Least Squares filtering of Speech Signals for Robust ASR ⋆

Vivek Tyagi [a,b,*] Christian Wellekens [a,b] Dirk T. M Slock [a,b]

[a] *Institute Eurecom, P.O Box: 193, Sophia-Antipolis, France.*

[b] *Swiss Federal Institute of Technology, Lausanne, Switzerland.*

Abstract

The behavior of the least squares filter (LeSF) is analyzed for a class of non-stationary signals that are either (a) composed of multiple sinusoids (voiced speech) whose frequencies, phases and the amplitudes may vary from block to block or, (b) are output of an all-pole filter excited by white noise input (unvoiced speech segments) and which are embedded in white noise. In this work, 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. We have used LeSF filter estimated on each block to enhance the speech signals embedded in white noise as well as other realistic noises such as factory noise and an aircraft cockpit noise. Automatic speech recognition (ASR) experiments on a connected numbers task, OGI Numbers95[29] show that the proposed LeSF based features provide a significant improvement in speech recognition accuracies in various non-stationary noise conditions when compared directly to the un-enhanced speech, spectral subtraction and noise robust CJ-RASTA-PLP features.

Key words: Least squares, Adaptive filtering, speech enhancement, robust speech recognition

1 Introduction

Speech enhancement, amongst other signal de-noising techniques, has been a topic of great interest for past several decades. The importance of such tech-

⋆ This work was supported by European Commission's 6th Framework Program project, DIVINES under the contract number FP6-002034.

* Corresponding author: Vivek Tyagi

Email addresses: tyagi@eurecom.fr, welleken@eurecom.fr, slock@eurecom.fr (Dirk T. M Slock).

niques in speech coding and automatic speech recognition systems can only be understated. Towards this end, adaptive filtering techniques have been shown to be quite effective in various signal de-noising applications. Some representative examples are echo cancellation[15], data equalization[16–18], narrow-band signal enhancement[14,19], beamforming[20–22], radar clutter rejection[23], system identification[24,25] and speech processing[14].

Most of the above mentioned representative examples require an explicit external noise reference to remove additive noise from the desired signal as discussed in [14]. 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 [14,11,9]. One of the earliest use of the least mean square (LMS) filtering for speech enhancement is due to Sambur[7]. 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 exposition in [7] helped to illustrate the efficacy of the LMS algorithm for enhancing naturally occurring signals such as speech. In [11], 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. Specifically, they have shown that for a stationary input consisting of multiple ($'N'$) sinusoids in white noise, the L-weight ALE, can be modeled by the $L \times L$ Wiener-Hopf matrix equation and that this matrix can be transformed into a set of $2N$ coupled linear equations. They have derived the analytical expression for the steady-state L-weight ALE filter as function of input SNR and the interference between the input sinusoids. It has been shown that the coupling terms between the input sinusoid pairs approach zero as the ALE filter length increases.

In [9], Anderson et al extended the above mentioned analysis for a stationary input consisting of finite band-width signals in white noise. These signals consist of white Gaussian noise (WGN) passed through a filter whose band-width α is quite small relative to the Nyquist frequency, but generally comparable to the bin width $1/L$. They have derived analytic expressions for the weights and the output of the LMS adaptive filter as function of input signal band-width and SNR, as well as the LMS filter length and bulk delay 'z^{-P}' (please refer to Fig. 1).

In this paper, we extend the previous work in [9,11] for enhancing a class of non-stationary signals that are composed of either (a) multiple sinusoids (voiced speech) whose frequencies and the amplitudes may vary from block

to block or (b) are the output of an all-pole filter excited by white noise input (unvoiced speech segments) and which are embedded in white noise. The number of sinusoids may also vary from block to block. The key difference in the approach proposed in this paper is that we relax the assumption of the input signal being stationary. The method of least squares may be viewed as an alternative to Wiener filter theory pg.483 [8]. Wiener filters are derived from *ensemble averages* and they require good estimates of the clean signal power spectral density (PSD) as well as the noise PSD. Consequently, one filter (optimum in a probabilistic sense) is obtained for all realizations of the operational environment, assumed to be wide-sense stationary. On the other hand, the method of least squares is *deterministic* in approach. Specifically, it involves the use of time averages over a block of data, with the result that the filter depends on the number of samples used in the computation. Moreover, the method of least squares does not require the noise PSD estimate. 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.

Working under the assumptions that the clean signal spectral vector and noise spectral vector are Gaussian distributed with k^{th} spectral value independent of j^{th} spectral value, Eprahaim and Malah derived the optimum minimum mean square error (MMSE) estimator of the clean speech's spectral amplitude (MMSE-STA)[3] and its log spectral amplitude (MMSE-LSA)[4]. This assumption is valid only if the clean signal and the noise are both stationary processes and the spectrum is estimated over an infinitely long window. Clearly the speech signal is neither a stationary process nor does it have a Gaussian distributed spectrum. Moreover, in most of the situations, the noise is not a stationary process. Besides this, MMSE-LSA, MMSE-STA, spectral subtraction (SS) and Wiener filter (WF) based techniques need a good estimate of noise spectrum. It is often claimed that the estimate of the noise PSD can be obtained from "non-speech" frames which can be detected using a pre-tuned threshold [4,3]. However, if the noise power changes (varying SNR conditions), there is no single threshold which can detect the non-speech frames. Moreover if the noise is non-stationary, the noise PSD estimate obtained through "non-speech" frames may not be able to track the noise statistics quite well as it is dependent on the availability of non-speech frames which are unevenly distributed in an utterance. Martin[2] has proposed a noise PSD estimator based on the minimum statistics. However even this techniques relies on certain parameters which need to be tuned depending on the degree of non-stationarity of the noise. Several researchers have tried to use a multitude of well-tailored tuning-parameters dependent on the a-prior knowledge of non-speech frames [1] ,

[1] For example, just by the design of the speech databases, the initial few frames always correspond to the silence and hence can be used for noise PSD estimation. However in a realistic ASR task such assumptions cannot be made.

highest and lowest SNR range, and several other ad-hoc weighting factors [2] pg. 438 [1] to achieve noise robustness in ASR.

Therefore it is desirable to develop a new enhancement technique that does not require an explicit noise PSD estimate. The least squares filter (LeSF) based techniques fall in this category as they do not explicitly require a noise PSD estimate. Although, the LeSF is optimal only in the case of the additive noise being white, the speech recognition experiments, reported in this paper, indicate that LeSF is also effecitve in case of non-white additive noises such as the factory noise and the aircraft cockpit noise. MFCC features computed from the LESF enhanced speech signal lead to significant ASR accuracy improvements in various noises as well as SNR conditions. 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'.

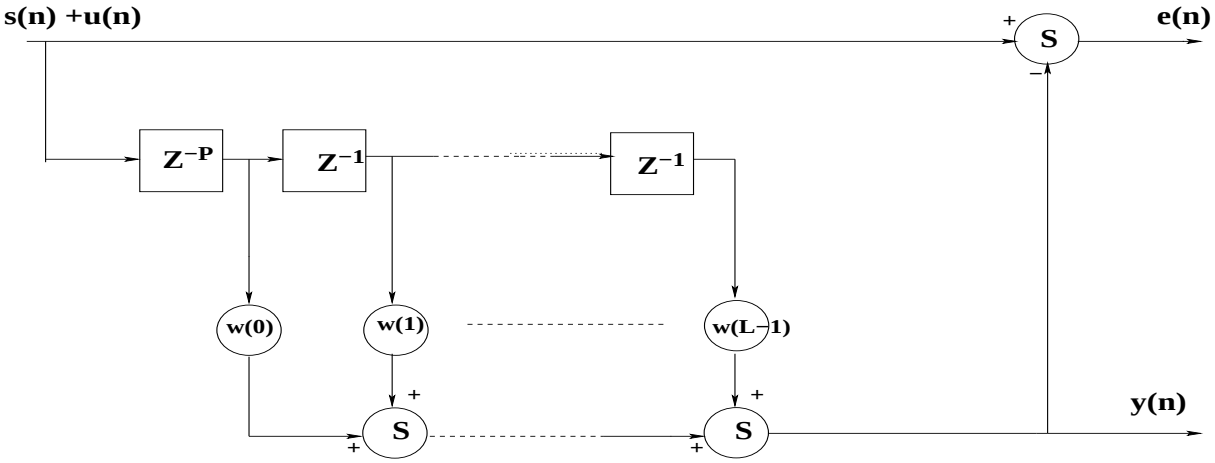

Fig. 1. The basic operation of the LeSF. The input to the filter is noisy speech, $(x(n) = s(n)+u(n))$, delayed by bulk delay $=P$. The filter weights w_k are estimated using the least squares algorithm based on the samples in the current frame. The output of the filter $y(n)$ is the enhanced signal.

