Comparison of Spectral Subtraction Methods used in Noise Suppression Algorithms

Mehdi Yektaeian
Department of Electrical and Computer Engineering
Isfahan University of Technology
ISFAHAN, IRAN
yektaeeian@yahoo.com

Rassul Amirfattahi
Department of Electrical and Computer Engineering
Isfahan University of Technology
ISFAHAN, IRAN
Fattahi@cc.iut.ac.ir

***Abstract*—The spectral subtraction method is a well-known noise reduction technique. Most implementations and variations of the basic technique advocate subtraction of the noise spectrum estimate over the entire speech spectrum. However, real world noise is mostly colored and does not affect the speech signal uniformly over the entire spectrum. In this paper, The artifacts introduced by STSA methods will describe, and how the spectral subtraction method is modified to counter these artifacts. The EVRC noise suppression algorithm will discuss in some detail. Finally, some of the methods will compare based on the attenuation given an estimate of the SNR and the results will show with diagrams.**

***Keywords*—spectral subtraction, envelop estimator, noise suppressor, noise reduction.**

I. INTRODUCTION

Spectral subtraction is a method to enhance the perceived quality of single channel speech signals in the presence of additive noise. It is assumed that the noise component is relatively stationary. specially, the spectrum of the noise component is estimated from the pauses that occur in normal human speech. The rest detailed treatment of spectral subtraction was performed by Boll [7,8]. Later papers expanded and generalized Boll's method to power subtraction, Wiener filtering and maximum likelihood envelope estimation.

II. BASIC SPECTRAL SUBTRACTION

Speech which is "contaminated" by noise can be expressed as

$$X(n) = S(n) + V(n) \tag{1}$$

where x(n) is the speech with noise, s(n) is the "clean" speech signal and v(n) is the noise process, all in the discrete time domain. What spectral subtraction attempts to do is to estimate s(n) from x(n). Since v(n) is a random process, certain approximations and assumptions must be made. One approximation is that the noise is (within the time duration of speech segments) a short-time stationary process. specially, it is assumed that the power spectrum of the noise remains cons-tant within the time duration of several speech segments (typically words or sentence fragments). Also, noise is assumed to be uncorrelated to the speech signal. This is an important assumption since ,the noise is estimated from pauses in the speech signal. finally, it is assumed that the human ear is fairly insensitive to phase, such that the effect of noise on the phase of s + v can be ignored. If the noise process is represented by its power spectrum estimate $|\hat{W}(f)|^2$, the power spectrum of the speech estimate $|\hat{S}(f)|^2$ can be written as:

$$|\hat{S}(f)|^2 = |X(f)|^2 - |\hat{W}(f)|^2 \tag{2}$$

since the power spectrum of two uncorrelated signals is additive. By generalizing the exponent from 2 to a, (2) becomes:

$$|\hat{S}(f)|^a = |X(f)|^a - |\hat{W}(f)|^a \tag{3}$$

This generalization is useful for writing the filter equation below [1,2]. The speech phase $\varphi_{\hat{S}}(f)$ is estimated directly from the noisy signal phase $\varphi_X(f)$:

$$\varphi_{\hat{S}}(f) = \varphi_X(f) \tag{4}$$

Thus a general form of the estimated speech in frequency domain can be written as:

$$\hat{S}(f) = (\max(|X(f)|^a - k|\hat{W}(f)|^a, 0))^{1/a} . e^{j\varphi x(f)} \tag{5}$$

where k > 1 is used to overestimate the noise to account for the variance in the noise estimate, as explained below. The inner term $|X(f)|^a - K|\hat{W}(f)|^a$ is limited to positive values, since it is possible for the overestimated noise to be greater than the current signal.

III. RELATED AND DERIVED METHODS

Since the development of the Spectral Subtraction method

by Boll [9], the basic problem has been attacked by changing the basic assumptions, in particular about the spectral magnitude of the noisy signal. Changing the basic assumption of (2) results in a different gain rule. For reference, some methods are presented here.

A. *The Wiener Filter*

Derived in a similar manner as the power spectral subtraction method, the Wiener Filter attempts to minimize the mean-squared error in frequency domain [5]. Writing $\Re(m, p)$ for the signal-to-noise ratio (SNR) of the m^{th} frequency bin, the generally cited form of the Wiener filter is :

$$G_W(m) = \frac{\Re(m)}{\Re(m)+1} \tag{6}$$

$\Re(m)$ is given as:

$$\Re(m) = \begin{cases} \dfrac{|X(m)|^2 - |\hat{W}(m)|^2}{|\hat{W}(m)|^2} & |X(m)|^2 > |\hat{W}(m)|^2 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and substituting in (6), we get

$$G_W(m) = \begin{cases} 1 - \dfrac{|\hat{W}(m)|^2}{|X(m)|^2} & |X(m)|^2 > |W(m)|^2 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

In fact, $|G_W(m) = \sqrt{G(m)}|$ with k = 1, a = 2,and $\alpha = 0$.

