAD\_Accuracy = \frac{correct\_frames}{total\_frames} \quad (5.2)$$

Parameters of the test signal are 8 kHz sampling rate, 128 samples per frame, 157 frames for signal length, and the length of the speech is 2.5 seconds or 20,000 samples

Visual inspection of the clean speech signal found 88 speech frames and 69 silent frames, which corresponds to 56% voice activity. The comparison to this reference for each noise type, SNR level, and SS iteration is reported in Table 5.2.

Table 5.2: % VAD accuracy

SNR (dB)	0	0	5	5	10	10
Iteration	1	2	1	2	1	2
Noise-free	95					
Road	91	91	95	94	96	95
Fan	81	89	90	94	94	96
Talker	73	75	76	75	74	74
AWGN	54	56	82	87	90	95

VAD accuracy, using the fan noise and AWGN, was improved by iterating the algorithm a second time. Road noise VAD accuracy

stayed about the same for both iterations, but the second iteration tended to classify more frames as speech. Interfering talkers produced consistently poor VAD accuracy across all SNR levels and iterations because the algorithm does not distinguish between desired and undesired talkers.

### 5.3 Conclusion

In summary, this paper has shown that using spectral subtraction prior to computing the VAD increased the VAD accuracy. The better performance of the VAD was used to reprocess the original signal using spectral subtraction, which achieved greater noise attenuation with less signal distortion. Perceptually weighted spectral subtraction was used to avoid the introduction of musical noise artifacts.

## 6. REFERENCES

- [1] Boll, S.F. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*- 27, no.2 (1979):113-120.
- [2] Deller, J.R. Jr., Hansen, J.H.L, and Proakis, J.G. *Discrete-Time Processing of Speech Signals*, (New York, NY: IEEE Press, 2000):266.
- [3] Ephraim, Y. and Malah, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, *IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-32*, no. 6, (1984):1109-1121.
- [4] Faneuff, J., *Spatial, Spectral, and Perceptual Nonlinear Noise Reduction for Hands-free Microphones in a Car*, Master's Thesis, Worcester Polytechnic Institute, July 2002. <http://www.wpi.edu/Pubs/ETD/Available/etd-0806102-214423/unrestricted/faneuff.pdf>
- [5] Fletcher, H. Auditory patterns, *Rev. Modern Phys.* 12, (1940):47-65.
- [6] Johnson, J.D. Transform Coding of Audio Signals Using Perceptual Noise Criteria. *IEEE Journal on Selected Areas in Communications* 6, no. 2, (February 1988):314-323.
- [7] Quackenbush, S.R, Barnwell, T.P., and Clements, M.A. *Objective Measures of Speech Quality*, Prentice Hall, Englewood Cliffs, NJ (1988):37-50.
- [8] Virag, N. Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System. *IEEE Transactions on Speech and Audio Processing* 7, no. 2., March 1999.
- [9] Wang, D.L. and Lim, J.S.. The unimportance of phase in speech enhancement. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 30, (August 1982):679-681.
- [10] Zwicker, E., Fastl, H., and Frater. *Psychoacoustics: Facts and Models*, 2<sup>nd</sup> Edition, Springer, (April 1999):149-173.