Adaptive Enhancement of Speech Signals for Robust ASR

Vivek Tyagi, Christian Wellekens

Institute Eurecom, Sophia Antipolis, France
tyagi@eurecom.fr, wellekens@eurecom.fr

Abstract

Behavior of the least squares filter (LeSF) is analyzed for a class of non-stationary signals that are composed of multiple sinusoids whose frequencies and the amplitudes may vary from block to block and which are embedded in white noise. Analytic expressions for the weights and the output of the LeSF are derived as a function of the block length and the signal SNR computed over the corresponding block. Recognizing that such a sinusoidal model is a valid approximation to the speech signals, we have used LeSF filter estimated on each block to enhance the speech signals embedded in white noise. ASR experiments on a connected digits task, OGI Numbers95 show that the proposed LeSF based features yield an increase in speech recognition performance in various non-stationary noise conditions when compared directly to the un-enhanced speech and noise-robust RASTA filtering technique. Besides achieving noise robustness, this filtering technique yields an enhanced speech signal as a by-product. This is particularly suitable for ASR in mobile telephony networks where the noise robust feature extraction module also performs the speech signal enhancement task without incurring additional computational load.

1. Introduction

Most of the adaptive filtering techniques require an explicit external noise reference to remove additive noise from the a desired signal as discussed in [8]. In situations where an external noise reference for the additive noise is not available, the interfering noise may be suppressed using a Wiener linear prediction filter (for stationary input signal and stationary noise) if there is a significant difference in the bandwidth of the signal and the additive noise [5]. One of the earliest use of the least mean square filtering for speech enhancement is due to Sambur[1]. In his work, the step size of the LMS filter was chosen to be one percent of the reciprocal of the largest eigenvalue of the correlation matrix of the first voiced frame. However, speech being a non-stationary signal, the estimation of the step size based on the correlation matrix of just single frame of the speech signal, may lead to divergence of the LMS filter output. Nevertheless, the treatment in [1] helped to illustrate the efficacy of the LMS algorithm for enhancing naturally occurring signals such as speech. In [5], Zeidler et. al. have analyzed the steady state behavior of the adaptive line enhancer (ALE), an implementation of least mean square algorithm that has applications in detecting and tracking narrow-band signals in broad-band noise. In [4], Anderson et al extended the above mentioned analysis for a stationary input consisting of finite band-width signals in white noise. In this paper, we extend the previous work in [4, 5] for enhancing a class of non-stationary signals that are composed of multiple sinusoids whose frequencies, phases and the amplitudes may vary from block to block and which are embedded in white noise. The key difference in the approach proposed in this paper is that we relax the assumption of the input signal being stationary. Therefore the input signal is blocked into frames and we analyze a L-weight least squares filter (LeSF), estimated on each frame which consists of N samples of the input signal. We have derived the analytical expressions for the impulse response of the L-weight least squares filter (LesF) as a function of the input SNR (computed over the current frame), effective band-width of the signal (due to finite frame length), filter length 'L' and frame length 'N'. Recognizing that such a time-varying sinusoidal model[7] is a reasonable approximation to the speech waveforms, we have applied the block estimated LeSF filter for de-noising speech signals embedded in broad-band noise. A Sinusoidal model is particularly suitable for voiced speech which consists of sinusoids with frequencies at the multiple of the fundamental frequency (pitch).