B. *Maximum Likelihood Envelope Estimator*

The Maximum Likelihood Envelope Estimator (MLEE) is based on the assumption that the speech signal is characterized by a deterministic waveform of unknown amplitude and phase [1]. The MLEE is characterized by its gain function:

$$G_{MLEE} = [\frac{1}{2} + \frac{1}{2}\sqrt{1 - \frac{|\hat{W}(m)|^2}{|X(m)|^2}}] \tag{9}$$

It should be notes that (9) was derived by estimating the a priori SNR. This leads directly to the Ephraim and Malah Noise Suppressor below.

III. The Ephraim and Malah Noise Suppressor

In [3], Ephraim and Malah presented a modification to the MLEE Filter by adding an estimator for the a priori SNR ($\Re_{prio}$) which uses exponential smoothing within the time domain. An examination of the algorithm by Cappe [3] concluded that this smoothing avoids the appearance of musical noise and signal distortion. However, removal of noise is not complete, and due to the smoothing, the signal component is incorrectly attenuated following signal transients. Cappe summarized the Ephraim and Malah Suppression Rule (EMSR) by:

$$G_{EMSR} = \frac{\sqrt{\pi}}{2}\sqrt{(\frac{1}{1+\Re_{post}})(\frac{\Re_{prio}}{1+\Re_{prio}})} \times M[(1+\Re_{post})(\frac{\Re_{prio}}{1+\Re_{prio}})] \tag{10}$$

where M stands for the function:

$$M(\theta) = \exp(\frac{-\theta}{2})[(1+\theta)I_0(\frac{\theta}{2}) + \theta I_1(\frac{\theta}{2})] \tag{11}$$

In the above equation, I0 and I1 represent the modified Bessel functions of zero and first order. Time and frequency indices have been omitted for clarity. The a priori SNR is calculated by:

$$\Re_{prio}(p) = (1-\alpha)R_{post}(p) + \alpha\frac{|G(p-1)X(p-1)|^2}{|\hat{W}|^2} \tag{12}$$

While the a posteriori SNR is the same as $\Re(m, p)$ given by (7). The value of α determines the time smoothing of the a priori SNR estimate, which on the basis of simulations was set to about 0.98. The a priori SNR is the dominant parameter, while the a posteriori SNR acts as a correction parameter when the a priori SNR is low. In [7], Scalart and Filho examined the use of a priori SNR estimation with standard (Boll, Wiener and MLEE) methods and also reported reduction in the amount of musical noise. This suggests that the smoothing operation plays a more significant role in the reduction of musical noise than the gain rule.

D. *Signal Subspace methods*

A new approach to noise reduction has been discussed by Ephraim and Van Trees [7], whereby the noisy signal is decomposed into a signal-plus-noise subspace and a noise subspace. The noise subspace is removed and the signal is generated from the remaining subspace by means of a linear estimation. Ephraim and Van Trees suggested the Discrete Cosine Transform (DCT) and Discrete Wavelet Transforms as approximations to the optimal, but computationally intensive Karhunen-Loeve Transform (KLT). Subjective tests showed that some distortion was introduced to the signal, which listeners found disturbing. Partially for this reason, the attention signal subspace approaches have received in literature was mainly in automatic speech recognition problems.

E. *Implementation of EVRC Noise Reduction*

The Enhanced Variable Rate Coder (EVRC) is the standard coder for use with the IS-95x Rate 1 air interface (CDMA) [5,8]. It employs an adaptive noise suppression filter, which is used as a baseline reference for the algorithm presented in this paper. Since it is a widely used "real-world" implementation of a noise reduction algorithm, it is worth examining in some detail. Some simplification for brevity was done to illustrate the algorithm more clearly, but as much as possible, the symbols used in the standard document are used. Conceptually, the EVRC's noise suppression is accomplished by summing the outputs of a bank of adaptive filters that span the frequency band of the input signal. The widths of the bands roughly approximate the ear's critical bands. The EVRC noise suppressor works on 10 milliseconds sections of speech data, using the overlap- add method [6], to obtain 104 sample vectors. These vectors are then zero-padded to 128 sample points and transformed using a 128-point Fast Fourier Transform (FFT), windowed by a smoothed trapezoidal window. Reconstruction is done using the overlap-add method, with no windowing. The 128 bins are grouped into 16 channels, approximating non-overlapping critical bands. The energy present in each channel is estimated by calculating the mean magnitude for all frequency bins with channel, and using a exponential average of the form:

$$E_C(m, ch) = \frac{1}{f_H - f_L + 1} \sum_{k=f_L}^{f_H} G_m(k) \qquad (13)$$

$$E(m, ch) = 0.45E(m-1, ch) + 0.55E_C(m, ch) \qquad (14)$$

where m is the index of the time frame, and f_L and f_H are the lowest and highest bin respectively of that particular channel. $G_m(k)$ is the k^{th} bin of the FFT of time frame m. Additionally, the channel energy estimate E(m,ch) is constrained to a minimum of 0.0625 to prevent conditions where a division by zero occurs. The channel energy estimate is then combined with the channel noise energy estimate(see below) to calculate the channel SNR estimate in dB units. The channel SNR values are also used to calculate a voice metric for each frame, which is used to determine if the current frame is noise only. If the frame is considered noise only, the current channel energy estimates are used to update the channel noise estimate E_N, again using exponential averaging. The channel noise estimate is constrained to a minimum of 0.0625.

$$E_N(m+1, ch) = 0.9E_N(m, ch) + 0.1E(m, ch) \qquad (15)$$

For the final channel gain calculation, an overall gain is calculated based on the total noise energy estimate.

$$\gamma_N = -10\log\left(\sum_{ch=0}^{15} E_N(m, ch)\right) \qquad (16)$$

which is constrained to the range $\gamma_N = -13, \ldots\ldots, 0$ .a quantized channel SNR is generated by:

$$\sigma''_Q(ch) = \text{round}\ \left(10\log\left(\frac{E(m, ch)}{E_N(m, ch)}\right)\right) / 0.375 \qquad (17)$$

the result of which is constrained to be between 6 and 89. Now the individual channel gains $\lambda(ch)$ can be computed.

$$\gamma_{dB}(ch) = 0.39(\sigma''_Q(ch) - 6) + \gamma_N \qquad (18)$$

$$\gamma(ch) = \min(1, 10^{\gamma_{dB}(ch))/20}) \qquad (19)$$

These channel gains are then applied to the FFT bins belonging to their respective channels, before the inverse FFT is performed. However, while the EVRC noise suppressor has a concept of critical bands, it does not make use of any other perceptual properties. There is no calculation of masking thresholds, all channels are calculated independently from each other. It should also be noted that the EVRC noise suppressor (and hence the entire coder) is preceded by a high pass filter whose 3 dB cutoff is at about 120 Hz and has a slope of about 80 dB/oct. This removes a large amount of noise which is commonly encountered in mobile applications (like car noise) while not greatly affecting speech quality.

D. *Comparison of Methods*

To compare short-time spectral amplitude (STSA) subtractive methods, the gain curve is the primary point of comparison. The gain curve shows the attenuation of any frequency bin for any given a posteriori SNR, that is the value of G(m) given R (m) from (7). Fig. 1(a) and Fig.1(b) show the gain curves for magnitude and power spectral subtraction respectively. From the plots, it can be seen that the parameter k is dominant in determining the slope of the curve. For small k the attenuation remains small even for very low SNR values.

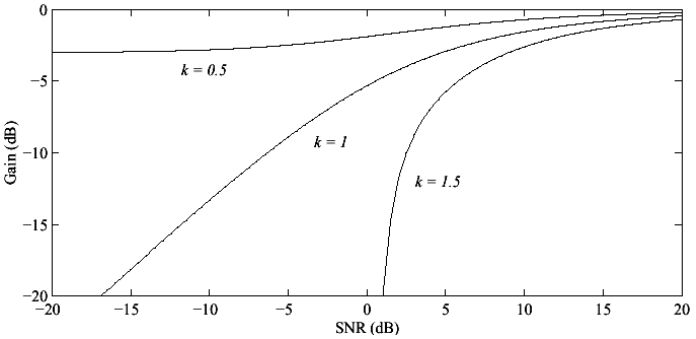
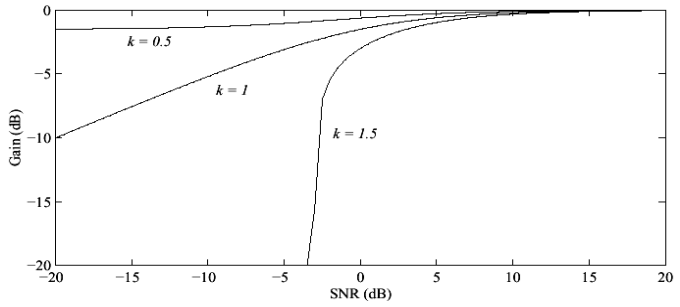


