KALMAN FILTERING FOR LOW DISTORTION SPEECH ENHANCEMENT IN MOBILE COMMUNICATION‡

Patrik Sörqvist, Peter Händel* and Björn Ottersten†*

*Audio and Visual Technology Research, Ericsson Radio Systems AB
S-164 80 Stockholm, Sweden
†Department of Signals, Sensors and Systems, Royal Institute of Technology
S-100 44 Stockholm, Sweden

ABSTRACT

This paper presents a model-based approach for noise suppression of speech contaminated by additive noise. A Kalman filter based speech enhancement system is presented and its performance is investigated in detail. It is shown that with a novel speech parameter estimation algorithm, it is possible to achieve 10dB noise suppression with a high total audible quality.

1. INTRODUCTION

A common signal processing problem is the enhancement of a signal from its noisy measurement. An example with practical relevance is speech enhancement in hands-free mobile telephony, where the speech may be severely contaminated by colored non stationary noise, such as the noise inside the compartment of a running car. The source is mainly the engine and the coupe fan at low car speeds, and the road and the wind at higher speeds.

The motivation for studying a Kalman filter based noise suppression system is that it can handle colored noise and has a reasonable numerical complexity. Additionally it is well suited to face the speech quality requirements in mobile telephony, that is low speech distortion, low distortion of the background noise and a low inherent time delay. One can note that a Kalman filter based enhancement system is implemented in the pacific digital cellular (PDC) half rate speech coding standards, [1].

A key issue in Kalman filtering is that the filtering algorithm relies on a set of parameters, that for this particular application are unknown and has to be replaced be estimated quantities. The choice of estimation method is of outmost importance, since the speech autoregressive (AR) parameters are estimated from noisy speech data, and the data is a non stationary process.

In order to obtain a Kalman filter output with high audible quality, it is important that the models are accurately estimated. For this application, the background noise may be considered long time stationary and consequently its parameters may be accurately estimated during speech pauses. Speech may be viewed as a short time stationary process, that is stationarity for 10-40ms (80-320 samples at 8kHz sampling rate). Thus, an instantaneous model of the speech has to be obtained from a short segment of noisy measurements.

The speech enhancement system studied consists of three main parts, that is a voice activity detector (VAD), an estimation module and a Kalman filter. The VAD and the estimation module both utilize data blocks of N samples for detection of speech activity and AR parameter estimation. Accordingly, the model parameters of the Kalman filter are updated with the same rate, that is new parameter estimates are down-loaded to the Kalman filter every 32ms (8 kHz sampling rate and $N = 256$, which is considered through the paper). This implies that an inherent time delay of one frame is introduced. A VAD based on the PDC half rate speech coding standards [1] is used. For each data frame, it provides a boolean flag indicatingthe presence or absence of speech.

The paper is organized as follows. In Section 2 the parametric models are defined and the Kalman filter reviewed. Section 3, treats estimation of the background noise parameters and methods for estimation of the speech parameters. An improved speech parameter estimation is derived in Section 4. The performance of the complete system is investigated in Section 5 and, the conclusions are given in Section 6.

2. DATA MODEL AND KALMAN FILTER

After analog preprocessing and AD conversion, the noisy speech is modeled as a sum of two AR processes, that is

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

where $x(n)$ denotes the measured signal, $s(n)$ the speech, and $v(n)$ denotes the colored noise. Further

$$\begin{aligned} s(n) &= \sum_{i=1}^{p} a_i\, s(n-i) + w(n) \\ v(n) &= \sum_{i=1}^{q} b_i\, v(n-i) + \eta(n) \end{aligned} \tag{2}$$

In (2), the design variables p and q denote the model order for $s(n)$ and $v(n)$, respectively. The noises $w(n)$ and $\eta(n)$

‡An audio file supporting this work is available in the CD rom proceedings. The file contains male and female speech contaminated by noise recorded in a running car, and the output from the studied noise reduction system.

