59/CW/IN/32

1999 NASA/ASEE SUMMER FACULTY FELLOWSHIP PROGRAM

JOHN F. KENNEDY SPACE CENTER
UNIVERSITY OF CENTRAL FLORIDA

NOISE REDUCTION USING FREQUENCY SUB-BAND BASED ADAPTIVE SPECTRAL SUBTRACTION

Dr. David Kozel
Associate Professor of Electrical Engineering
Purdue University Calumet
Department of Engineering
Hammond, IN

NASA/KSC
NASA Colleague: Frederick McKenzie
Communications/RF and Audio

July 31, 1999

Abstract

A frequency sub-band based adaptive spectral subtraction algorithm is developed to remove noise from noise-corrupted speech signals. A single microphone is used to obtain both the noise-corrupted speech and the estimate of the statistics of the noise. The statistics of the noise are estimated during time frames that do not contain speech. These statistics are used to determine if future time frames contain speech. During speech time frames, the algorithm determines which frequency sub-bands contain useful speech information and which frequency sub-bands contain only noise. The frequency sub-bands, which contain only noise, are subtracted off at a larger proportion so the noise does not compete with the speech information. Simulation results are presented.

1. Introduction

It is desired to incorporate adaptive noise suppression into the communications equipment on the Emergency Egress Vehicle and the Crawler-Transporter. In the case of the Emergency Egress Vehicle, the spectral content of the noise source changes as a function of the speed of the vehicle and its engine. In the case of the Crawler-Transporter, the noise a person hears will vary with his location relative to the Crawler-Transporter and if the hydraulic leveling device on the Crawler-Transporter is being used. Due to the varying nature of the noise, an adaptive algorithm is necessary for both applications. Furthermore, the noise frequencies produced by both applications are in the voice band range, so standard filtering techniques will not work. To remove noise from a noise-corrupted speech signal; a frequency sub-band based adaptive spectral subtraction algorithm is developed. In the following sections, a brief overview of spectral subtraction and its limitations is given, the frequency sub-band based adaptive spectral subtraction algorithm is described in detail along with the advantage to using frequency sub-bands, and simulation results are presented and discussed.

2. Spectral Subtraction

Spectral subtraction assumes that noise-corrupted speech is composed of speech plus additive noise.

$$x(t)=s(t) + n(t) \tag{1}$$

Where:

x(t) = noise-corrupted speech
s(t) = speech
n(t) = noise

Taking the Fourier Transform of equation (1),

$$|X(f)|e^{j\theta_X} = |S(f)|e^{j\theta_S} + |N(f)|e^{j\theta_N} \tag{2}$$

When no reference microphone is used, the magnitude and phase of the noise are unavailable when speech is present. The phase of the noise-corrupted speech is commonly used to approximate the phase of the speech. This is equivalent to assuming that the noise-corrupted speech and the noise are in phase. The average magnitude of the noise, $|\hat{N}(f)|$, is usually used to approximate the magnitude of the noise. Since the noise spectrum will in general have sharper peaks than the average noise spectrum, a multiple, μ, of the average noise spectrum is subtracted. This is done to reduce "musical-noise" which is caused from these random peaks. Solving for the estimated speech spectrum,

$$|\hat{S}(f)|e^{j\theta_X} = (|X(f)| - \mu|\hat{N}(f)|)e^{j\theta_X} \tag{3}$$

The inverse Fourier Transform yields the estimated speech:

$$\hat{s}(t) = \mathcal{F}^{-1}\{\hat{S}(f)\} \tag{4}$$

2.1 Limitations of Spectral Subtraction

When using any algorithm, it is important to understand its limitations and restrictions. Since the noise and speech have no physical dependence, the assumption that the noise and speech are in phase at any or all frequencies has no basis. Rather, they can be thought of as two independent random processes. The phase difference between them at any frequency has an equal probability of being any value between zero and 2π radians. Thus, the noise and speech vectors at one frequency may add with a phase shift while simultaneously at a different frequency may subtract with a different phase shift. Thus, subtracting an assumed in-phase noise signal from the noise-corrupted speech has the same probability of reducing the particular frequency component of the speech even further as it does of bringing it back to its proper level. Furthermore, it is almost certain to cause some distortion in the phase. The amount of error produced at each frequency depends upon the relative phase shift and the relative magnitudes of the speech and noise vectors. As noted in [1], for each spectral frequency that the magnitude of the speech is much larger than the corresponding magnitude of the noise, the error is negligible. For the consonant sounds of relatively low magnitude, the error will be much larger. This is true even if the magnitude of the noise at each frequency could be exactly determined during speech.

