

NEW METHODS FOR ADAPTIVE NOISE SUPPRESSION

Levent Arslan¹, Alan McCree, and Vishu Viswanathan

Systems and Information Science Laboratory
P. O. Box 655474, Mail Stop 238
Texas Instruments
Dallas, TX 75265, U.S.A.

ABSTRACT

We propose three new adaptive noise suppression algorithms for enhancing noise-corrupted speech: smoothed spectral subtraction (SSS), vector quantization of line spectral frequencies (VQ-LSF), and modified Wiener filtering (MWF). SSS is an improved version of the well-known spectral subtraction algorithm, while the other two methods are based on generalized Wiener filtering. We have compared these three algorithms with each other and with spectral subtraction on both simulated noise and actual car noise. All three proposed methods perform substantially better than spectral subtraction, primarily because of the absence of any musical noise artifacts in the processed speech. Listening tests showed preference for MWF and SSS over VQ-LSF. Also, MWF provides a much higher mean opinion score (MOS) than does spectral subtraction. Finally, VQ-LSF provides a relatively good spectral match to the clean speech, and may, therefore, be better suited for speech recognition.

1. INTRODUCTION AND MOTIVATION

With the advent of digital cellular telephones, the role of noise suppression in speech processing problems such as speech coding and speech recognition has taken on an increased importance. This increased importance is due not only to customer expectation of high performance even in high car noise situations but also to the need to move progressively to lower data rate speech coding algorithms to accommodate the ever-increasing number of cellular telephone customers. While the higher data rate speech coding algorithms tend to maintain robust performance even in high noise conditions, that is usually not the case with lower data rate speech coding algorithms; the speech quality from the latter tends to degrade drastically in high noise. While noise suppression to prevent such speech quality losses is important, it must be achieved without introducing any undesirable artifacts or speech distortions or any significant loss of speech intelligibility. These performance goals for noise suppression have been around for many years, but they have now come to the forefront in the digital cellular

¹With Department of Electrical Engineering, Duke University, Durham, NC 27708-0291, U.S.A.; work was performed as part of a TI Summer Intern project.

telephone application, which is the context in which this research is being undertaken.

With these performance goals in mind, we have developed, implemented, and tested and compared against the well-known spectral subtraction method in simulated and actual car noise conditions three new adaptive noise suppression algorithms. Tests conducted included objective methods using the signal-to-noise ratio (SNR), the Itakura-Saito distance measure, and speech recognition performance improvement, and subjective speech quality evaluations using informal listening tests, mean opinion score (MOS) tests, and A-B comparison listening tests. Because of our emphasis on the digital cellular telephone application, we included in our tests speech output from the 8 kbits/s VSELP coder, which is the North American Digital Cellular Full-rate Standard (IS-54). All three proposed algorithms, which are described below along with test results, use the same approach for adaptive estimation of the noise power spectrum; this estimation approach is also described below.

2. SMOOTHED SPECTRAL SUBTRACTION

If the additive noise $n(t)$ is stationary and uncorrelated with the clean speech signal $s(t)$, then the power spectrum of the noisy speech $y(t)$ is the sum of the power spectra of $s(t)$ and $n(t)$:

$$\begin{aligned}y(t) &= s(t) + n(t) \\ P_y(w) &= P_s(w) + P_n(w)\end{aligned}$$

Therefore, the clean speech spectrum can be estimated by simply subtracting the noise spectrum from the noisy speech spectrum, which is the basis of the spectral subtraction technique [1]:

$$\hat{P}_s(w) = P_y(w) - P_n(w)$$

In practice, this technique can be applied frame by frame to the input signal using a Fast Fourier Transform (FFT) algorithm to estimate the power spectrum. After the clean speech spectrum is estimated by spectral subtraction, the clean speech time signal is generated via inverse FFT from this magnitude spectrum and the phase of the original signal.

The spectral subtraction method can substantially reduce the noise level of the noisy input speech, but it introduces an annoying distortion of its own. This distortion is due to fluctuating tonal noises in the output signal, a phenomenon

commonly called musical noise. As a result, the processed speech may sound worse than the original noisy speech and is unacceptable to many listeners.