2 Least Squares filter (LeSF) for signal enhancement

The basic operation of the LeSF is illustrated in figure (1) and it can be understood intuitively as follows. The autocorrelation sequence of the additive noise $u(n)$ that is broad-band decays much faster for higher lags than that of the speech signal. Therefore the use of a large filter length ('L') and the delay P

[2] such as raising the Wiener filter or spectral subtraction gain function to a certain power which is empirically tuned, dependent on the SNR conditions.

causes de-correlation between the noise components of the input signal, namely $(u(n-L-P+1), u(n-L-P+2), ..., u(n-P))$ and the noise component of the reference signal, namely $(u(n))$. *It is worth noting that a longer filter length L will also help to cause a de-correlation between the noise appearing at the k^{th} filter tap, namely, $u(n-k-P+1)$ (where, $k \sim L$) and the noise component of the reference signal, namely, $u(n)$. This is due to the fact that the broad-band noise's auto-correlation coefficients decay quite rapidly for higher lags.* 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 while rejecting as much of broad-band noise (whose spectrum lies away from the formant positions). Denoting the clean and the additive noise signals by $s(n)$ and $u(n)$ respectively, we obtain the noisy signal $x(n)$.

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

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, $n \in [0, N-1]$.

$$e(n) = x(n) - y(n) \tag{2}$$

$$\text{where } y(n) = \sum_{i=0}^{L-1} w(i)x(n-P-i) \tag{3}$$

Let $\mathbf{A}$ denote the $(N+L)\times L$ data matrix[8] of the input frame $\mathbf{x} = [x(0), x(1), ...x(N-1)]$ and $\mathbf{d}$ denote the $(N+L)\times 1$ desired signal vector which in this case is signal $\mathbf{x}$ appended by L zeros. The LeSF weight vector $\mathbf{w}$ is then given by

$$\mathbf{w} = \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{d} \tag{4}$$

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 $\mathbf{x}$ estimated over the frame length [8].

$$\left[\mathbf{A}^H\mathbf{A}\right]_{i,j} = r(|i-j|) \tag{5}$$

$$= \sum_{n=0}^{N-|i-j|} x(n)x(n+|i-j|) \tag{6}$$

In practice, $\mathbf{A}^H\mathbf{A}$ can always be assumed to be non-singular due to presence of additive noise[8] for filter length $L < N$. The weight vector $\mathbf{w}$ in (4) can be

obtained using Levinson Durbin algorithm[8] without incurring a significant computational cost.

3 LeSF applied to Speech

In this section, we will analytically solve (4) to obtain the LeSF **w**. We model voiced speech using sinusoidal model[13], while unvoiced speech is modeled by a source-filter model. However, we show that the functional form of the equations remain the same except for a change in the parameter values.

3.1 Voiced Speech

As proposed in [13], voiced speech signals can be modeled as a sum of multiple sinusoids whose amplitudes, phases and frequencies can vary from frame to frame. Let us assume that a given frame of speech signal **s(n)** can be approximated as a sum of M sinusoids. The number of sinusoids M may vary from block to block. Then the noisy signal $x(n)$ can be expressed as

$$x(n) = \sum_{i=1}^{M} A_i \cos(\omega_i n + \phi_i) + u(n) \tag{7}$$

where $n \in [0, N-1]$ and $u(n)$ is a realization of white noise. Then the k^{th} lag autocorrelation can be shown to be,

$$\begin{aligned} r(k) &= \textstyle\sum_{n=0}^{N-k-1} x(n)x(n+k) \\ &\simeq \textstyle\sum_{i=1}^{M}(N-k)A_i^2 cos(2\pi f_i k) + N\sigma^2\delta(k) \end{aligned} \tag{8}$$

where it is assumed that the noise $u(n)$ is white, ergodic and uncorrelated with the signal s(n) and $N \gg 1/(f_i - f_j)$ for all frequency pairs (i, j). The latter condition ensures that all the interference terms between all the sinusoids pairs (i, j) sum up to zero. The LeSF weight vector $w(k)$ is then obtained as the solution of the Normal equations,

$$\begin{aligned} \textstyle\sum_{k=0}^{L-1} r(l-k)w(k) &= r(l+P) \\ l &\in [0, 1, 2..L-1] \end{aligned} \tag{9}$$

An enhancement example is illustrated in Fig.2. The first pane displays the magnitude spectrum of a clean, voiced speech frame. The second pane shows

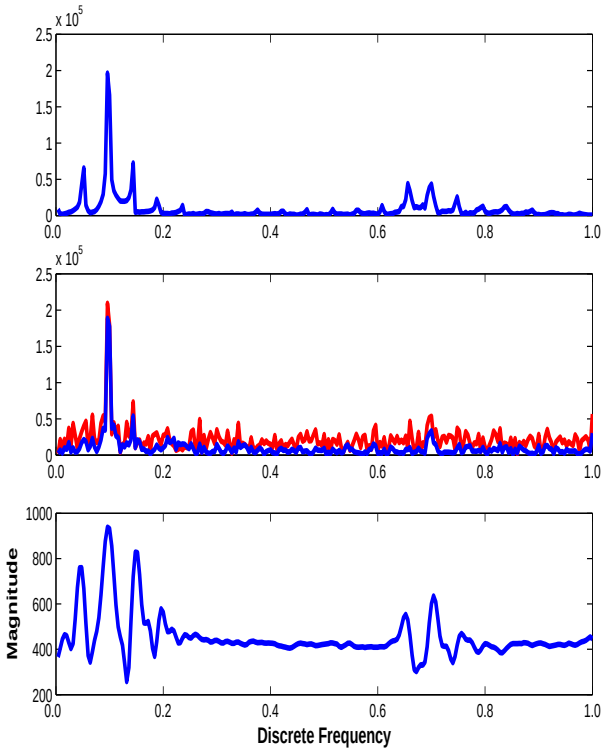

Fig. 2. *The first pane displays spectral magnitude of a clean speech segment. Second pane displays the spectral magnitude of the same segment corrupted by white noise (red curve), whereas blue curve corresponds to the spectral magnitude of the enhanced signal. The third pane displays the frequency response of the LeSF filter that enhanced the noisy segment and was estimated over the noisy segment itself.*

spectra of the same segment embedded in white noise (red curve) and the spectrum of the enhanced signal (blue curve). As can be noticed, the broad band white noise has been attenuated while the harmonics have been retained. The third pane shows the magnitude response of the LeSF filter used for enhancing this segment and it was estimated over the noisy segment itself. As can be noticed the LeSF filter automatically puts the pass-bands around the harmonics, thus enhancing the signal while rejecting the broad-band noise. In the sections to follow, we will present further examples and performance specifications of the LeSF filter for enhancing noisy speech signals.

The set of L linear equations described in (9) 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' [10]. $S_{xx}(z)$ is given by,

$$S_{xx}(z) = \sum_{k=-\infty}^{\infty} r(k) z^{-k} \tag{10}$$

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

$$S_{xx}(z) = G\frac{\prod_{m=1}^{M}(z - e^{-\beta_m + j\Psi_m})(z^{-1} - e^{-\beta_m - j\Psi_m})}{\prod_{m=1}^{M}(z - e^{-\alpha_m + j\omega_m})(z^{-1} - e^{-\alpha_m - j\omega_m})} \tag{11}$$

If the signal $\mathbf{x}$ is real, then so is 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 $z = exp(\pm\alpha_m \pm \omega_m)$ and $z = exp(\pm\beta_m \pm \Psi_m)$. Anderson et. al.[9] have derived the general form of the solution to (9) for input signal with rational power spectra such as that described by (11). In this case, the LeSF weights are given by,

$$w(k) = \sum_{m=1}^{M} \Big(B_m e^{-\beta_m k} \cos(\Psi_m k) + C_m e^{+\beta_m k} \cos(\Psi_m k) \Big) \tag{12}$$

As can be seen, LeSF consists of an exponentially decaying term and an exponentially growing term attributed to reflection [14], 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 (12) into (9).