3. The Value of Sub-bands

For a given range of frequencies, say zero to six kilohertz, each speech sound is only composed of some of the frequencies. No sound is composed of all of the frequencies. If the spectrum is divided into frequency sub-bands, the frequency sub-bands containing just noise can be removed when speech is present. Furthermore, during speech the power level of the frequency sub-bands that contain speech will increase by a larger proportion than the power level of the entire spectrum. Thus, speech will be easier to detect by looking at the sub-band power change than by looking at the overall power change. This is especially true of the consonant sounds, which are of lower power, but are concentrated in one or two frequency sub-bands. By dividing the signal into frequency sub-bands, frequency bands that do not contain useful information can be removed so that the noise in those frequency sub-bands does not compete with the speech information in the useful sub-bands.

3.1 Adaptive Spectral Subtraction Algorithm

Details of the frequency sub-band based adaptive spectral subtraction algorithm are described in this section. The signal is sampled, windowed with a hamming window, and zero padded by the same procedure described in [1]. Each time frame of signal overlaps the previous time frame by 50 percent. An "m" point Fast Fourier Transform is taken, and the magnitude of the frequency response is separated from the phase angle. The magnitude response is partitioned into frequency sub-bands as shown in Table 1. The

range of frequencies in each sub-band is chosen in accordance with the bark scale [2] to account for the hearing characteristics of the human ear.

Sub-band	Start Bin	Stop Bin	Number of Bins	Beginning Frequency (Hz)		Ending Frequency (Hz)
1	1	16	16	0	-	388
2	17	21	5	388	-	505
3	22	26	5	505	-	622
4	27	32	6	622	-	763
5	33	38	6	763	-	904
6	39	45	7	904	-	1069
7	46	53	8	1069	-	1257
8	54	62	9	1257	-	1468
9	63	72	10	1468	-	1703
10	73	84	12	1703	-	1985
11	85	98	14	1985	-	2314
12	99	114	16	2314	-	2690
13	115	133	19	2690	-	3136
14	134	156	23	3136	-	3676
15	157	186	30	3676	-	4381
16	187	224	38	4381	-	5273
17	225	256	32	5273	-	6025

Table 1. Frequency Ranges of the Frequency Sub-bands

To key into the communication system, the user is required to press and hold a push-to-talk button while speaking into the microphone. Thus, it is assumed that speech is not present when the push-to-talk is not pressed. For each time frame, L, when the push to talk is not pressed, the signal is just noise.

$$|X_L(kf)| = |N_L(kf)| \quad \text{for frequency bins } k = 1, \ldots, m \tag{5}$$

While the push-to-talk is not pressed, the statistics of the noise are determined, and the algorithm is initialized. The statistics of the noise are updated every n_A time frames until a push-to-talk occurs. n_A is chosen large enough to provide reliable noise statistics and small enough to be updated before each push-to-talk. The average noise magnitude for each frequency bin is determined using the sample mean.

$$\overline{N}(kf) = \frac{1}{n_A}\sum_{L=1}^{n_A}|N_L(kf)| \qquad \text{for frequency bin } k = 1, \ldots, m \tag{6}$$

The power in frequency sub-band ν for time frame L is

$$P_{L\nu} = \sum_{k=\beta_\nu}^{\xi_\nu}|X_L(kf)|^2 \tag{7}$$

Where β_ν and ξ_ν are the beginning and ending frequency bins for sub-band ν. The average power in frequency sub-band ν over the n_A time frames is estimated using the sample mean.

$$P_{A\nu} = \frac{1}{n_A}\sum_{L=1}^{n_A}P_{L\nu} \qquad \text{for sub-band } \nu = 1, \ldots, \eta \tag{8}$$