*†P. Händel is currently on leave at Signal Processing Laboratory, Tampere University of Technology, Finland. B. Ottersten is currently on leave at ArrayComm Inc., San Jose, CA.

*†The authors email addresses are: patrik.sorqvist@era-t .ericsson.se, ph@cs.tut.fi, and otterste@s3.kth.se.

are assumed white zero mean with variances σ_w^2 and σ_η^2, respectively. The parameters in (2) are all unknown. In general, the noise in a car compartment is not stationary, but it is appropriate to consider it as a stationary process for periods of 1-2 seconds. The statistics of speech changes even faster and the model cannot be considered stationary for more than fractions of a second. Due to the short-time stationarity of $s(n)$, $\{a_i\}$ may be assumed time invariant for 10-40ms. The noise parameters $\{b_i\}$ may typically be assumed constant for 1-2s.

Note that, in speech pauses (1) is reduced to $x(n) = v(n)$. Thus, in speech pauses the estimation of $\{b_i\}$ is trivial.

The problem of interest is to extract the speech from the noisy measurement. The purpose of modeling the data as AR processes is to use a model-based method to enhance the degraded speech. The Kalman *filter* solves the problem of estimating the signal $s(n)$ from observations of $\{x(k)\}_{k=0}^{n}$, [2]. The Kalman filter relies on a state space formulation. Let

$$\theta(n) = (s(n-p+1)\cdots s(n), v(n-q+1)\cdots v(n))^T \quad (3)$$

then, the state space representation of (1)-(2) is

$$\begin{aligned} \theta(n+1) &= \mathbf{F}\theta(n) + \mathbf{G}\mathbf{z}(n) \\ x(n) &= \mathbf{h}^T\theta(n) \\ s(n) &= \mathbf{h}_2^T\theta(n) \end{aligned} \quad (4)$$

In (4), $\mathbf{F}$ is a $(r \times r)$ matrix with $r = p + q$, $\mathbf{G}$ is a $(r \times 2)$ matrix, $\mathbf{h}$ and $\mathbf{h}_2$ are column vectors of length r. Further, $\mathbf{z}(n) = (\ w(n)\ \eta(n)\ )^T$, and $\mathbf{Q} = \mathbf{G}E\left[\mathbf{z}(n)\,\mathbf{z}(n)^T\right]\mathbf{G}^T = \mathbf{G}\mathbf{R}_1\mathbf{G}^T$. In $\mathbf{h}$ just two elements are non zero, that is $\mathbf{h}(p) = 1$ and $\mathbf{h}(r) = 1$. The vector $\mathbf{h}_2$ contains zeros except $\mathbf{h}_2(p) = 1$. Explicit expressions for $\mathbf{F}$, $\mathbf{G}$, and $\mathbf{R}_1$ are given below.

$$\mathbf{F} = \left(\begin{array}{c|c} \overbrace{\begin{array}{ccc} 0 & 1 & \\ & \ddots & \\ & & 1 \\ a_p & \cdots & a_1 \end{array}}^{p} & \begin{array}{c} \\ \mathbf{0}_{p,q} \\ \\ \end{array} \\ \hline \mathbf{0}_{q,p} & \begin{array}{cc} 0 & 1 \\ & \ddots \\ & 1 \\ \underbrace{b_q \quad \cdots \quad b_1}_{q} & \end{array} \end{array}\right) \quad (5)$$

$$\mathbf{G} = \left(\begin{array}{cccc|cccc} \overbrace{0}^{} & \cdots & 0 & 1 & \overbrace{0}^{} & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1 \end{array}\right)^T \quad (6)$$

$$\mathbf{R}_1 = \begin{pmatrix} \sigma_w^2 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix} \quad (7)$$

Equation (6)-(7) implies that $\mathbf{Q}$ is a sparse matrix with only two nonzero elements, that is $\mathbf{Q}(p,q) = \sigma_w^2$ and $\mathbf{Q}(r,r) = \sigma_\eta^2$.