Figure1. Gain curves of spectral subtraction algorithms
(a) Magnitude Spectral Subtraction (a = 1)



Figur1. (b) Power Spectral Subtraction (a = 2)

For k = 1:5 (in general, for k > 1), the spectral subtraction as a noise gate, cutting off completely (assuming $\alpha = 0$) if the SNR drops below:

$$\Re_{\text{off}} = 10\log(K^{\alpha/2} - 1) \quad (\text{dB}) \quad K > 1 \qquad (20)$$

Figure 2 shows the gain curves of some of the other methods described in the previous section. As expected the curve for the Wiener filter is very similar to the power spectral subtraction with k = 1. It is an interesting feature of the Wiener filter that as the SNR decreases, the filter gain becomes equal to the SNR. The other curves on Fig.2 show the gain curves for the MLEE method and the EVRC noise suppressor.

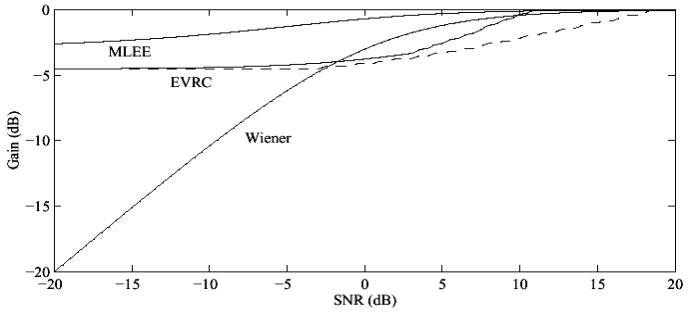


Fig2. Gain curves of selected other methods

The MLEE curve provides very little attenuation, with a maximum attenuation of 3 dB. It is therefore of little use if the intent is to provide significant noise removal. For reference, the EVRC noise suppressor was included. Like the Ephraim and Malah Noise Suppressor, the gain is dependent not only on the a posteriori SNR, but on other values as well. In the case of the EVRC noise suppressor, the gain is not only subject to temporal smoothing, but also on the overall estimate of the noise, as can be seen from equations (15),(19). The two EVRC curves on fig.2 show the gain assuming fixed signal power, but varying noise power (solid line) and fixed noise but varying signal power (dotted line). Both curves are based on the noise power being constant across the whole spectrum.

SUMMARY

In this paper, some methods for reducing or removing acoustic noise are introduced. In particular, methods based on short time Fourier transforms are examined. The problems of noise estimation are briefly discussed. The artifacts introduced by STSA methods are described, and how the spectral subtraction method is modified to counter these artifacts. The EVRC noise suppression algorithm is discussed in some detail. This gives insight into a "real-world" implementation of a noise reduction algorithm. Finally, some of the methods were compared based on the attenuation given an estimate of the SNR.

REFERENCES

[1] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*. Wiley Teub-ner, 1996.

[2] S. J. Godsill and P. J. W. Rayner, Digital Audio Restoration. Springer Verlag, 1998.

[3] J. John R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals.* New York: IEEE Press, 2000.

[4] G. Soulodre, Adaptive *Methods for Removing Camera Noise from Film Soundtracks*. PhD thesis, McGill University, Montreal, Canada, 1998.

[5] T. Painter and A. Spanias, "*Perceptual coding of digital audio*," Proc. IEEE, vol. 88, pp. 451-513, Apr. 2000.

[6] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," IEEE Trans. Speech and Audio Processing, vol. 5, pp. 497-514, Nov. 1997.

[7] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Perceptual filters for audio signal enhancement," J. Audio Eng. Soc., vol. 45, pp. 22-35, Jan/Feb 1997.

[8] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," IEEE Trans. Speech and Audio Processing, vol. 7, pp. 126-137, Mar. 1999.

[9] T. Haulick, K. Linhard, and P. Schr. ogmeier, "Residual noise suppression using psychoacoustic criteria," in Eurospeech 97, (Rhodes, Greece), pp. 1395-1398, Sept. 1997.