The standard deviation of the power in frequency sub-band ν over the n_A time frames is estimated using the square root of the sample variance.

$$\sigma_\nu = \sqrt{\frac{1}{(n_A - 1)} \sum_{L=1}^{n_A} (P_{A\nu} - P_{L\nu})^2} \qquad \text{for sub-band } \nu = 1, \ldots, \eta \quad (9)$$

The threshold proportions for duration and burst speech in each frequency sub-band are dependent on the standard deviation of the power in that frequency sub-band and externally adjustable proportions, α_d and α_b.

$$\tau_{d\nu} = (1 + \alpha_d \sigma_\nu) \qquad \text{for sub-band } \nu = 1, \ldots, \eta \quad (10)$$

$$\tau_{b\nu} = (1 + \alpha_b \sigma_\nu) \qquad \text{for sub-band } \nu = 1, \ldots, \eta \quad (11)$$

Once an average value for the noise is determined, the maximum ratio of noise to average noise over the sub-band

$$MR_{L\nu} = \max_{\text{over } k=\xi_\nu, \ldots, \beta_\nu} \left(\frac{|N_L(kf)|}{|\overline{N}(kf)|} \right) \text{ for sub-bands } \nu = 1, \ldots, \eta \quad (12)$$

and the running average of $MR_{L\nu}$

$$AMR_\nu = (1 - \mu) AMR_\nu + \mu MR_{L\nu} \text{ for sub-bands } \nu = 1, \ldots, \eta \quad (13)$$

are determined.

When the push-to-talk is pressed, the algorithm must determine if speech is present during that particular time frame. For each time frame, L, the noise flags for the sub-bands, γ_ν, the noise flag counter, γ_C, and the noise flag record vector, γ_R, are initialized to the following values:

$$\gamma_\nu = 1 \qquad \text{for sub-band } \nu = 1, \ldots, \eta \quad (14)$$

$$\gamma_C = 0 \quad (15)$$

$$\gamma_R(1) = 0 \quad (16)$$

Then, for sub-band ν,

if{ [all $P_\nu(L, \ldots, L+\delta_d) > \tau_{d\nu} P_{A\nu}$] or [all $P_\nu(L-\delta_d, \ldots, L) > \tau_{d\nu} P_{A\nu}$]
or [all $P_\nu(L-\delta_c, \ldots, L+\delta_c) > \tau_{d\nu} P_{A\nu}$] or [$P_\nu(L) > \tau_{b\nu} P_{A\nu}$]} (17)

set

$$\gamma_\nu = 0 \quad (18)$$

$$\gamma_C = \gamma_C + 1 \quad (19)$$

$$\gamma_R(\gamma_C) = \nu \quad (20)$$

Equations (17) through (20) are repeated for sub-band $\nu = 1, \ldots, \eta$. In equation (17), the time frame shifts, δ_d and δ_c, required for duration speech are based upon the minimum time duration required for most speech sounds [3, p.62]. The time frame shift, δ_d, is used to detect the beginning and ending of speech sounds. The frame shift, δ_c, detects isolated speech sounds. The burst speech threshold proportion, $\tau_{b\nu}$, should be larger than the duration speech threshold proportion, $\tau_{d\nu}$, but the time required shorter since bursts generally have more energy but don't last as long. Equation (17) looks into the future (i.e., $P_\nu(L, \cdots, L+\delta_d)$) by processing frames of data but holding back decisions on them for δ_d time frames.