The musical noise problem is best understood by interpreting spectral subtraction as a time-varying linear filter [2]. First, we rewrite the spectral subtraction equation as:

$$\begin{aligned}\hat{S}(w) &= H(w)Y(w) \\ H(w) &= \sqrt{\frac{P_y(w) - P_n(w)}{P_y(w)}} \\ \hat{s}(t) &= F^{-1}\{\hat{S}(w)\}\end{aligned}$$

where $Y(w)$ is the Fourier transform of the noisy speech, $H(w)$ is the time-varying linear filter, and $\hat{S}(w)$ is the estimate of the Fourier transform of the clean speech. Therefore, spectral subtraction consists of applying a frequency-dependent attenuation to each frequency in the noisy speech power spectrum, where the attenuation varies with the noisy signal to noise ratio (NSNR) at each frequency; NSNR = $\frac{P_y(w)}{P_n(w)}$. Since the frequency response of the filter $H(w)$ varies with each frame of the noisy speech signal, it is simply a time-varying linear filter. The left hand curve in Figure 1 shows the attenuation vs. NSNR for the spectral subtraction method. This illustrates that the amount of suppression varies rapidly with the NSNR at a given frequency, especially when the signal and noise are nearly equal in power. When the input signal contains only noise, musical noise is generated because the estimated NSNR at each frequency fluctuates due to measurement error, producing attenuation filters with random variation across frequencies and over time.

Our proposed Smoothed Spectral Subtraction (SSS) method involves three separate improvements over spectral subtraction. First, a clamp is applied to the filter $H(w)$ so that it cannot go below a minimum value, say, -10 dB. This prevents the noise suppression filter from fluctuating around very small gain values, and also reduces potential speech signal distortion. Second, the noise power spectrum estimate is artificially increased by a small margin, say, 5 dB, so that small errors in noisy signal spectral estimates do not lead to fluctuating attenuations [3]. These two modifications to the attenuation rule result in the curve plotted on the right in Figure 1, which has the same amount of attenuation for any small value of NSNR. Third, instead of using the FFT-derived estimates of the noisy speech and noise spectra directly in the attenuation rule, we use smoothed versions of the power spectra. We use a moving average smoothing in frequency; a smoothing window size of 32 (for an FFT size of 256) was found to work well. This smoothing reduces the variance of the spectral estimates, which prevents musical noises from occurring. As a combined result of these three improvements, the SSS algorithm is able to attenuate the acoustic background noise by 10 dB without introducing any musical noise artifacts.

3. LSF VECTOR QUANTIZATION BASED NOISE SUPPRESSION

We have also developed a vector quantization (VQ) based iterative Wiener filtering technique. This method, which

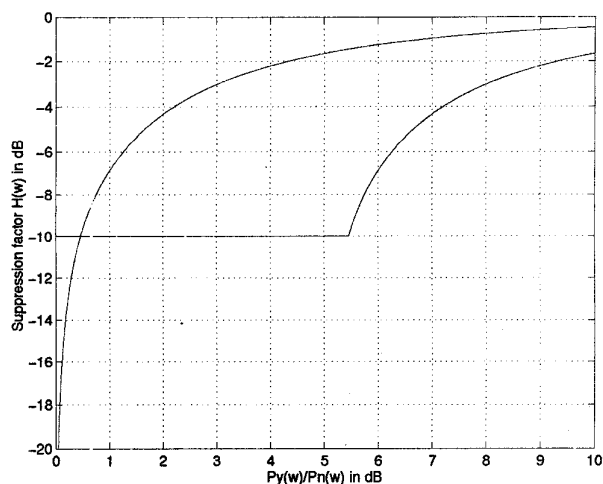


Figure 1: Attenuation curves for spectral subtraction and smoothed spectral subtraction.

we refer to as VQ-LSF, uses a generalized Wiener filter $H(w)$ [4]:

$$H(w) = \left(\frac{\hat{P}_s(w)}{\hat{P}_s(w) + \alpha P_n(w)} \right)^\beta$$

where $\hat{P}_s(w)$ is the clean speech power-spectrum estimate, α is the noise suppression factor (we use $\alpha = 10$), and β is the power of the filter (we use $\beta = 0.5$). The clean speech power spectrum is estimated as an LPC model spectrum as follows. We estimate the line spectral frequencies (LSF's) [5] of the clean speech based on the LSF's of the noisy speech data, by calculating a weighted average of codebook LSF's according to their perceptually weighted distances from the noisy speech LSF's. The VQ-LSF method iteratively improves the clean speech power spectral estimate by repeating the above estimation method on Wiener-filtered noisy speech. Typically, 5 or 6 iterations may be required. A similar technique has previously been proposed in [7], which uses a weighted sum of the LPC spectra based on their forward probabilities (computed using the hidden Markov modeling approach) for each mixture LPC spectrum. VQ-LSF is more efficient than the approach given in [7] both in the amount of computation and in the size of memory required.

We use a VQ codebook of LSF's of size 256. We calculate the distance of the noisy frame LSF's from each of the codebook entries. The calculation of the distance is based on a perceptual weighting called the inverse harmonic mean [6], and is given by:

$$d_k = \sum_{i=1}^P w_i (LSF_{ni} - LSF_{ki}) \quad k = 1, \dots, 256$$

$$w_i = \frac{1}{LSF_{ni} - LSF_{nc}}$$

where d_k is the distance corresponding to the k^{th} codeword, P is the number of LSF's, LSF_n refers to the noisy frame

LSF's, LSF_k refers to LSF's from the k^{th} codeword, w_i is the weight of the i^{th} LSF, and LSF_{nc} is the closest neighbor of LSF_{ni} . With this perceptually weighted distance, LSF's which are close to each other, and therefore have higher chance of being formants, are allowed to dominate the distance measure.