To be able to use the general form of the solution of the LeSF filter as in (12), we need a pole-zero model of the input autocorrelation in the form as described in (11). For sufficiently large frame length N, such that filter length $L \ll N$, we can make the following approximation.

$$\begin{aligned} (N - k) &\simeq Ne^{-k/N} \\ k \in [0, 1, 2, \ldots, L] &\text{ and } L \ll N \end{aligned} \tag{13}$$

The above can be verified by using the Taylor series expansion of $Ne^{-\alpha k}$ and using only the linear term as $k \ll N$. We call $(\alpha = 1/N)$ as α^{voiced}. Using this approximation in (8), we get,

$$r(k) = Ne^{-\alpha^{voiced} k} \sum_{i=1}^{M} A_i^2 cos(\omega_i k) + N\sigma^2\delta(k) \tag{14}$$

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

$$
\begin{aligned}
S_{xx}(z) = & \sum_{m=1}^{M} \frac{NA_i^2(1-e^{-2\alpha})}{2} \times \\
& \Big(\frac{1}{(z-e^{-\alpha_m+j\omega_m})(z^{-1}-e^{-\alpha_m-j\omega_m})} \\
& + \frac{1}{(z-e^{-\alpha_m-j\omega_m})(z^{-1}-e^{-\alpha_m+j\omega_m})}\Big) + N\sigma^2 \\
\text{where } \alpha_m = & \alpha^{voiced} = 1/N \ \forall\, m \in [1.M]
\end{aligned}
\tag{15}
$$

3.2 Unvoiced Speech

We model unvoiced speech $s(n)$ as the output of an all pole transfer function (whose poles are at $z = e^{-\alpha_i^{unvoiced} \pm \omega_i}$) excited by a white noise signal $e(n)$. Specifically,

$$
S(z) = \frac{E(z)}{\prod_{i=1}^{Q}(z-e^{-\alpha_i^{unvoiced}+j\omega_i})(z-e^{-\alpha_i^{unvoiced}-j\omega_i})} \tag{16}
$$

where $S(z), E(z)$ are the z-transforms of unvoiced speech signal $s(n)$ and white noise excitation signal $e(n)$ respectively. Then it can be shown that the autocorrelation coefficients of the unvoiced speech are also decaying exponentials (pg. 118,[8]) i.e

$$
r_{unvoiced}(k) = \sum_{i=1}^{Q} e^{-\alpha_i^{unvoided} k} cos(\omega_i k), \tag{17}
$$

where the decaying factor $\alpha_i^{unvoiced} > \alpha^{voiced} = 1/N$ (where N is the block length). This is due to the fact that voiced speech has sharper spectral peaks than the unvoiced speech. Consequently the autocorrelation coefficients of the unvoiced speech decay much faster than those of the voiced speech. However, the functional form for the autocorrelation coefficients of the voiced and unvoiced speech is the same, except that $\alpha^{voiced} < \alpha^{unvoiced}$. In presence of white noise, the power spectral density of the noisy unvoiced speech segment is given by,

$$S_{xx}(z) = \sum_{i=1}^{Q} \frac{NA_i^2(1-e^{-2\alpha_i})}{2} \times \left(\frac{1}{(z - e^{-\alpha_i + j\omega_i})(z^{-1} - e^{-\alpha_i - j\omega_i})} + \frac{1}{(z - e^{-\alpha_i - j\omega_i})(z^{-1} - e^{-\alpha_i + j\omega_i})} \right) + N\sigma^2 \tag{18}$$

where α_i is a decay factor of the i^{th} pole pair. We note that the functional form of the power spectral densities in (18) and (15) are the same except that α_i in (18) will in general be greater than α^{voiced} in (15). Therefore the functional form of the LeSF filter $\mathbf{w}$ in (12) remains the same for both voiced and unvoiced speech. Its just that for the unvoiced speech the bandwidth of the pass-bands of the LeSF will be wider than that of voiced LeSF. In Fig. 3, we show a transfer function with two complex-pole pairs (at conjugate symmetric positions) that is used to synthesize unvoiced speech by exciting it with white noise. First pane shows the pole-zero plot. Second pane shows the frequency response of this all-pole model. In the third pane, blue, red and green curves are the FFT magnitudes of the clean speech, noisy speech corrupted by white noise at SNR -3dB and the LeSF enhanced speech respectively. The fact that the green curve matches the blue curve closely, shows that the LeSF has been able to filter out the noise component.

3.3 Analytic form of LeSF

From now onward we will not make any distinction between the exponential decay factors α^{voiced} and $\alpha^{unvoiced}$ as the functional form of the equations remain the same. Therefore the following discussion is valid for both voiced speech and unvoiced speech.

To be able to use the general form of the solution of the LeSF filter as in (12), we need a pole-zero model of the input autocorrelation in the form as described in (11). Under the approximation that the decaying exponentials are widely spaced along the unit circle, the power spectrum $S_{xx}(z)$ in (15) that consists of sum of certain terms can be approximated by a ratio of the product of terms (of the form $(z - e^{\rho + j\theta})$), leading to a rational 'z' transform. Specifically, as explained in [10,9] and making the following assumptions,

- The pole pairs in (15) lie sufficiently close to the unit circle (easily satisfied as $\alpha \simeq 0$.)
- All the frequency pairs (ω_i, ω_j) in (15) are sufficiently separated from each other such that their contribution to the total power spectrum do not overlap significantly.

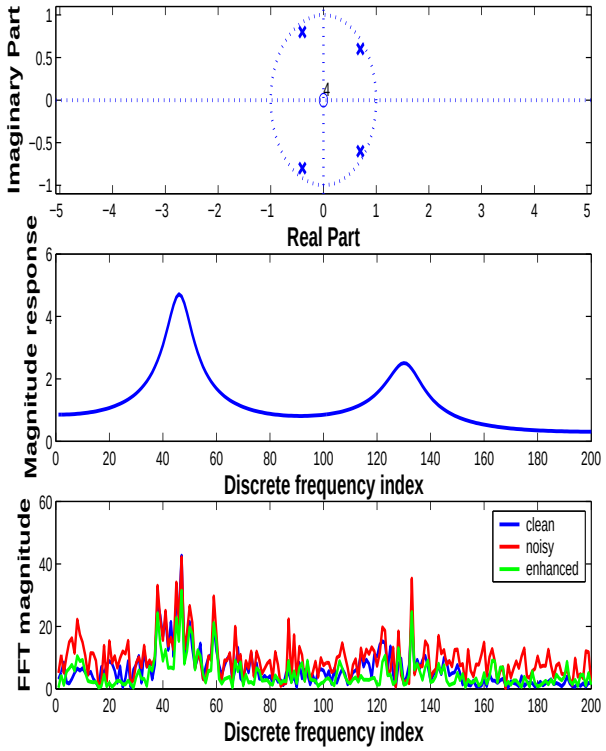

Fig. 3. *A example of a two-formant vocal-tract frequency response which is excited by white noise to synthesize unvoiced speech.*

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

$$S_{xx}(z) = \sigma^2 \frac{\prod_{m=1}^{M}(z - e^{-\beta_m + j\omega_m})(z - e^{+\beta_m + j\omega_m})}{\prod_{m=1}^{M}(z - e^{-\alpha_m + j\omega_m})(z - e^{+\alpha_m + j\omega_m})} \times \frac{(z - e^{+\beta_m - j\omega_m})(z - e^{-\beta_m - j\omega_m})}{(z - e^{+\alpha_m - j\omega_m})(z - e^{-\alpha_m - j\omega_m})} \tag{19}$$

where $\alpha_m = 1/N$

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 (12), which has been derived at length in [9], the solution of the LeSF weight vector to the present problem is,

$$w(n) = \sum_{m=1}^{M} \left(B_m e^{-\beta_m n} + C_m e^{+\beta_m n} \right) \cos \omega_m (n + P) \tag{20}$$

The values of β_m, B_m and C_m can be determined by substituting (20) and (14) in (9). The l^{th} equation in the linear-system described in (9) 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 that can be neglected. The detailed analytic solution is presented in the appendix A. *We note that the analytic solution for the filter weights $w(n)$ has been developed only for the special when the noise is white. Unfortunately, it is not possible to derive a closed form (analytic) solution of the filter weights $w(n)$ for non-white noises. However, as the filter weight $w(n)$ is a continuous function of the noise autocorrelation coefficients (A.6) and the white noise is the limiting case of the broad-band noise when the bandwidth becomes infinite (or equal to the Nyquist frequency for the discrete systems, as is the case here), we expect the following. The filter weight $w(n)$ for a non-white and broad-band noise will approximately follow a behavior similar to the case of the $w(n)$ when the noise is white. However, if the noise is significantly narrowband, then this discussion does not hold true.*

- "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. [3]
- 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,

$$\cosh \beta_i = \cosh \alpha + \frac{\rho_i}{2\gamma_i + \rho_i + 2} \sinh \alpha \tag{21}$$

There are two interesting cases. First case is when the sinusoid at frequency ω_i is significantly stronger than other sinusoids such that γ_i is quite low. This is illustrated in figure (4), where we plot the bandwidth β_i of the LeSF's passband 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 (15), we note that α is reciprocal of frame length N. The vertical line in figure (4) corresponds to the case when $\rho_i = \gamma_i$. We note that for a given partial SNR ρ_i, the LeSF bandwidth becomes narrower as the frame length N increases, indicating a better selectivity of the LeSF filter.