After using equation (17) to check all of the sub-bands, if [(γ_C>1) or (γ_R(1) > 14)], the frame is considered to be a speech frame. During speech frames, the ratio of the sum of noise-corrupted speech to sum of average noise

$$R_{L\nu} = \frac{\sum_{k=\beta_\nu}^{\xi_\nu} |X_L(kf)|}{\sum_{k=\beta_\nu}^{\xi_\nu} |\overline{N}_L(kf)|} \quad \text{for frequency sub-bands } \nu = 1, \ldots, \eta \tag{21}$$

is updated. Then, the speech estimate is determined using

$$|\hat{S}_L(kf)| = |X_L(kf)| - \min[R_{L\nu}, AMR_\nu](1 + \alpha_p \sigma_\nu)(1 + \alpha_f \gamma_\nu)\overline{N}(kf)$$

$$\text{for } \nu = 1, \ldots, \eta \text{ and } k = \xi_\nu, \ldots, \beta_\nu \tag{22}$$

If the magnitude of the estimated speech is less than zero for any frequency, it is set equal to zero. In equation (22), the proportion of the average noise subtracted is weighted by the minimum of $R_{L\nu}$ and AMR_ν. $R_{L\nu}$ is large during strong vowel sounds, but small during weaker consonant sounds. AMR_ν is the running average of the proportion needed to remove all of the noise. This proportion will remove too much speech information during weaker consonant sounds. The above weights are multiplied by σ_ν to account for the variation in the noise. The noise flag, γ_ν, increases the proportion subtracted when speech is not present in a frequency sub-band.

If the time frame is not a speech frame, it is a noise frame. During noise frames,

$$|N_L(kf)| = |X_L(kf)| \quad \text{for frequency bins } k = 1, \ldots, m, \tag{23}$$

and the following values are updated. The maximum ratio of noise to average noise over each frequency sub-band

$$MR_{L\nu} = \max_{\text{over } k=\xi_\nu, \ldots, \beta_\nu} \left(\frac{|N_L(kf)|}{|\overline{N}(kf)|} \right) \quad \text{for frequency sub-bands } \nu = 1, \ldots, \eta. \tag{24}$$

The running average of $MR_{L\nu}$

$$AMR_\nu = (1 - \mu)AMR_\nu + \mu MR_{L\nu} \quad \text{for } \nu = 1, \ldots, \eta \tag{25}$$

The running average of the power

$$P_{A\nu} = (1 - \mu)P_{A\nu} + \mu P_{L\nu} \quad \text{for frequency sub-bands } \nu = 1, \ldots, \eta, \tag{26}$$

and the running average of the noise at each frequency

$$\overline{N}(kf) = (1 - \mu)\overline{N}(kf) + \mu |N_L(kf)| \quad \text{for } k = 1, \ldots, m. \tag{27}$$

Also, the estimated speech signal is set to zero.

$$|\hat{S}_L(kf)| = 0 \quad \text{for } k = 1, \ldots, m \tag{28}$$

At this point the algorithm checks to see if the push-to-talk is still being pressed. If it is, the process is repeated starting at equation (14). If it is not, the algorithm goes back to the initialization stage, equation (5), to update the statistics of the noise and obtain new threshold proportions.

4. Results and Discussion

The algorithm developed in Section 3 was tested using noise-corrupted speech collected at 12.05 K Hz from the Emergency Egress Vehicle [4]. To generate each time frame, the data was windowed with a hamming window of length 256 points and zero padded to 512 points. Each frame of data overlapped the previous frame of data by 50 percent. The section of data contained the words, "pond", "key", "so", and "wren" chosen from the list given in the Diagnostic Rhyme Test (DRT) [5]. Spectrograms of the original signal containing the noise-corrupted speech and the signal after frequency sub-band based adaptive spectral subtraction are shown in Figure 1.