2. Least Squares filter (LeSF) for signal enhancement

The LeSF filter consists of L weights and the filter coefficients w_k for k $\in$ $[0, 1, 2..L-1]$ are estimated by minimizing the energy of the error signal $e(n)$ over the current frame, $n \in [0, N-1]$.

$$e(n) = x(n) - y(n) \tag{1}$$

$$\text{where } y(n) = \sum_{i=0}^{L-1} w(i)x(n-P-i) \tag{2}$$

The basic operation of the LeSF can be understood intuitively as follows. The autocorrelation sequence of the additive noise $w(n)$ that is broad-band decays much faster for higher lags than that of the speech signal. Therefore the bulk delay P causes de-correlation between the noise components of the input. The LeSF filter responds by adaptively forming a frequency response which has pass-bands centered at the frequencies of the formants of the speech signal. Let **A** denote the $(N+L) \times L$ data matrix[2] of the input frame $\mathbf{x} = [x(0), x(1), ...x(N-1)]$ and **d** denote the $(N+L) \times 1$ desired signal vector which in this case is just a delayed version of signal **x**. The LeSF weight vector **w** is then given by

$$\mathbf{w} = \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{d} \tag{3}$$

As is well known, $\mathbf{A}^H\mathbf{A}$ is a symmetric $L \times L$ Toeplitz matrix whose (i, j) element is the temporal autocorrelation of the signal vector **x** estimated over the frame length [2].

$$\begin{aligned} \left[\mathbf{A}^H\mathbf{A}\right]_{i,j} &= r(|i-j|) && (4) \\ &= \textstyle\sum_{n=0}^{N-|i-j|} x(n)x(n+|i-j|) && (5) \end{aligned}$$

In practice, $\mathbf{A}^H\mathbf{A}$ can always be assumed to be non-singular due to presence of additive noise[2] for filter length $L < N$.

The weight vector **w** in (3) can be obtained using Levinson Durbin algorithm[2] without incurring a significant computational cost.

3. LeSF applied to Sinusoidal model of Speech

As proposed in [7], speech signals can be modeled as a sum of multiple sinusoids whose amplitudes and frequencies can vary from frame to frame. Lets assume that a given frame of speech signal **x** can be approximated as a sum of M sinusoids.

$$\begin{aligned} x(n) &= s(n) + \nu(n) \qquad (6) \\ x(n) &= \sum_{i=1}^{M} A_i \cos(2\pi\omega_i n + \phi_i) + \nu(n) \end{aligned}$$

where $n \in [1, N]$ and $\nu(n)$ is a realization of white noise. Then the k^{th} lag autocorrelation can be shown to be,

$$\begin{gathered} r(k) = \sum_{n=1}^{N-k} x(n)x(n+k) \\ \simeq \sum_{i=1}^{M} (N-k) A_i^2 cos(2\pi f_i k) + N\sigma^2 \delta(k) \end{gathered} \qquad (7)$$

where it is assumed that $N \gg 2\pi/(f_i - f_j)$ for all frequency pairs (i, j) and the noise $\nu(n)$ is white, ergodic and uncorrelated with the signal x(n). The LeSF weight vector is then obtained as the solution of the Normal equations,

$$\begin{gathered} \sum_{k=0}^{L-1} r(l-k) w(k) = r(l+P) \\ l \in [0, 1, 2..L-1] \end{gathered} \qquad (8)$$

The set of L linear equations described in (8) can be solved by elementary methods if the z-transform ($S_{xx}(z)$) of the symmetric autocorrelation sequence ($r(k)$) is a rational function of z [3].

$$S_{xx}(z) = \sum_{k=-\infty}^{\infty} r(k) z^{-k} \qquad (9)$$

Consider then, a conjugate symmetric rational z transform with M pairs of zeros and M pairs of poles.

$$S_{xx}(z) = G \frac{\prod_{m=1}^{M} (z - e^{-\beta_m + j\Psi_m})(z^{-1} - e^{-\beta_m - j\Psi_m})}{\prod_{m=1}^{M} (z - e^{-\alpha_m + j\omega_m})(z^{-1} - e^{-\alpha_m - j\omega_m})} \qquad (10)$$