The Kalman filter estimate is given by [2]

$$\begin{aligned} \tilde{\mathbf{P}}(n) &= \mathbf{F}\mathbf{P}(n-1)\mathbf{F}^T + \mathbf{Q} \\ \mathbf{k}(n) &= \frac{\tilde{\mathbf{P}}(n)\mathbf{h}}{\mathbf{h}^T\tilde{\mathbf{P}}(n)\mathbf{h}} \\ \hat{\theta}(n) &= \mathbf{F}\hat{\theta}(n-1) + \mathbf{k}(n)\left(x(n) - \mathbf{h}^T\mathbf{F}\hat{\theta}(n-1)\right) \quad (8) \\ \mathbf{P}(n) &= [\mathbf{I}_{p+q} - \mathbf{k}(n)\mathbf{h}^T]\tilde{\mathbf{P}}(n) \\ \hat{s}(n) &= \mathbf{h}_2^T\hat{\theta}(n) \end{aligned}$$

In (8), $\tilde{\mathbf{P}}(n)$ is the covariance matrix for the prediction error, $\mathbf{P}(n)$ is the covariance matrix for the estimation error, $\mathbf{k}(n)$ is the Kalman gain which controls the step-size, and $\hat{\theta}(n)$ is the estimated state space vector. Since updated values of ($\{a_i\}$, $\{b_i\}$, σ_w^2, $\sigma_e ta^2$) are down-loaded every 32ms, the Kalman filter is re-initialized according to the final values from the previous frame.

The Kalman filter provides the best linear unbiased estimate of $\theta(n)$. Unfortunately, this is not the same as maximizing the audible quality. Problems may arise from cases when the Kalman filter is to sharp, that is when the noise and parts of the speech is removed. This can be handled by reducing the distance from the poles to origin of coordinates in the estimated model. The noise variance is also a parameter that acts on the filters sharpness. If σ_η^2 is scaled by δ, $\delta < 1$, the effect of the filter is reduced.

The computational complexity of the Kalman filter may be significantly reduced because of the sparse structure of the matrices $\mathbf{F}$, $\mathbf{Q}$ and $\mathbf{h}$, see [3] where a linear transformation of the state vector, $\theta(n)$, is considered. The numerical benefit of such an optimization is that the total complexity is reduced from $r^3 + 10r^2$ to $10r^2$ additions and multiplications per processed sample.

3. PARAMETER ESTIMATION

The problems concerning parameter estimation have to be carefully examined. Poor estimates of the speech or noise AR parameters result in model errors in the Kalman filter. This implies that the enhancement of the speech from the noisy measurement yields poor audible quality.

Estimation of $\{b_i\}$ and σ_n^2 in (2) is performed during speech pauses. Since a fixed 32ms block length is utilized, an averaging of the autocorrelation function (ACF) is used. The ACF for the noise is estimated as $\hat{\gamma}(k) = 1/N\sum_{i=1}^{N-k} v(k+i)v(i)$ and updated at frame level according to $\hat{\gamma}_\ell(k) = \rho\hat{\gamma}_{\ell-1}(k) + (1-\rho)\hat{\gamma}(k)$ where $\hat{\gamma}_\ell(k)$ denotes the running average in frame ℓ based on $\hat{\gamma}(k)$. The Levinson-Durbin algorithm is used for AR parameter calculation, [4].

The estimation of the noise parameters during speech pauses requires pauses with certain intervals. If the VAD detects speech activity for longer time than the stationarity holds for the noise, the model will loose in accuracy. However, one may use the Kalman filters estimates of the noise to update the noise model during speech activity.

There are several methods to estimate the speech parameters from the degraded speech, that is the problem of estimating $\{a_i\}$ and σ_w^2 in (2) from (1) where the properties of the noise are partially known. Some methods are outlined below.