In figure (5), we plot the bandwidth β_i as a function of ρ_i for the cases when complementary signal SNR is high at $\gamma_i = 10db$ and is low at $\gamma_i = -6.99db$.

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

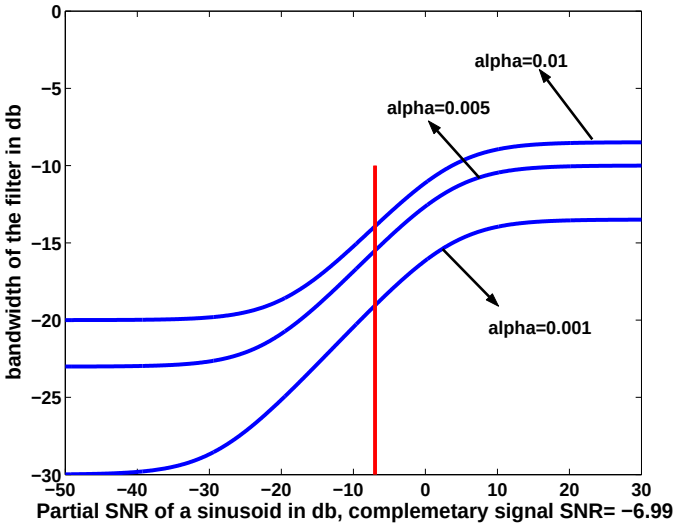

Fig. 4. *Plot of the filter bandwidth β_i centered around frequency ω_i as a function of partial sinusoid SNR ρ_i for a given complementary signal SNR $\gamma_i = -6.99db$ and "effective" input bandwidth $\alpha(alpha) = 0.01, 0.005, 0.001$ respectively. The vertical line meets the three curves when $\rho_i = \gamma_i$.*

The two dots correspond to the case when $\rho_i = \gamma_i$. We note that $\gamma_i = 10db$ corresponds to a signal with high overall SNR [4] . Therefore the cross-over point $(\gamma_i = \rho_i)$ for low γ_i occurs at narrower bandwidth as compared to high γ_i case. This is so because in the former case the overall signal SNR is low and thus the LeSF filter has to have narrower pass-bands to reject as much of noise as possible.

B_i and C_i in (20) 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.

$$B_i = \frac{2e^{-\beta_i}e^{-\alpha P}(\alpha+\beta_i)^2(\beta_i-\alpha)}{((\alpha+\beta_i)^2 - e^{-2\beta_i L}(\beta_i-\alpha)^2)}$$
$$C_i = \frac{2e^{-\beta_i(2L+1)+1}e^{-\alpha P}(\alpha+\beta_i)(\beta_i-\alpha)^2}{((\alpha+\beta_i)^2 - e^{-2\beta_i L}(\beta_i-\alpha)^2)} \qquad (22)$$

We note from (21) 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.

In Fig.6, the magnitude response of the LeSF filter is plotted for various SNR. The input in this case consist of three sinusoids at normalized frequencies (

[4] As overall SNR of the signal $= 10\log_{10}(10^{10\gamma_i} + 10^{10\rho_i})$

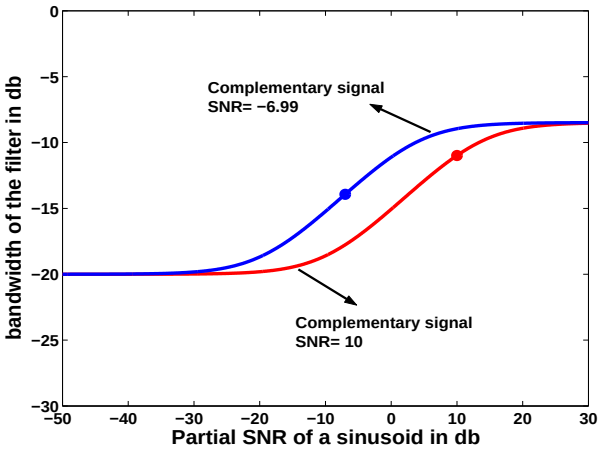

Fig. 5. *Plot of the filter bandwidth* β_i *centered around frequency* ω_i *as a function of partial sinusoid SNR* ρ_i *for given complementary signal SNRs* $\gamma_i = -6.99db, 10db$ *respectively. The "effective" input bandwidth* $\alpha(alpha) = 0.01$ *for both the curves. The two dots correspond to the cases when the partial SNR* ρ_i *is equal to complementary signal SNR* γ_i.

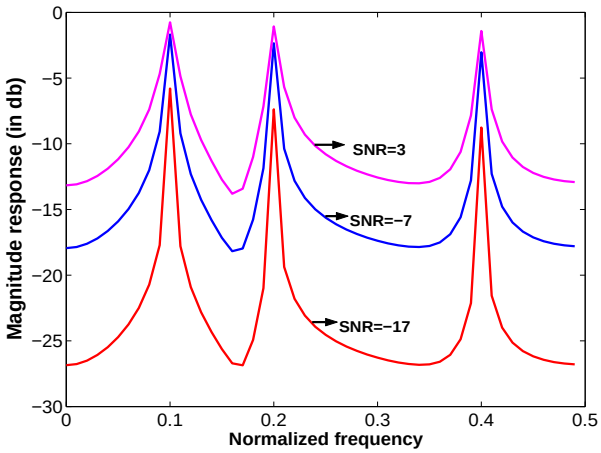

Fig. 6. *Plot of the magnitude response of the LeSF filter as a function of the input SNR. The input consists of three sinusoids at normalized frequencies (0.1, 0.2, 0.4) with relative strength* $(1 : 0.6 : 0.4)$ *respectively.*

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 [9,11] for the case of stationary inputs.

In Fig.7, we plot the spectrograms of a clean speech utterance. Fig.8 and Fig.9 display the same utterance embedded in white noise at SNR 6dB and its LeSF enhanced version respectively. As can be see from these spectrograms, LeSF has been able to reject the additive white noise to a large extent while retaining most of the speech signal. Fig.10 and Fig.11 display the same utterance

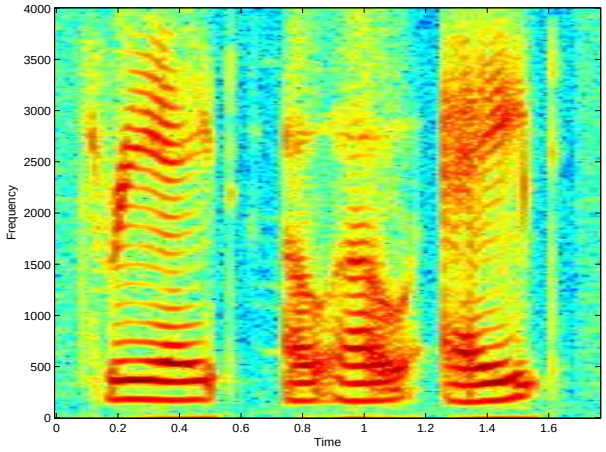

Fig. 7. *Clean spectrogram of an utterance from the OGI Numbers95 database*

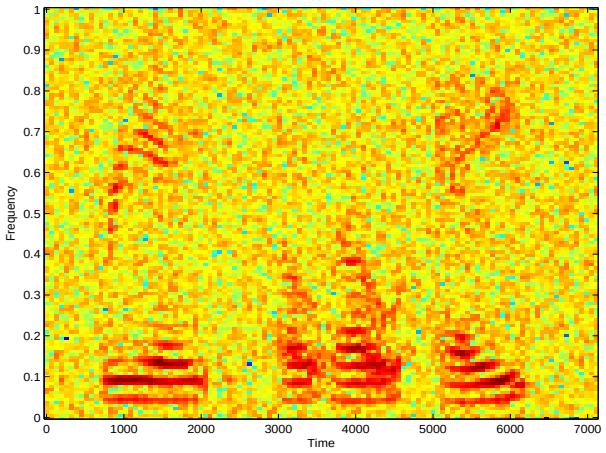

Fig. 8. *Spectrogram of the utterance corrupted by White noise at 6dB SNR.*

embedded in F16-cockpit noise at SNR 6dB and its LeSF enhanced version respectively. As can be seen from the spectrograms, except for the narrow-band noise component centered around 2.5KHz, the LeSF filter has been able to reject significant amount of additive F-16 cockpit noise [29] from the noisy speech signal. By design, the LeSF puts pass-bands around those regions of the input signal's spectrum that has high spectral energy density. As a consequence of that, the LeSF has not been able to well attenuate the narrow band noise in the spectrograms of Fig.10 and Fig.11. However, it has attenuated the broad-band noise components quite well.