If the signal **x** is real, then so its autocorrelation sequence, $r(k)$. In this case the power spectrum, $S_{xx}(z)$, has quadruplet sets of poles and zeros because of the presence of conjugate pairs at $z = exp(\pm\alpha_m - \omega_m)$ and $z = exp(\pm\beta_m - \Psi_m)$. Anderson et. al.[4] have derived the general form of the solution to (8) for input signal with rational power spectra such as that described by (10). In this case, the LeSF weights are given by,

$$w(k) = \sum_{m=1}^{M} \left(B_m e^{-\beta_m k + j\Psi_m k} + C_m e^{+\beta_m k + j\Psi_m k} \right) \qquad (11)$$

As can be seen, LeSF consists of an exponentially decaying term and an exponentially growing term attributed to reflection [8], that occurs due to finite filter length L. The value of the coefficients B_m and C_m can be determined by solving the set of coupled equations obtained by substituting the expression for $w(k)$ given in (11) into (8).

To be able to use the general form of the solution of the LeSF filter as in (11), we need a pole-zero model of the input autocorrelation in the form as described in (10). For sufficiently large frame length N, such that filter length $L \ll N$, we can make the following approximation.

$$\begin{gathered} (N-k) \simeq N e^{-k/N} \\ k \in [0, 1, 2, \ldots, L] \text{ and } L \ll N \end{gathered} \qquad (12)$$

Using this approximation in (7), we get,

$$r(k) = N e^{-k/N} \sum_{i=1}^{M} A_i^2 cos(\omega_i k) + N\sigma^2 \delta(k) \qquad (13)$$

In this form, $r(k)$ corresponds to a sum of multiple decaying exponential sequences and its z transform takes up the form,

$$\begin{aligned} S_{xx}(z) &= \frac{1}{\prod_{m=1}^{M} (z - e^{-\alpha_m + j\omega_m})(z^{-1} - e^{-\alpha_m - j\omega_m})} \\ &\quad \times \frac{1}{(z - e^{-\alpha_m - j\omega_m})(z^{-1} - e^{-\alpha_m + j\omega_m})} + \sigma^2 \\ \text{where } \alpha_m &= 1/N \end{aligned} \qquad (14)$$

As explained in [3, 4] and making the following assumptions,

- The pole pairs in (14) lie sufficiently close to the unit circle (easily satisfied as $\alpha \simeq 0$.)
- All the frequency pairs (ω_i, ω_j) in (14) are sufficiently separated from each other such that their contributions to the total power spectrum do not overlap significantly.

the z transform of the total input can be expressed as,

$$\begin{aligned} S_{xx}(z) &= \sigma^2 \frac{\prod_{m=1}^{M} (z - e^{-\beta_m + j\omega_m})(z^{-1} - e^{-\beta_m - j\omega_m})}{\prod_{m=1}^{M} (z - e^{-\alpha_m + j\omega_m})(z^{-1} - e^{-\alpha_m - j\omega_m})} \\ &\quad \times \frac{(z - e^{+\beta_m + j\omega_m})(z^{-1} - e^{+\beta_m - j\omega_m})}{(z - e^{+\alpha_m + j\omega_m})(z^{-1} - e^{+\alpha_m - j\omega_m})} \qquad (15) \\ \text{where } \alpha_m &= 1/N \end{aligned}$$

Corresponding to each of the sinusoidal component in the input signal there are four poles at locations $z = e^{\pm\alpha \pm \omega_m}$ and there are four zeros on the same radial lines as the signal poles but at different distances away from the unit circle. Using the general solution described in (11), which has been derived at length in [4], the solution of the LeSF weight vector to the present problem is,

$$w(n) = \sum_{m=1}^{M} \left(B_m e^{-\beta_m n} + C_m e^{+\beta_m n} \right) \cos \omega_m (n + P) \qquad (16)$$

The values of β_m, B_m and C_m can be determined by substituting (16) and (13) in (8). The l^{th} equation in the linear-system described in (8) has terms with coefficients $exp(-\beta_m l)$, $exp(+\beta_m l)$, $exp(-\alpha l) \cos(\omega_m(l+P))$ and $exp(\alpha l) \cos(\omega_m(l+P))$. Besides these, there are two other kind of terms.