The most straightforward method is to estimate the parameters directly from the degraded speech, that is the

speech is approximated with the received noisy speech. This approach works well for data frames with high SNR, but for moderate and low SNR the performance is severely degraded since the approximation, that $x(n) \approx s(n)$, is no longer valid. This results in poor estimates and consequently the Kalman filter will not work as intended.

An alternative is iterative direct estimation, proposed in [3]. In a first step the estimated AR speech parameters obtained from the noisy data are used in the Kalman filter. The output from the Kalman filter, $\hat{s}(n)$, is used to enhance the parameter estimates, that is, a new set of speech AR parameters is estimated from $\hat{s}(n)$ and used in a second Kalman filter. This method has the same kind of problem as direct estimation, for low SNRs the output from the Kalman filter is often distorted. The iterative method does not guarantee an improved estimate in terms of audible quality.

Neither of the previous methods make use of the fact that the noise parameters are estimated during speech pauses. Such an approach is outlined below. Rewrite (1) and (2) as

$$x(n) = \frac{B(q^{-1})w(n) + A(q^{-1})\eta(n)}{A(q^{-1})B(q^{-1})} \tag{9}$$

The parameters $\{b_i\}$ and σ_η^2, are unknown, but due to the long-time stationarity of the background noise, accurate estimates $\hat{B}(q^{-1})$ and $\hat{\sigma}_\eta^2$ are available. Then, filtering $x(n)$ with $\hat{B}(q^{-1})$ gives

$$y(n) = \hat{B}(q^{-1})x(n) \approx \frac{C(q^{-1})}{A(q^{-1})}e(n) \tag{10}$$

where $C(q^{-1})$ is a polynomial of degree $\max(p,q)$ and $e(n)$ is a white noise. From (10) it is evident that $\{a_i\}$ may be estimated as the AR part of an ARMA process. Efficient algorithms may be found, for example, in [4]. Our experience is that in practice it is hard to get accurate estimates of $\{a_i\}$ due to the finite block length N.

4. IMPROVED SPEECH PARAMETER ESTIMATION

In this section, the spectral features of the speech will be used to obtain accurate estimates of the speech parameters.

The polynomial $A(q^{-1})$ determines the characteristic of the speech, therefore it is natural to consider the power spectral density (PSD) corresponding to $A(q^{-1})$ as an important feature of the speech. The proposed method estimates the PSD for the speech. The parameters $A(q^{-1})$ and σ_w^2 are then implicitly calculated from an estimated speech PSD.

The PSDs are evaluated in M equidistant points. The choice of M can be done independently of the size of the speech analyses frame, N. Clearly M has to be greater than p. From (1)-(2) it follows that

$$\Phi_x(k) = \Phi_s(k) + \Phi_v(k) \qquad k = 1, \ldots, M \tag{11}$$

where $\Phi(k)$ is a short notation for $\Phi(2\pi k/M)$, and where $\Phi_s(k) = \sigma_w^2 / \mid A(e^{j\omega}) \mid^2, \omega = 2\pi k/M$. Since $\Phi_v(k)$ can be estimated during speech pauses, it is natural to estimate the speech PSD as

$$\hat{\Phi}_s(k) = \hat{\Phi}_x(k) - \delta(k)\hat{\Phi}_v(k) \quad k = 1, \ldots, M \tag{12}$$

where $\hat{\Phi}_x(k)$ is an estimate based on data in the present frame. In (12), $\delta(k)$ is a possibly frequency dependent design variable. The variable, $\delta(k)$ may be used in order to optimize the performance, for example

$$\delta_{opt}(k) = \arg\min E\left[\hat{\Phi}_s(k) - \Phi_s(k)\right]^2 \tag{13}$$

where $\hat{\Phi}_s(k)$ is given in (12). In a first approximation, see [5] for details