4 Gain of the LeSF filter

The LeSF filter output consist of filtered sinusoids and the filtered noise signal. For the input signal described by (7) that is filtered by a LeSF filter with coefficients as in (20), the output filtered signal power P_{signal} and the output

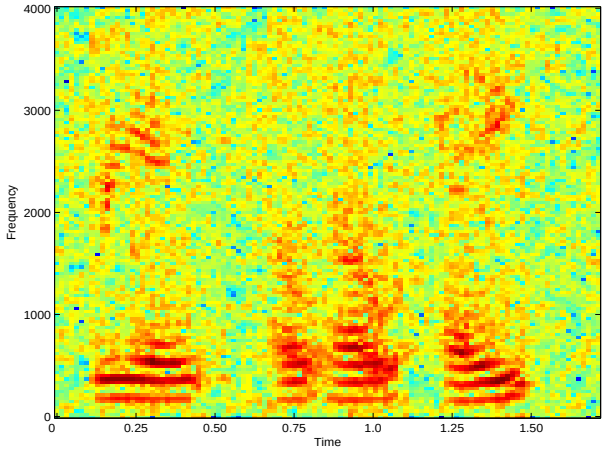

Fig. 9. *Spectrogram of the noisy utterance (white noise) enhanced by a* ($L = 100$) *tap LeSF filter that has been estimated over blocks of length* ($N = 500$).

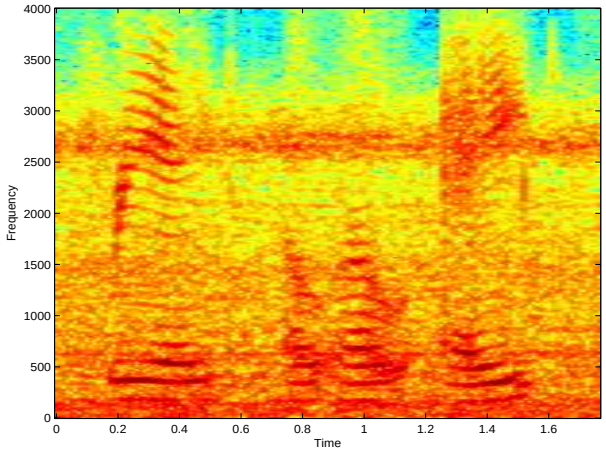

Fig. 10. *Spectrogram of the utterance corrupted by F16-cockpit noise at 6dB SNR.*

filtered noise power P_{noise} are approximately [5] given by,

[5] assuming $N \gg L$ such that initial L samples can be used to initialize the filter

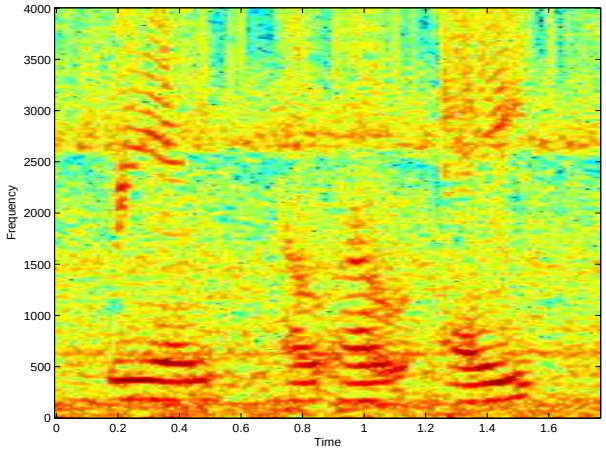

Fig. 11. *Spectrogram of the noisy utterance (F16-cockpit noise) enhanced by a* $(L = 100)$ *tap LeSF filter that has been estimated over blocks of length* $(N = 500)$.

$$P_{signal} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} w(i)w(j)r(|i-j|) \tag{23}$$

$$= \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} w(i)w(j) \times \frac{(N-|i-j|)}{N} \sum_{m=1}^{M} A_m^2 cos(2\pi f_m(i-j)) \tag{24}$$

$$P_{noise} = \sum_{n=0}^{L-1} \sigma^2 w^2(n)$$

$$= \sum_{i=1}^{M} \left[B_i^2 e^{-\beta_i L} + C_i^2 e^{\beta_i L}\right] \frac{\sigma^2 \sinh(\beta_i L)}{2\beta_i} + \sigma^2 BCL \tag{25}$$

The output SNR of the LeSF filter is given by (23) divided by (25). The LeSF gain is given by the ratio of the output SNR to the input $SNR = \sum_{i=1}^{M} A_i^2/(2\sigma^2)$.

In Fig. 12, we plot the LeSF broadband gain as a function of input SNR, for a fixed block length of $N = 500$ and varying filter lengths $(L = 100, 80, 60)$. As can be noted from the Fig. 12, the LeSF broadband gain approaches a horizontal asymptote for decreasing input SNR. This is in agreement with the fact that the bandwidth β_i of the LeSF filter approaches a horizontal asymptote (Fig. 4) for decreasing SNR. We note from Fig. 12 that the LeSF broadband gain increases as the filter length L increases for a fixed block length $N = 500$. However, since the L tap LeSF filter is estimated using the N samples from a given block, the filter length cannot be increased arbitrarily and is limited from above by the block length 'N'.

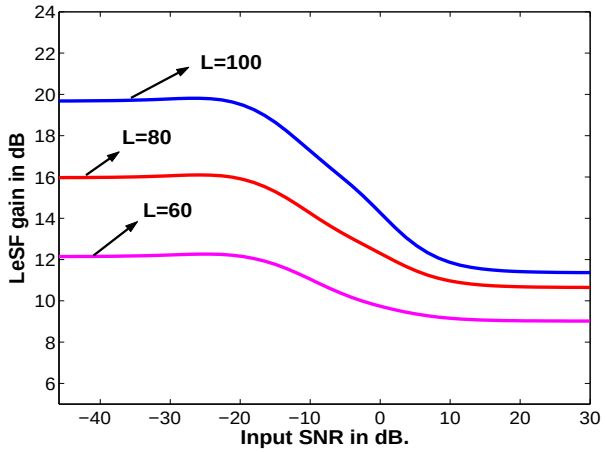

Fig. 12. *LeSF gain plotted as a function of input SNR for fixed block length* $N = 500$ *and various filter lengths* $L = 100, 80, 60$.

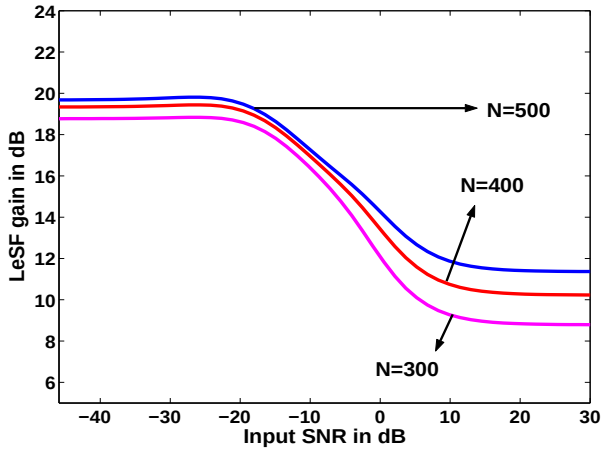

Fig. 13. *LeSF gain plotted as a function of input SNR for fixed filter length* $L = 100$ *and various block lengths* $N = 300, 400, 500$.
wer

In Fig. 13, we plot the LeSF broadband gain as a function of input SNR, for a fixed filter length $L = 100$ and varying block lengths ($N = 300, 400, 500$). We note that the LeSF broadband gain increases as the block length N increases. However, for a non-stationary signal such as speech, as the block length increases, the corresponding power spectrum will become more broadband. Therefore we will not be able to model the corresponding block as a sum of a small number of sinusoids M as done in (7). As a result the number of sinusoids M will be large and possibly closely spaced to each other, leading to significant interference terms between the constituent sinusoids in (8) and (14).

5 Experiments and Results

In order to assess the effectiveness of the proposed algorithm, speech recognition experiments were conducted on the OGI Numbers[28] corpus. This database contains spontaneously spoken free-format connected numbers over a telephone channel. The lexicon consists of 31 words. [6] Hidden Markov Model and Gaussian Mixture Model (HMM-GMM) based speech recognition systems were trained using public domain software HTK[27] 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 White, Factory and F-16 cockpit noise from the Noisex92 [29] database.