- "Non-stationary" terms that are modulated by a sinusoid at frequency $2\omega_m$ where $m \in [1, M]$. For $\omega_m \neq 0$, $\omega_m \neq \pi$, their total contribution is approximately zero.[1]
- Interference terms that are modulated by a sinusoid at frequency $\Delta\omega = (\omega_i - \omega_j)$ where $(i, j) \in [1, \ldots, M]$. If filter length $L \gg 2\pi/\Delta\omega$, these interference terms approximately sum up to zero and hence can be neglected.

The coefficients of the terms $exp(-\beta_m l)$, $exp(+\beta_m l)$ are the same for each of the L equations and setting them to zero leads to just one equation which relates β_m to α and the SNR. Let ρ_i denote the "partial" SNR of the sinusoid at frequency ω_i i.e $\rho_i = A_i^2/\sigma^2$ and the complementary signal SNR be denoted as $\gamma_i = (\sum_{m=1, m\neq i}^{M} A_i^2)/\sigma^2$. Then we have the following relation,

$$\cosh \beta_i = \cosh \alpha + \frac{\rho_i}{2\gamma_i + \rho_i + 2} \sinh \alpha \qquad (17)$$

There is an interesting case when the sinusoid at frequency ω_i is significantly stronger than the other sinusoids such that γ_i is quite low. This is illustrated in figure (1), where we plot the bandwidth β_i of the LeSF's pass-band that is centered around ω_i as a function of the partial SNR of the i^{th} sinusoid, ρ_i. The complementary signal's SNR is quite low at $\gamma_i = -6.99db$. We plot curves for different "effective" input sinusoid's bandwidth α. From (14), we note that α is reciprocal of frame length N. The vertical line in figure (1) corresponds to the case when $\rho_i = \gamma_i$. We note that for a given partial SNR ρ_i, the LeSF bandwidth (β_i) becomes narrower as the frame length N increases, indicating a better selectivity of the LeSF filter.

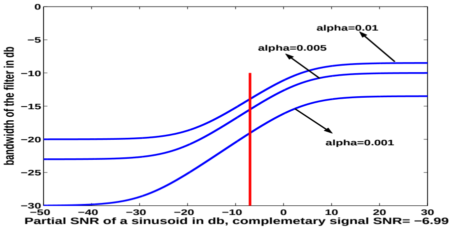

Figure 1: *Plot of the filter bandwidth β_i centered around frequency ω_i as a function of partial sinusoid SNR ρ_i for a given complementary signal SNR $\gamma_i = -6.99db$ and "effective" input bandwidth $\alpha(alpha) = 0.01, 0.005, 0.001$ respectively. The vertical line meets the three curves when $\rho_i = \gamma_i$.*

B_i and C_i in (16) are determined by equating their respective coefficients. The "non-stationary" interference terms between all of the pairs of the frequency (ω_i, ω_j), can be neglected if $(\omega_i - \omega_j) >> 2\pi/L$. This requires that LeSF's frequency resolution $(2\pi/L)$ should be able to resolve the constituent sinusoids.

$$\begin{aligned} B_i &= \frac{2e^{-\beta_i} e^{-\alpha P} (\alpha + \beta_i)^2 (\beta_i - \alpha)}{((\alpha + \beta_i)^2 - e^{-2\beta_i L} (\beta_i - \alpha)^2)} \\ C_i &= \frac{2e^{-\beta_i (2L+1)+1} e^{-\alpha P} (\alpha + \beta_i)(\beta_i - \alpha)^2}{((\alpha + \beta_i)^2 - e^{-2\beta_i L} (\beta_i - \alpha)^2)} \end{aligned} \qquad (18)$$

We note from (17) that the various sinusoids are coupled with each other through the dependence of their bandwidth β_i on the complementary signal SNR γ_i. As a consequence of that B_i, C_i are also indirectly dependent on the powers of the other sinusoids through β_i.