$$\mathrm{Var}(\hat{\Phi}_v(k)) \approx \frac{1-\rho}{2}\frac{2q}{N}\Phi_v^2(k) \tag{14}$$

$$\mathrm{Var}(\hat{\Phi}_x(k)) \approx \frac{2p}{N}\Phi_x^2(k) \tag{15}$$

where N is the frame length, p and q the model orders in (2) and $2/(1-\rho)$ roughly determines the number of frames used to estimate the noise parameters. Let $\hat{\Phi}_v(k) = \Phi_v(k) + \Delta_v(k)$ and $\hat{\Phi}_x(k) = \Phi_x(k) + \Delta_x(k)$ where $\Delta_v(k)$ and $\Delta_x(k)$ are stochastic quantities that have zero means and variances given by (14) and (15), respectively. Further $\Delta_v(k)$ and $\Delta_x(k)$ are approximately uncorrelated [5], that follows from the fact that $\hat{\Phi}_v(k)$ and $\hat{\Phi}_x(k)$ are calculated from different sets of data, and that the input data is non-stationary. Further, let $\tilde{\Phi}_s(k)$ denote the PSD estimation error, that is $\tilde{\Phi}_s(k) = \hat{\Phi}_s(k) - \Phi_s(k)$. Then, a straight forward calculation gives

$$\tilde{\Phi}_s(k) = \Phi_v(k)(1-\delta(k)) + \Delta_x(k) - \delta(k)\Delta_v(k) \tag{16}$$

Note that the first part of (16) is deterministic. Thus the mean square error (MSE) is

$$E\left[\tilde{\Phi}_s^2(k)\right] \approx (1-\delta(k))^2\Phi_v^2(k) + \frac{2p}{N}\Phi_x^2(k) + \delta^2(k)\frac{1-\rho}{N}q\Phi_v^2(k) \tag{17}$$

Now minimize $E\left[\tilde{\Phi}_s^2(k)\right]$ with respect to $\delta(k)$. From (17) it follows that

$$\delta_{opt} = \frac{N}{N + (1-\rho)q} \tag{18}$$

Note that δ_{opt} is frequency independent and $\delta_{opt} < 1$.

Since (12) contains estimated quantities, the PSD subtraction may result in negative values for some k. To prevent negative values it is appropriate to set a fixed minimum level, or to use a fixed maximum reduction.

The parameters $\{a_i\}$ and σ_w^2, with corresponding PSD as close to $\hat{\Phi}_s(\omega)$, in (12), as possible, are sought. It is not possible to calculate $\{a_i\}$ or σ_w^2 directly, therefore an iterative algorithm is unavoidable. A recursive prediction error method (RPEM) is used to fit the given situation, [4]. In initializing the RPEM algorithm, the parameters from the degraded speech are used. The number of iterations is highly dependent on the initial values. For frames with low SNR, the lower frequencies often have very poor initial estimates since background noise in mobile communications often has lowpass properties. This fact can be used to motivate interrupt conditions such as low error and small gain. The stability of the algorithm can be checked by ensuring that the roots of the AR polynomial is located inside the unit circle. If divergence is detected the output parameters is set to the initial values.

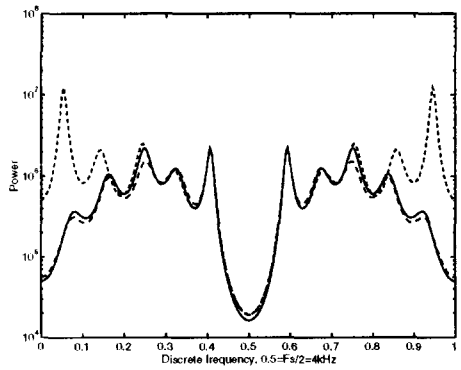

Figure 1: The RPEM algorithm performance for $p = 10$ and $\delta = 1$. The solid line represents the speech PSD, the dashed line the estimated speech PSD and the dotted line the PSD when no post processing is applied.