5.1 Bulk Delay P

Noting that the autocorrelation coefficients of a periodic signal are themselves periodic with the same period (hence they do not decay with the increasing lag), Sambur[7] has used a bulk delay equal to the pitch period of the voiced speech for its enhancement. However, for the un-voiced speech a high bulk delay will result in a significant distortion by the LeSF filter as its autocorrelation coefficients decay much more rapidly than those of the voiced speech. Therefore, we kept the bulk delay at '$P = 1$' as a good choice for enhancing both the voiced and un-voiced speech frames.

5.2 Block length N and filter length L

Speech signals were blocked into frames of (N=500) samples (62.5ms) each and a (L=100) tap LeSF filter was derived using (4), through the Levinson-Durbin algorithm, for each frame that could be either voiced or unvoiced. *The relatively high order* ($L = 100$) *of the LeSF filter is required for a twofold reason. Firstly, it provides sufficiently high frequency resolution* ($2\pi/L$) *to resolve the constituent sinusoids in case of the voiced speech. Secondly, it causes decorrelation between the noise appearing at the* k^{th} *filter tap* ($u(n-k-P+1), k \sim L$) *and the noise component of the reference signal* $u(n)$. Each 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

[6] With confusable words such as nine, ninety and nineteen, eight, eighty and eighteen and so forth

MFCC features and noise robust CJ-RASTA-PLP[6] features. The MFCC feature vector computation is the same for the baseline and the LeSF-MFCC features. The only difference is that the MFCC baseline features are computed directly from the noisy speech while the LeSF-MFCC features are computed from LeSF enhanced speech signal. We also compared our technique with the soft-decision spectral subtraction based technique. In [5], authors have used a speech presence probability in conjunction with spectral subtraction to achieve noise robustness. This can be seen as a soft-decision spectral subtraction which has been shown to be superior than hard-decision spectral subtraction by Mcaulay et al.[12]. As the train set, test set and the factory noise environment in our experiments and those of [5] are the same, we quote the ASR results for the factory noise directly from [5]. The authors in [5] propose three features based on the soft-spectral subtraction, which vary in their pre-processing steps and are termed POST-FILT, POWER-FILT and PSIL. In table 1, we quote their ASR word error rates in the factory noise environment, directly from the results reported in[5]. We note that the proposed technique outperforms all three soft-decision spectral subtraction variants.

The speech recognition results for the baseline MFCC, CJ-RASTA-PLP and the proposed LeSF-MFCC, in various levels of noise are given in Tables 2, 3 and 4. All the reported features in this paper have cepstral mean subtraction (CMS). The proposed LeSF processed MFCC performs significantly better than others in various kinds of heavy noise conditions (SNR 6,0). 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.

Table 1

Word error rate results for factory noise using soft-decision spectral subtraction. All features have cepstral mean subtraction.

SNR	LeSF MFCC	POST-FILT	POWER-FILT	PSIL
Clean	6.6	8.1	8.3	7.1
12 dB	11.3	16.2	17.0	15.7
6 dB	20.0	30.7	31.2	28.7
0 db	41.3	63.1	61.9	58.2

Table 5 shows the word error rate of the LeSF enhanced MFCC features for a fixed block length of 500 samples ($62.5ms$) and varying LeSF filter length 'L'. We note that the word error rate decreases as the filter length increases. This is so because a higher filter length results in a sharper frequency response of the LeSF filter(narrower band-width of the passbands), thereby enabling it to reject as much of the broad-band noise as possible that lies away from the frequencies of the constituent sinusoids of the clean signal.

Table 2
Word error rate results for factory noise. Parameters of the LeSF filter, L=100 and N=500. All features have cepstral mean subtraction.

SNR	MFCC	CJ-RASTA-PLP	LeSF MFCC
Clean	5.7	7.8	6.6
12 dB	12.3	12.2	11.3
6 dB	27.1	23.8	20.0
0 db	71.0	59.8	41.3

Table 3
Word error rate results for white noise. Parameters of the LeSF filter, L=100 and N=500. All features have cepstral mean subtraction.

SNR	MFCC	CJ-RASTA-PLP	LeSF MFCC
Clean	5.7	7.8	6.6
12 dB	16.4	14.7	17.3
6 dB	34.6	29.0	24.1
0 db	80.3	66.0	40.4

Table 4
Word error rate results for F16-cockpit noise. Parameters of the LeSF filter, L=100 and N=500. All features have cepstral mean subtraction.

SNR	MFCC	CJ-RASTA-PLP	LeSF MFCC
Clean	5.7	7.8	6.6
12 dB	13.6	14.2	12.5
6 dB	28.4	25.3	21.0
0 db	72.3	59.2	41.0

6 Conclusion

We consider a class of non-stationary signals as input that are composed of either (a) multiple sinusoids (voiced speech) whose frequencies and the amplitudes may vary from block to block or, (b) output of an all-pole filter excited by white noise input (unvoiced speech segments) and which are embedded in 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[13] and the source-filter model

Table 5
Word error rate results for factory noise for varying length, $L = 100, 50, 20$ of the LeSF filter. The block length, N is 500 (62.5ms).

SNR	LeSF L=20	LeSF L=50	LeSF L=100
Clean	9.3	7.3	6.6
12 dB	14.3	12.3	11.3
6 dB	24.4	22.0	20.0
0 db	46.5	43.0	41.3

are a reasonable approximation to the voiced speech and unvoiced speech respectively, we have applied the block estimated LeSF filter for de-noising speech signals embedded in the realistic[29] broad-band noise as commonly encountered on a factory floor and an aircraft cockpit. The proposed technique leads to a significant improvement in ASR performance as compared to noise robust CJ-RASTA-PLP[6], speech presence probability based spectral subtraction[5] and the MFCC features computed from the unprocessed noisy signal.

A Appendix: Solution to the equation

A.1 Autocorrelation over a block of samples

We consider the signal $x(n)$ as

$$x(n) = \sum_{i=1}^{M} A_i \cos(\omega_i n + \phi_i) + u(n) \tag{A.1}$$

where $n \in [0, N-1]$ and $u(n)$ is a realization of white noise. Then the k^{th} lag autocorrelation is given by,

$$
\begin{aligned}
r(k) &= \textstyle\sum_{n=0}^{N-k-1} x(n)x(n+k) \\
&= \textstyle\sum_{n=0}^{N-k-1}(\sum_{i=1}^{M} A_i \cos(\omega_i n + \phi_i) + u(n))(\sum_{j=1}^{M} A_j \cos(\omega_j(n+k) + \phi_j) + u(n+k)) \\
&= \textstyle\sum_{n=0}^{N-k-1}(\sum_{i=1}^{M} A_i^2 \cos(\omega_i n + \phi_i)\cos(\omega_i(n+k) + \phi_i) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1}\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M} A_i A_j \cos(\omega_i n + \phi_i)\cos(\omega_j(n+k) + \phi_j) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1}\sum_{j=1}^{M} A_j u(n)\cos(\omega_j(n+k) + \phi_j) + \sum_{n=0}^{N-k-1}\sum_{i=1}^{M} A_i u(n+k)\cos(\omega_i n + \phi_i) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1} u(n)u(n+k) \\
&= \textstyle\sum_{n=0}^{N-k-1}(\sum_{i=1}^{M} \frac{A_i^2}{2}(\cos(\omega_i k) + \cos(2\omega_i n + \omega_i k + 2\phi_i))) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1}\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M} \frac{A_i A_j}{2}(\cos((\omega_i - \omega_j)n - \omega_j k + \phi_i - \phi_j) + \cos((\omega_i + \omega_j)n + \omega_j(k) + \phi_i + \phi_j)) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1}\sum_{j=1}^{M} A_j u(n)\cos(\omega_j(n+k) + \phi_j) + \sum_{n=0}^{N-k-1}\sum_{i=1}^{M} A_i u(n+k)\cos(\omega_i n + \phi_i) \\
&\quad + \textstyle\sum_{n=0}^{N-k-1} u(n)u(n+k) \\
&= \underbrace{(N-k)\sum_{i=1}^{M}\frac{A_i^2}{2}\cos(\omega_i k)}_{SIG} + \sum_{i=1}^{M}\underbrace{\frac{A_i^2}{2}\sum_{n=0}^{N-k-1}\cos(2\omega_i n + \omega_i k + 2\phi_i))}_{A} \\
&\quad + \textstyle\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}\underbrace{\frac{A_i A_j}{2}\sum_{n=0}^{N-k-1}\cos((\omega_i - \omega_j)n - \omega_j k + \phi_i - \phi_j)}_{B} \\
&\quad + \textstyle\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}\underbrace{\frac{A_i A_j}{2}\sum_{n=0}^{N-k-1}\cos((\omega_i + \omega_j)n + \omega_j(k) + \phi_i + \phi_j)}_{C} \\
&\quad + \textstyle\sum_{j=1}^{M}\underbrace{\sum_{n=0}^{N-k-1} A_j u(n)\cos(\omega_j(n+k) + \phi_j)}_{D} + \sum_{i=1}^{M}\underbrace{\sum_{n=0}^{N-k-1} A_i u(n+k)\cos(\omega_i n + \phi_i)}_{E} \\
&\quad + \underbrace{\sum_{n=0}^{N-k-1} u(n)u(n+k)}_{F}
\end{aligned}
\tag{A.2}
$$