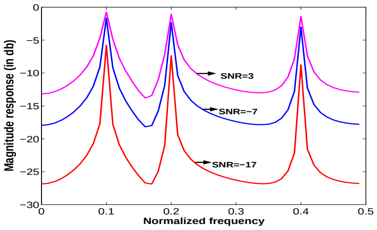

Figure 2: *Plot of the magnitude response of the LeSF filter as a function of the input SNR. The input consists of three sinusoids at normalized frequencies (0.1, 0.2, 0.4) with relative strength* $(1 : 0.6 : 0.4)$ *respectively.*

In figure (2), the magnitude response of the LeSF filter is plotted for various SNR. The input in this case consist of three sinusoids at normalized frequencies (0.1, 0.2, 0.4). The frame length is $N = 500$ and filter length is $(L = 100)$. As the signal SNR decreases, the bandwidth of the LeSF filter starts to decrease in order to reject as much of noise as possible. The LESF filter's gain decreases with decreasing SNR. Similar results were reported in [4, 5] for the case of stationary inputs.

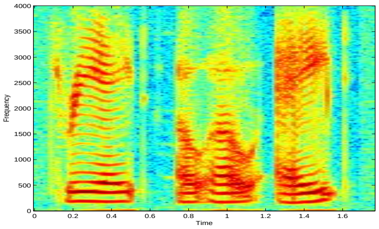

Figure 3: *Clean spectrogram of an utterance from the OGI Numbers95 database*

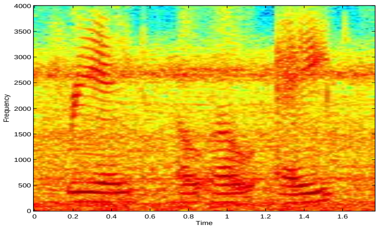

Figure 4: *Spectrogram of the utterance corrupted by F16-cockpit noise at 6dB SNR.*

In Fig. 3, we plot the spectrograms of a clean speech utterance. Fig. 4 and Fig. 5 display the same utterance embedded in F16-cockpit noise at SNR 6dB and its LeSF enhanced version respectively. As can be seen from the spectrograms, the LeSF filter has been able to reject significant amount of additive F-16 cockpit noise [9] from the speech signal.

4. Experiments and Results

In order to assess the effectiveness of the proposed algorithm, speech recognition experiments were conducted on the OGI

[1] due to self cancelling positive and negative half periods of a sinusoid.

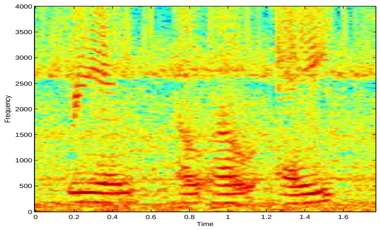

Figure 5: *Spectrogram of the noisy utterance enhanced by a* $(L = 100)$ *tap LeSF filter that has been estimated over blocks of length* $(N = 500)$.

Numbers corpus. This database contains spontaneously spoken free-format connected numbers over a telephone channel. The lexicon consists of 31 words. Speech signals were blocked into frames of 500 samples (62.5ms) each and a 100 tap LeSF filter was derived using (3) for each frame. The relatively high order $(L = 100)$ of the LeSF filter is required to be able to have sufficiently high frequency resolution $(2\pi/L)$ to resolve constituent sinusoids. The speech frame was then filtered through its corresponding LeSF filter to derive an enhanced speech frame. Finally MFCC feature vector was computed from the enhanced speech frame. These enhanced LeSF-MFCC were compared to the baseline MFCC features and RASTA-PLP features with the same window size (62.5ms). Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK on the clean training set from the original Numbers corpus. The system consisted of 80 tied-state triphone HMM's with 3 emitting states per triphone and 12 mixtures per state.