Based on this distance measure, we formulate a probability function which indicates how likely each codeword entry is to be the LSF's of the undegraded speech for that frame:

$$p_k = \frac{e^{-\gamma d_k}}{\sum_{i=1}^{256} e^{-\gamma d_i}} \quad k = 1, \dots, 256$$

where γ is a constant that controls the dynamic range for the probabilities. Large values of γ mean more emphasis on the weights of the higher probability codewords ($\gamma = 0.002$ was used in our experiments). Then the clean speech LSF's are estimated by:

$$LSF_{ci} = \sum_{k=1}^{256} p_k LSF_{ki}, \quad \text{with } i = 1, 2, \dots, P$$

Next, we convert the estimated clean speech LSF's to LPC coefficients \hat{a}_k , and calculate the LPC spectrum from:

$$\hat{P}_s(w) = \frac{\hat{g}_s^2}{|1 - \sum_{k=1}^P \hat{a}_k e^{-j\omega k}|^2}$$

where the gain of the LPC spectrum, \hat{g}_s^2 , is calculated from the following expression:

$$\hat{g}_s^2 = R_0 - \sum_{i=1}^P a_i R_i$$

where R_i is i^{th} autocorrelation lag, and a_i is the i^{th} linear predictor coefficient of the noisy speech. We use this estimate in the Wiener filter, and filter the noisy speech frame at each iteration. As in smoothed spectral subtraction, we use a clamp on the filter $H(w)$ so it does not go below a preset minimum (-10 dB).

4. MODIFIED WIENER FILTERING

Our third method, which we call modified Wiener filtering (MWF), is also based on the generalized Wiener filter. However, it uses a noise suppression factor that is time-varying and is computed based on the frame-by-frame SNR. As before, the lowest Wiener filter gain is clamped to a preset minimum threshold. The clean speech power spectral estimate is calculated from the LPC model spectrum of the noisy speech $P_y(w)$ with only a gain modification:

$$\hat{P}_s(w) = \frac{E_y - E_n}{E_y} P_y(w)$$

where E_y and E_n are, respectively, energies of the noisy speech and noise. The Wiener filter expression then reduces to:

$$H(w) = \sqrt{\frac{P_y(w)}{P_y(w) + \frac{E_y}{E_y - E_n} \alpha P_n(w)}}$$

Next, we make the factor multiplying $P_n(w)$ in the above expression inversely dependent on SNR (i.e., $\frac{E_s}{E_n}$ or $\frac{E_y - E_n}{E_y}$ under the assumption that signal and noise are uncorrelated), and allow it to change from frame to frame. This will ensure stronger suppression for noise-only frames and weaker suppression during voiced-speech frames which are not corrupted as much to begin with. The desired SNR dependence is achieved simply by replacing α with $\frac{E_n}{E_y} \alpha$. Then the expression for $H(w)$ becomes:

$$H(w) = \sqrt{\frac{P_y(w)}{P_y(w) + \frac{E_n}{E_y - E_n} \alpha P_n(w)}}$$

Unlike other Wiener filtering approaches, the MWF method is non-iterative and hence computationally attractive. The SNR-dependent noise suppression factor gives MWF the ability to suppress those parts of the degraded signal where speech is not likely to be present, and not to suppress and hence not to distort the voiced speech as much. As compared to spectral subtraction, the MWF method suppresses the noise to substantially lower levels without introducing noticeable artifacts and speech signal distortions.

5. ADAPTIVE NOISE SPECTRUM ESTIMATION

Most noise estimation methods update an average noise power spectrum during non-speech periods, but their performance depends upon accurate estimation of speech versus non-speech intervals, which is a difficult problem especially in high-noise conditions. We have developed a robust noise estimation method that does not require speech detection. At each frequency, the noise estimate is updated for each new input frame by moving towards the new power spectrum estimate. To keep the adaptive estimator from adjusting too quickly to increasing levels, the new estimate is not allowed to exceed 1.006 times the previous estimate or to be smaller than 0.978 times the previous estimate. Thus, the noise estimate cannot increase faster than 3 dB per second or decrease faster than 12 dB per second. As a result, the noise estimates will only slightly increase during short speech segments, and will rapidly return to the correct value during pauses between words. This approach is simple to implement, and is robust in actual performance since it makes no assumptions about the characteristics of either the speech or the noise signals.