Let us consider the under-braced terms A, B, C. They are the sums of the samples of a cosine wave at frequencies, $2\omega_i$, $\omega_i - \omega_j$, $\omega_i + \omega_j$ respectively. If (N-k) is much greater than the $\frac{2\pi}{\omega_i}$ and $\frac{2\pi}{\omega_i - \omega_j}$ for all frequency pairs (i, j), then the sums A, B, C will contain several periods of their corresponding cosine waves. Sum over 'Q' periods of the samples of a cosine wave at any nonzero frequency is zero. This can be seen by the following integral which approximates the sum

of the samples in A, B, C,

$$\int_{t=0}^{t=Q2\pi/\omega} A\cos(\omega t+\phi)dt = 0 \tag{A.3}$$

This happens as the negative and positive swings of the cosine cancel each other. Therefore A, B, C are approximately zero. Let $(N-k) = Q \times T + \Delta$, where Q is an integer and T is the period of a certain sinusoid at frequency ω and Δ is the left over part as $N-k$ is not an exact multiple of T. Hence $\Delta < T$. Then we have,

$$\begin{aligned}\sum_{n=0}^{n=N-k} A\cos(\omega n+\phi) &= \sum_{n=0}^{n=QT} A\cos(\omega n+\phi) + \sum_{n=QT+1}^{n=N-k} A\cos(\omega n+\phi)\\ &= \sum_{n=QT+1}^{n=N-k} A\cos(\omega n+\phi)\\ &< A\times\Delta \ll A\times(N-k)\end{aligned} \tag{A.4}$$

This proves the A, B, C can safely be ignored in comparison to SIG term which is proportional to (N-k). Moreover D, E are also approximately zero as the noise $u(n)$ is assumed uncorrelated with signal $s(n)$. The term $F = N\sigma^2\delta(k)$ as the noise is assumed to be white. Hence ignoring A, B, C, D, E, we get

$$\begin{aligned}r(k) &= \sum_{n=0}^{N-k-1} x(n)x(n+k)\\ &\simeq \sum_{i=1}^{M}(N-k)A_i^2 cos(2\pi f_i k) + N\sigma^2\delta(k)\\ &\simeq \sum_{i=1}^{M}(N\exp(-\alpha k))A_i^2 cos(2\pi f_i k) + N\sigma^2\delta(k)\end{aligned} \tag{A.5}$$

where $\alpha = 1/N$ and hence $\alpha \ll 1$.

A.2 Solving the least squares matrix equation

In the section, we will analytically solve the LeSF equation for the form of the autocorrelation coefficients given above in (A.5). The $(L \times L)$ matrix LeSF equation is reproduced below

$$\begin{pmatrix} r(0) & r(1) & r(2) & \cdots & r(L-1)\\ r(1) & r(0) & r(1) & \cdots & r(L-2)\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ r(L-1) & r(L-2) & \cdots & \cdots & r(0)\end{pmatrix} \times \begin{pmatrix} w(0)\\ w(1)\\ \cdots\\ w(L-1)\end{pmatrix} = \begin{pmatrix} r(P)\\ r(P+1)\\ \cdots\\ r(P+L-1)\end{pmatrix} \tag{A.6}$$

where the $w(0)$, $w(1), \cdots, w(L-1)$ are the LeSF filter tap weights and the autocorrelation coefficients $r(k)$ are given by (A.5). In [9], it has been shown that the functional form of filter tap weights in (A.6) for the form of the $r(k)$ as in (A.5), is given by,

$$w(k) = \sum_{i=1}^{M-1} \left(B_i \exp(-\beta_i k) + C_i \exp(\beta_i k)\right) \cos(\omega_i (k+P)) \tag{A.7}$$

where P is the bulk delay. In (A.7), the quantities C_i, B_i, β_i are unknown. Our objective is to solve (A.6) for the unknown quantities in the filter tap weights w in closed form. Towards this end, let's consider the $(p+1)^{th}$ equation in the system of the equations (A.6) which is reproduced below,

$$\sum_{k=0}^{k=p-1} r(p-k)w(k) + r(0)w(p) + \sum_{k=p+1}^{L-1} r(k-p)w(k) = r(P+p) \tag{A.8}$$

Next, we substitute the functional forms of $r(k)$ and $w(k)$ in (A.8). We collect the terms that correspond to the i^{th} sinusoid together and call them as "self-terms" while the terms that have contribution from the i^{th} and the j^{th} sinusoid are called "cross terms". As, we will show that some of these cross terms can be ignored in comparison to the "self-terms", thus facilitating analytical solution. The "self-terms" due to the i^{th} sinusoid, on the right hand side of (A.8) are,

$$\sum_{k=0}^{p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\cos(\omega_i(k+P))\left[A_i^2\exp(-\alpha(p-k))\cos(\omega_i(p-k))\right]$$
$$+\left[B_i\exp(-\beta_i p)+C_i\exp(\beta_i p)\right]\cos(\omega_i(p+P))\left[\sum_{i=1}^{M}A_i^2+\sigma^2\right]$$

$$\sum_{k=p+1}^{L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\cos(\omega_i(k+P))\left[A_i^2\exp(-\alpha(k-p))\cos(\omega_i(k-p))\right]$$
$$=\underbrace{\frac{1}{2}A_i^2\cos(\omega_i(p+P))\sum_{k=0}^{k=p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(p-k))}_{stationary}$$
$$+\underbrace{\frac{1}{2}A_i^2\sum_{k=0}^{k=p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]exp(-\alpha(p-k))\cos(2\omega_i k-\omega_i(p-P))}_{non-stationary}$$
$$+\underbrace{\left[B_i\exp(-\beta_i p)+C_i\exp(\beta_i p)\right]\cos(\omega_i(p+P))(\sum_{i=1}^{M}A_i^2+\sigma^2)}_{stationary}$$
$$\underbrace{\frac{1}{2}A_i^2\cos(\omega_i(p+P))\sum_{k=p+1}^{k=L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(k-p))}_{stationary}$$
$$+\underbrace{\frac{1}{2}A_i^2\sum_{k=p+1}^{k=L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]exp(-\alpha(k-p))\cos(2\omega_i k-\omega_i(p-P))}_{non-stationary} \tag{A.9}$$

The "non-stationary" terms approximately sum up to zeros due to the self-canceling positive and negative swings of the sinusoid at frequency $(2\omega_i)$ and hence can be ignored in comparison to the "stationary terms". Similarly in the $(p+1)^{th}$ equation, there are "cross-terms" that get contribution from the i^{th} and the j^{th} sinusoid. These terms are given below.

$$
\begin{aligned}
&\textstyle\sum_{k=0}^{p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\cos(\omega_i(k+P))\left[A_j^2\exp(-\alpha(p-k))\cos(\omega_j(p-k))\right]\\
&+\textstyle\sum_{k=p+1}^{L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\cos(\omega_i(k+P))\left[A_j^2\exp(-\alpha(k-p))\cos(\omega_j(k-p))\right]\\
&=\underbrace{\frac{1}{2}A_j^2\sum_{k=0}^{k=p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(p-k))\left[\cos((\omega_i-\omega_j)(k)+\omega_i P+\omega_j p)\right]}_{non-stationary}\\
&+\underbrace{\frac{1}{2}A_j^2\sum_{k=0}^{k=p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(p-k))\left[\cos((\omega_i+\omega_j)(k)+\omega_i P-\omega_j p)\right]}_{non-stationary}\\
&+\underbrace{\frac{1}{2}A_j^2\sum_{k=p+1}^{k=L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(k-p))\left[\cos((\omega_i-\omega_j)k+\omega_i P+\omega_j p)\right]}_{non-stationary}\\
&+\underbrace{\frac{1}{2}A_j^2\sum_{k=p+1}^{k=L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(k-p))\left[\cos((\omega_i+\omega_j)(k)+\omega_i P-\omega_j p)\right]}_{non-stationary}
\end{aligned}
\tag{A.10}
$$