To verify the robustness of the features to noise, the clean test utterances were corrupted using Factory and F16 cockpit noise from the Noisex92 [9] database. The speech recognition results for the baseline MFCC, RASTA-PLP and the proposed LeSF-MFCC, in various levels of noise are given in Tables 1 and 2. The proposed LeSF processed MFCC performs significantly better than others in all noise conditions. The slight performance degradation of the LeSF-MFCC in the clean is due to the fact that the LeSF filter being an all-pole filter does not model the valleys of the clean speech spectrum well. As a result, the LeSF filter sometimes amplifies the low spectral energy regions of the clean spectrum. Besides achieving noise robustness, this filtering technique yields an enhanced speech signal as a by-product. This is particularly suitable for ASR in mobile telephony networks where the noise robust feature extraction module also performs the speech signal enhancement task without incurring additional computational load.

Table 1: *Word error rate results for factory noise*

SNR	MFCC	PLP-JRASTA	LeSF MFCC
Clean	5.7	7.8	6.8
12 dB	12.3	12.2	12.0
6 dB	27.1	23.8	21.0
0 db	71.0	59.8	42.6

5. Conclusion

We consider a class of non-stationary signals as input that are composed of multiple sinusoids whose frequencies and the amplitudes may vary from block to block and which are embedded in the white noise. We have derived the analytical expressions for the impulse response of the L-weight least squares filter (LeSF) as a function of the input SNR (computed over the current frame), effective band-width of the signal (due to finite frame length), filter length 'L' and frame length 'N'. Recognizing that such a time-varying sinusoidal model[7] is a reasonable approximation to the speech waveforms, we have applied the block estimated LeSF filter for de-noising speech signals embedded in broad-band noise. The proposed technique leads to a significant improvement in ASR performance as compared to RASTA-PLP and the usual MFCC features.

Table 2: *Word error rate results for f16 noise*

SNR	MFCC	PLP-JRASTA	LeSF MFCC
Clean	5.7	7.8	6.8
12 dB	13.6	14.2	12.4
6 dB	28.4	25.3	20.6
0 db	72.3	59.2	41.2

6. Acknowledgment

This work has been supported by EC 6^{th} Framework project DIVINES under the contract number FP6-002034.

7. References

[1] M. R. Sambur, " Adaptive noise canceling for Speech signals," In IEEE Trans. on ASSP, vol. ASSP-26, No.5, October 1978.

[2] S. Haykin, Adaptive Filter Theory, Prentice-Hall Publishers, N.J., USA, 1993.

[3] E. Satorius, J. Zeidler and S. Alexander, " Linear predictive digital filtering of narrowband processes in additive broad-band noise, " Naval Ocean Systems Center, San Diego, CA, Tech. Rep. 331, Nov. 1978.

[4] C. M. Anderson, E. H. Satorius and J. R. Zeidler, " Adaptive Enhancement of Finite Bandwidth Signals in White Gaussian Noise, " In IEEE Trans. on ASSP, Vol. ASSP-31, No.1, February 1983.

[5] J. R. Zeidler, E. H. Satorius, D. M. Chabries and H. T. Wexler, " Adaptive Enhancement of Multiple Sinusoids in Uncorrelated Noise, " In IEEE Trans. on ASSP, Vol. ASSP-26, No. 3, June 1978.

[6] H. Hermansky, N. Morgan, " Rasta Processing of Speech," IEEE Trans. on SAP, vol.2, no.4, October 1994.

[7] R. J. McAulay and T. F. Quatieri, " Speech Analysis/Synthesis Based on a Sinusoidal Representation, " In IEEE Trans. on ASSP, Vol. ASSP-34, No. 4, August 1986.

[8] B Widrow et. al., " Adaptive noise cancelling: Principles and applications, " Proc. IEEE, vol.65, pp 1692-1716, Dec 1975.

[9] A. Varga, H. Steeneken, M. Tomlinson and D. Jones, " The NOISEX-92 study on the effect of additive noise on automatic speech recognition, " Technical report, DRA Speech Research Unit, Malvern, England, 1992.