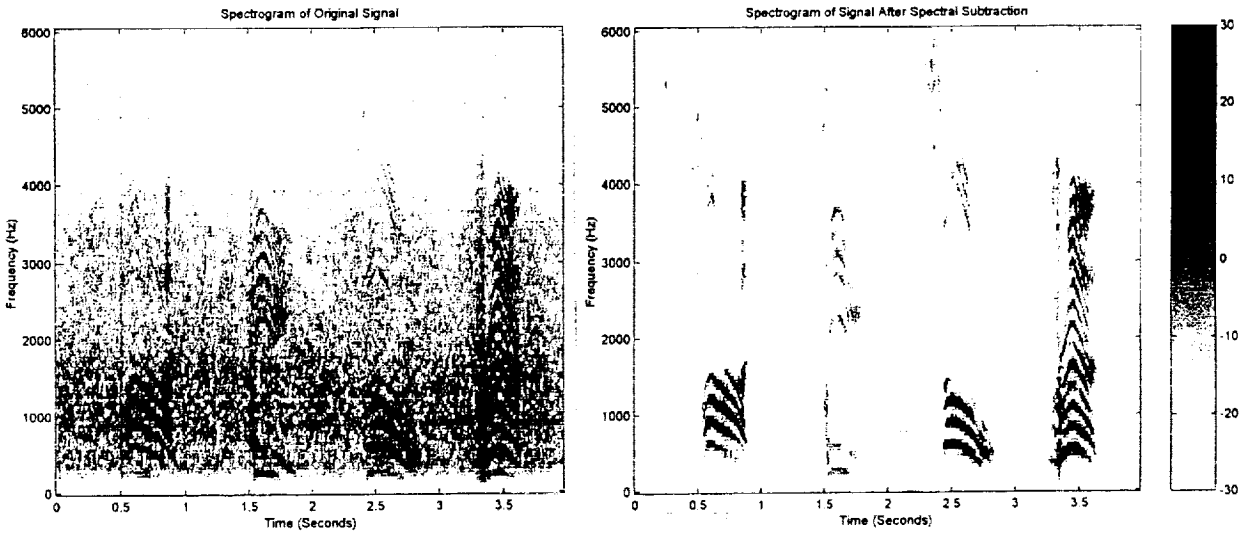

Figure 1. Spectrogram of Original Signal and Signal After Frequency Sub-band Based Spectral Subtraction

The original signal was pre-filtered [6] to compensate for the effects of the anti-aliasing filter, which was required for the A/D converter and the power reduction in speech at higher frequencies [7, p.238]. The ratio of power to average power in the noise-corrupted speech signal for frequency sub-bands 7, 13, and 17 is shown in Figure 2 along with the corresponding long and short term speech power thresholds, the sub-band noise flag, and overall noise flag for each time frame of the data sequence. As can be seen by the power of the signal relative to the power thresholds in each frequency sub-band, one frequency sub-band may contain speech information during a given time frame, while another does not. Frequency sub-band 7 contains the "on" sound of the word "pond", the "k" sound of the word "key", and the "o" sound of the word "so" in time frames approximately 50 - 75, 138 - 145, and 230 – 250, respectively. The noise flag for frequency sub-bands 13 and 17 for the same time frames indicate that these frequency sub-bands do not contain speech information during these time frames. Frequency sub-band 13 contains the "d" sound of the word "pond" and the "e" sound of the word "key" in time frames approximately 75 - 85 and 150 – 165, respectively. The noise flag for frequency