If $(\omega_i-\omega_j)\gg\frac{2\pi}{L}$ and $(\omega_i+\omega_j)\neq 2\pi$, then these "non-stationary" terms approximately sum up to zero too. Therefore, all the cross-terms noted above can be safely neglected on the right hand side of the equation (A.8). Therefore neglecting all the non-stationary terms in (A.9), (A.10), we get,

$$
\begin{aligned}
&\underbrace{\frac{1}{2}A_i^2\cos(\omega_i(p+P))\sum_{k=0}^{k=p-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(p-k))}_{stationary}\\
&+\underbrace{\left[B_i\exp(-\beta_i p)+C_i\exp(\beta_i p)\right]\cos(\omega_i(p+P))(\sum_{i=1}^{M}A_i^2+\sigma^2)}_{stationary}\\
&+\underbrace{\frac{1}{2}A_i^2\cos(\omega_i(p+P))\sum_{k=p+1}^{k=L-1}\left[B_i\exp(-\beta_i k)+C_i\exp(\beta_i k)\right]\exp(-\alpha(k-p))}_{stationary}\\
&=\textstyle\sum_{i=1}^{M}\exp(-\alpha(p+P))A_i^2 cos(\omega_i(p+P))
\end{aligned}
\tag{A.11}
$$

Next, we collect all the terms in (A.11) with the coefficients $\exp(-\beta_i p)$,$\exp(+\beta_i p)$ for each of the i^{th} sinusoid and set them to zero as there are no terms on the right hand side of (A.11) with these coefficients. Consider the terms with the coefficient $\exp(-\beta_i p)$, which are given below, and is set to zero as explained

above.

$$\sum_{i=1}^{M}\left[-\frac{A_i^2}{2}\frac{\cos(\omega_i(p+P))B_i\exp(-\beta_i p)}{(1-\exp(-(\beta_i-\alpha)))}+B_i\exp(-\beta_i p)\cos(\omega_i(p+P))\left(\sum_{i=1}^{M}A_i^2+\sigma^2\right)\right]$$

$$+\sum_{i=1}^{M}\left[\frac{A_i^2}{2}\frac{B_i\exp(-\beta_i p)\exp(-(\alpha+\beta_i))\cos(\omega_i(p+P))}{(1-\exp(-(\alpha+\beta_i)))}\right]$$

$$=0 \tag{A.12}$$

Therefore for each "i", we get the relationship,

$$\cosh\beta_i=\cosh\alpha+\frac{\rho_i}{2\gamma_i+\rho_i+2}\sinh\alpha \tag{A.13}$$

where, ρ_i denotes the "partial" SNR of the sinusoid at frequency ω_i i.e $\rho_i = A_i^2/\sigma^2$ and the complementary signal SNR is denoted as $\gamma_i = (\sum_{m=1,m\neq i}^{M} A_i^2)/\sigma^2$. The coefficients of the terms $exp(-\beta_i p)$, $exp(+\beta_i p)$ are the same for each of the L equations and setting them to zero leads to just one equation which relates β_i to α, ρ_i and γ_i. Next step is to solve for B_i and C_i. Towards this end, we equate the coefficient of $\exp(-\alpha p)$ on both the left and right hand sides of (A.11). This leads to,

$$\begin{aligned}\frac{A_i^2\cos(\omega_i(p+P))B_i\exp(-\alpha p)}{1-\exp(-(\alpha-\beta_i))}+\frac{A_i^2\cos(\omega_i(p+P))C_i\exp(-\alpha p)}{1-\exp(-(\alpha+\beta_i))}\\ =A_i^2\exp(-\alpha p)\exp(-\alpha P)\cos(\omega_i(p+P))\end{aligned} \tag{A.14}$$

Similarly we set the coefficient of $\exp(+\alpha p)$ to zero as there is no term with this coefficient on the right hand side of (A.11). This leads to,

$$\begin{aligned}&\frac{A_i^2\cos(\omega_i(p+P))B_i\exp(-(\alpha+\beta_i))\exp(-\alpha L+\alpha p+\alpha-\beta_i L+\beta_i)}{1-\exp(-(\alpha+\beta_i))}\\ &+\frac{A_i^2\cos(\omega_i(p+P))C_i\exp(-\alpha+\beta_i)\exp(-\alpha L+\alpha p+\alpha+\beta_i L-\beta_i)}{1-\exp(-\alpha+\beta_i)}=0\end{aligned} \tag{A.15}$$

There are two unknowns B_i and C_i in (A.14) and (A.15). Solving them simultaneously gives,

$$B_i = \frac{2e^{-\beta_i}e^{-\alpha P}(\alpha+\beta_i)^2(\beta_i-\alpha)}{((\alpha+\beta_i)^2 - e^{-2\beta_i L}(\beta_i-\alpha)^2)}$$
$$C_i = \frac{2e^{-\beta_i(2L+1)+1}e^{-\alpha P}(\alpha+\beta_i)(\beta_i-\alpha)^2}{((\alpha+\beta_i)^2 - e^{-2\beta_i L}(\beta_i-\alpha)^2)} \quad \text{(A.16)}$$

This concludes the analytic solution of the LeSF equation.

References

[1] H. K Kim, and R. C Rose, "Cepstrum domain acoustic feature compensation based on decomposition of speech and noise for ASR in noisy environments, " IEEE Tran. on SAP, vol. 11, No. 5, September 2003.

[2] R. Martin, " Noise Power Spectral Density Estimation based on optimal smoothing and minimum statistics, " IEEE Trans. on SAP, Vol. 9, No. 5, July 2001.

[3] Y. Ephraim and D. Malah, " Speech enhancement using a minimum mean square error short-time spectral magnitude estimator, " IEEE Trans. on ASSP, vol. ASSP-32, pp.1109-1121, 1984.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log spectral amplitude estimator, " IEEE Tran. on ASSP, vol. ASSP-33, No. 2, April 1985.

[5] G. Lathoud, M. M. Doss and B. Mesot, " A spectrogram model for enhanced source localization and noise robust ASR, " to appear in the Proc. of Eurospeech 2005, Lisbon, Portugal.

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

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

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

[9] 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.

[10] 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.

[11] 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.

[12] R. J. McAulay and M. L. Malpass, " Speech Enhancement Using a Soft-Decision Noise Suppression Filter, " In IEEE Trans. on ASSP, Vol. ASSP-28, No. 2, April 1980.

[13] 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.

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

[15] M. Sondhi and D. Berkley, " Silencing echoes on the telephone network," Proc. of IEEE, vol.68, pp948-963, Aug. 1980.

[16] A Gersho, " Adaptive equalization of highly dispersive channels for data transmission, " Bell Syst. Tech. J., vol.48, pp.55-70, Jan. 1969.

[17] E. Satorius and S. T. Alexander, " Channel equalization using adaptive lattice algorithms, " IEEE Trans. Commun. vol. 27, pp.899-905, June 1979.

[18] E. Satorius and J. Pack, " Application of least squares lattice algorithms for adaptive equalization, " IEEE Trans. on Commun. vol. COM-29, pp.136-142, Feb. 1981.

[19] N. Bershad, P. Feintuch, F. Reed and B. Fisher, " Tracking characteristics of the LMS adaptive line-enhancer -Response to a linear chirp signal in noise, " IEEE Trans. on ASSP, vol. ASSP-28, pp504-517, Oct. 1980

[20] L. J. Griffiths, " A simple adaptive algorithm for real time processsing in antenna arrays, " Proc. of IEEE, vol. 57, pp.1696-1704, Oct. 1969.

[21] O.L. Frost, " An algorithm for linearly constrained adaptive array processing , " Proc. of IEEE, vol. 60, pp.926-935, Aug. 1972.

[22] R. Compton, " Pointing accuracy and dynamic range in a steered beam antenna array, " IEEE. Trans. on Aerosp. Electron. Syst., vol. AES-16, pp.280-287, may 1980.

[23] C. Gibson, and S. Haykin, " Learning characteristics of adaptive lattice filtering algorithms, " IEEE Trans. ASSP, vol. ASSP-28, pp.681-692, Dec. 1980.

[24] L. Rabiner, R. Crochiere and J. Allen , " FIR system modeling and identification in the presence of noise and band-limited inputs, " IEEE Trans. on ASSP, vol. ASSP-26, pp.319-333, Aug 1978.

[25] L. Marple, " Efficient least squares FIR system identification, " IEEE Trans. on ASSP, vol.ASSP-29, pp.62-73, Feb. 1981.

[26] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences, " IEEE Trans. on ASSP, Vol. ASSP-28, No. 4, August 1980.

[27] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK Book, Cambridge University, 1995.

[28] R. A. Cole, M. Fanty, and T. Lander, "Telephone speech corpus at CSLU," Proc. of ICSLP, Yokohama, Japan, 1994.

[29] 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.