6. RESULTS FROM EXPERIMENTAL EVALUATIONS

We generated noisy speech data by adding white noise or actual car noise to clean speech files. In addition, we used noisy speech data collected directly in a moving car. We compared the three proposed noise suppression methods against each other and against traditional power spectral subtraction. Tests conducted included objective methods using the signal-to-noise ratio (SNR), the Itakura-Saito distance measure, and speech recognition performance improvement, and subjective speech quality evaluations using

informal listening tests, mean opinion score (MOS) tests, and A-B comparison listening tests. All three methods produced substantial noise attenuation (10 dB or more) without producing musical noises or speech distortions. Each method was found to be significantly better than spectral subtraction in informal listening tests.

The VQ-LSF method does not suppress the noise in very low frequencies (e.g., 0-300 Hz), with the result that the residual noise in the enhanced speech sounds as though there is low-frequency emphasis. The low-frequency emphasis may not be desirable for listening purposes; however as shown in Table 1, of all methods tested, VQ-LSF provides the lowest Itakura-Saito distance computed between the clean speech and the enhanced speech and hence provides the best overall spectral match to the clean speech spectrum. Since we expect this to result in improved speech recognition performance, we performed a simple feasibility demonstration. We used a hidden Markov model-based isolated word recognition system on a 20-word vocabulary. The system was trained with data from 4 speakers, and tested with 5 different speakers. The recognition rate was 97.0% in a noise-free environment. When the original speech was degraded with white Gaussian noise at 10 dB SNR, the recognition rate dropped down to 71.6%. With the VQ-LSF algorithm as a front-end to recognition, the recognition rate improved to 83.6%.

As both SSS and MWF performed well in all tested conditions, we conducted A-B preference (forced choice) comparisons between the two on 12 sentences (6 male and 6 female speakers, one sentence each) collected in a car driven along a highway. A panel of 15 listeners preferred MWF over SSS 73% of the time before VSELP processing and 58% of the time after VSELP processing. The performance differences between the two methods are, however, not that large.

Based on these comparisons, we performed further testing to compare the MWF method with standard spectral subtraction using the same 12 highway noise sentences. We calculated the SNR values to be 14 dB for the noisy speech, 18 dB for the speech enhanced using spectral subtraction, and 23 dB for the MWF-enhanced speech. We conducted MOS tests using 15 listeners, also using the same 12 sentences. The MOS scores are shown in Table 2. The average MOS score is 2.7 for noisy speech, 1.9 for traditional spectral subtraction, and 3.4 for MWF.

REFERENCES

- [1] S. F. Boll, "Suppression of Acoustic Noise in Speech using Spectral Subtraction", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-27, No 2, pp.113-120, April 1979.
- [2] R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter" Subtraction", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No 2, pp. 74-82, April 1980.
- [3] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 208-211, April 1979.

- [4] J. Lim and A. V. Oppenheim, "All-pole Modeling of Degraded Speech", in *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-26, No 3, pp. 197-210, June 1978.
- [5] F. K. Soong and B. H. Juang, "Line Spectrum Pair (LSP) and Speech Data Compression", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1.10.1-1.10.4, 1984.
- [6] R. Laroia, N. Phamdo, and N. Farvardin, "Robust and Efficient Quantization of Speech LSP Parameters Using Structured Vector Quantizers", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 641-644, 1991.
- [7] Y. Ephraim, "A Minimum Mean Square Error Approach for Speech Enhancement", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 829-832, 1990.

Processing Type	Distance
Noisy Speech	0.355
Standard Spectral Subtraction	0.257
Smoothed Spectral Subtraction	0.273
Modified Wiener Filter	0.267
VQ-LSF	0.201

Table 1: Itakura-Saito distance measure values between the original clean speech and the enhanced speech files for white Gaussian noise with 10 dB SNR.

Listener	Degraded	Spec Sub	MWF
subject 1	2.333	1.833	3.667
subject 2	2.333	2.917	3.667
subject 3	1.917	2.750	2.833
subject 4	2.667	2.000	3.083
subject 5	3.083	2.000	2.583
subject 6	3.417	2.167	4.083
subject 7	3.000	1.750	3.333
subject 8	3.250	2.500	3.667
subject 9	3.083	1.917	3.917
subject 10	3.000	1.083	3.583
subject 11	2.500	1.333	3.417
subject 12	2.400	1.214	3.364
subject 13	2.667	1.000	3.750
subject 14	1.917	1.000	2.417
subject 15	3.000	2.333	4.083
Mean	2.704	1.853	3.430
St. Dev	0.463	0.625	0.508

Table 2: Mean Opinion Scores on a scale of 1 to 5 over 12 sentences recorded in a car on a highway (6 male, 6 female) for 3 different conditions: i) Original noisy speech, ii) enhanced speech using standard spectral subtraction, iii) enhanced speech using MWF method.