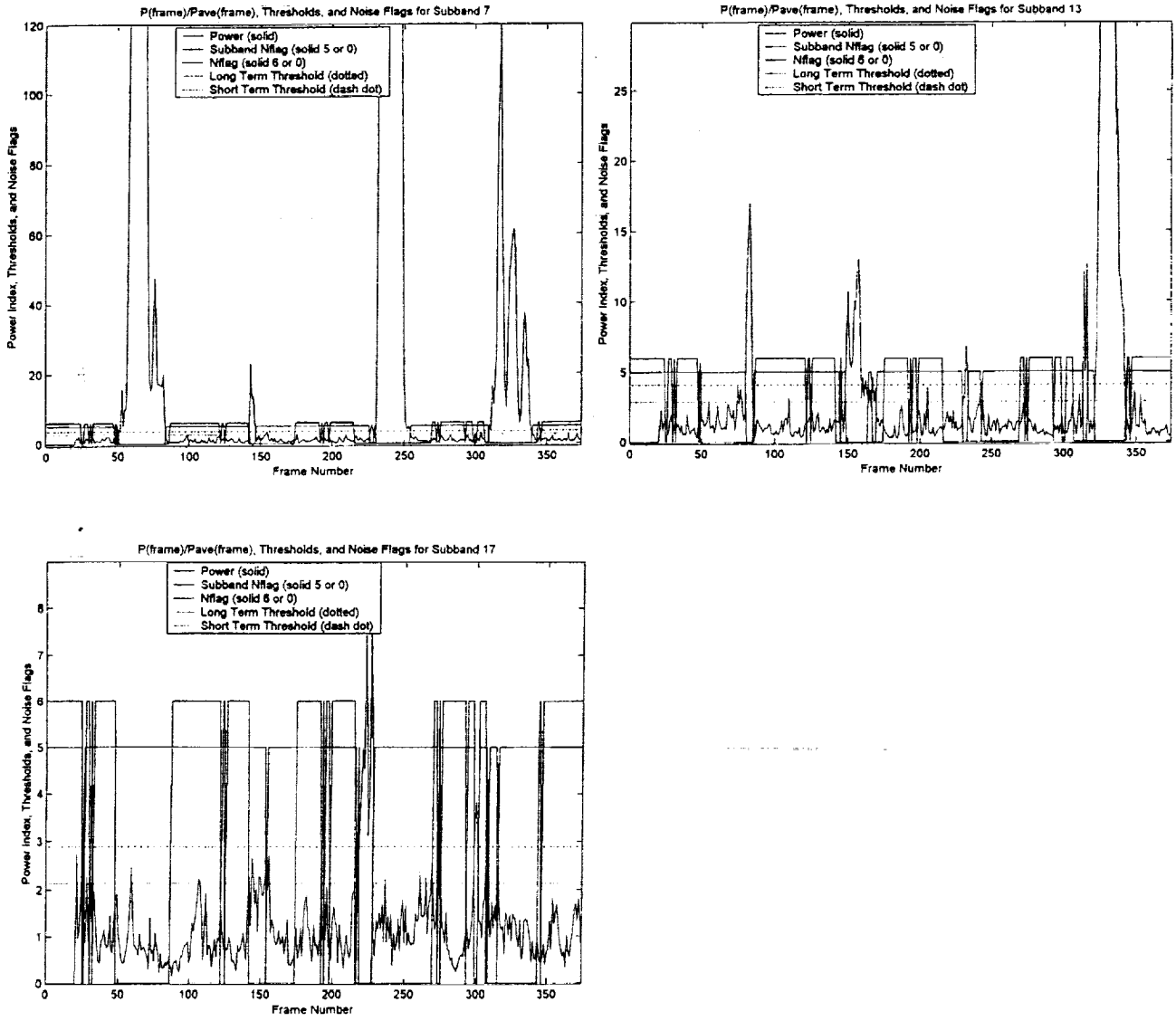

Figure 2. Power/(Average Power), Speech Power Thresholds, and Noise Flags for Frequency Sub-bands 7, 13, and 17, Respectively.

sub-bands 7 and 17 for the same time frames indicate that they do not contain speech information during these time frames. Finally, frequency sub-band 17 contains the "s" sound of the word "so" in time frames approximately 220 - 230. The noise flag for frequency sub-bands 7 and 13 for the same time frames indicate that they do not contain speech information during these time frames. According to equation (17), the noise in the frequency sub-bands that do not contain speech information will be subtracted off at a much greater proportion than the noise in the frequency sub-bands that contain speech during that particular time frame. This is done to essentially remove all noise in frequency sub-bands that do not contain speech information while preserving as much speech information as possible when removing noise from frequency sub-bands that

contain speech information. Comparing the magnitude scales for the different sub-bands in Figure 2, it is apparent that a very small overall relative power increase occurs for some of the consonant sounds such as the "s" in the word so. These power increases would be difficult to detect if sub-bands were not used.

A plot of the noise and average noise as a function of frequency for the final time frame is displayed in Figure 3. It is apparent that a multiple of the average noise must be subtracted from the noise in order to remove the spectral noise peak values. Due to the nature of the noise being considered, these spectral peaks vary in frequency and magnitude from time frame to time frame. When speech is present, the amount of over subtraction for frequency sub-bands containing speech information must be limited or too much of the speech information will be removed with the noise. Figure 4 displays R_{Lv}, MR_v, and AMR_v as a function of time frame for frequency sub-band 13.

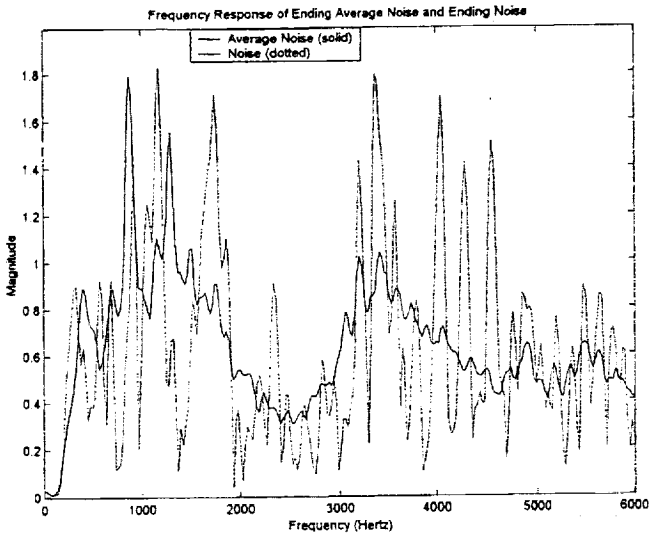

Figure 3. Frequency Response of |Noise| and Average |Noise| for Final Time Frame

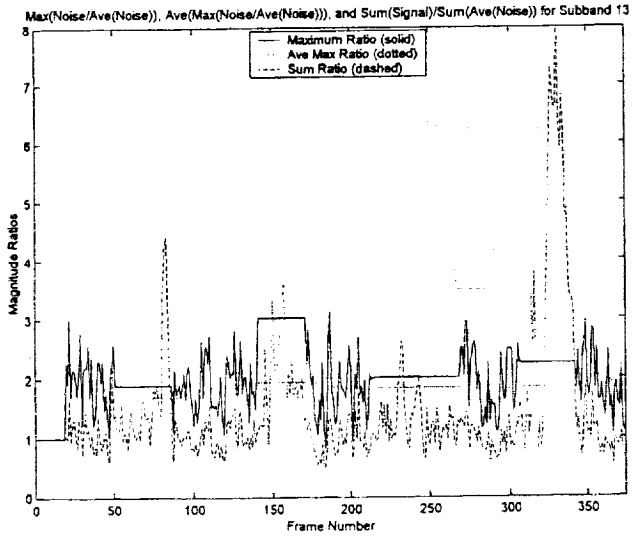

Figure 4. Maximum(|Noise|/(Average |Noise|), Average Maximum(|Noise|/(Average |Noise|), Sum(|Signal|)/Sum(Average |Noise|) for Frequency Sub-band 13

The minimum of R_{Lv} and AMR_v is used to limit the amount of noise removed in a frequency sub-band when speech is present.

5. Conclusion

Figure 1 demonstrates that the algorithm removes noise from the frequency sub-bands that do not contain speech information, while preserving the speech information in the frequency sub-bands that contain speech. Places for improvement in the algorithm include an estimate of the ratio of noise power to speech power so that the user would not have to set the parameter, α; the use of feedback to estimate MR_v so that it does not need to be calculated; and a better estimate of the instantaneous noise when speech is present. All of these goals can be achieved by using multiple microphones.

References

[1] Kozel, David, NASA/ASEE Summer Faculty Fellowship Program Research Reports: NASA CR-202756; 1996, p143-157.

[2] E. Zwicker and H. Fastl, Psychoacoustics Facts and Models, Springer-Verlag, 1990.

[3] Digital Signal Processing Applications with the TMS320C30 Evaluation Module: Selected Application Notes, literature number SPRA021, 1991.

[4] Kozel, David, NASA/ASEE Summer Faculty Fellowship Program Research Reports: NASA CR-1999-208546; 1998, p103-112.

[5] Voiers, William, D., Evaluating Processed Speech using the Diagnostic Rhyme Test, Speech Technology, January/February, 1983, p30-39.

[6] Kozel, David, NASA/ASEE Summer Faculty Fellowship Program Research Reports: NASA CR-207197; 1997, p113-122.

[7] Davis, Don, and Davis, Carolyn, Sound System Engineering Second Edition, Macmillan Publishing Co., New York, NY, 1987.