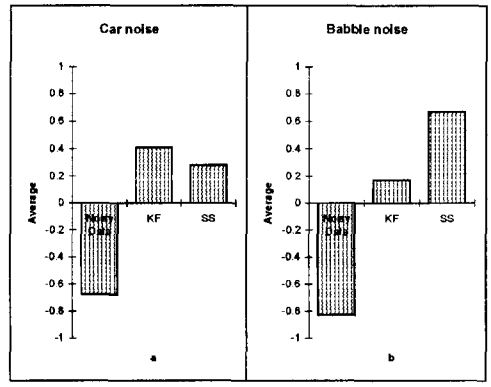

Figure 2: Method preferences for different kinds of background noise. a) car noise SNR=10dB, b) babble noise.

5. PERFORMANCE

This section treats the experimental results of the speech enhancement system. Genuine data is used, that is car noise or "cocktail party" noise is added to pre-recorded speech.

The performance of the method in section 4 is illustrated in Figure 1, where the solid line represents the speech PSD (calculated from the clean speech signal using $p = 10$), the dashed line the estimated speech PSD (using $p = 10$ and $\delta = 1$. The dotted line indicates the degraded speech PSD, which is the initial values for the RPEM algorithm. It is clear that a significant improvement is achieved.

Our experiments indicate that the improved estimates correspond to a direct parameter estimation of noisy speech with 12dB higher SNR, in terms of MSE, for SNR$<$ 10dB.

The performance of the complete system has been studied in detail. The conclusion is to use direct estimation of the speech AR parameters when the frame SNR is high, and improved estimation, such as the method above, when the frame SNR is low. This requires a criterion based on the frame SNR, that decides whether to use direct estimation or one of the more advanced methods previously discussed.

A listening test has been carried out in order to evaluate the performance of the speech enhancement system. The 12 experienced test persons listened to three different versions (Kalman filtered, enhanced by a low distortion spectral subtraction (SS) method [5] and the original noisy speech) of the same sentences, but two at a time. The comparisons were carried out in random order. Each time a sentence was chosen, one point was awarded to the type of file that was chosen (KF, SS or noisy speech) and the other file type lost one point. The average results for the different subtests are displayed in Figure 2. An average of 1 indicates that the file was preferred every time, 0 indicates preferred every second time and -1 that the file was rejected every time.

In the first subtest the file contained two female speakers and two male speakers contaminated by car noise. The results displayed in Figure 2a, show that a significant quality improvement is achieved when noise reduction is applied.

In the second subtest the background noise was babble noise, that is a non-stationary noise such as "cocktail party" noise. The results in Figure 2b, show that SS was preferred compared to KF, but KF was far more popular then the original noisy data.

6. CONCLUSIONS

In this paper, a Kalman filter based speech enhancement system is studied. The system comprises of three parts, a voice activity detector, a parameter estimation unit, and a Kalman filter.

Depending on the SNR in the processed data frame, two different strategies for speech AR parameter estimation are employed. For high frame SNRs the AR parameters are estimated from the noisy speech, while for low SNR these parameters are post-processed in order to reduce the estimation error due to the additive noise. An algorithm for this purpose has been outlined in some detail. It is based on standard signal processing building blocks such as FFT and the autocorrelation method, in combination with a low complexity iterative scheme.

It has been demonstrated that the complete noise reduction system is able to reduce the background noise level, in the mobile telephony scenario, with approximately 10dB without introducing any speech distortion or distortion of the background noise. Thus, it may be a suitable alternative for front-end noise reduction in the mobile telephony scenario.

7. REFERENCES

[1] T. Ohya, H. Suda and T. Miki, "5.6 kbits/s PSI-CELP of the half-rate PDC speech coding standard", *1994 IEEE 44th Vehicular Technology Conference*, Stockholm, Sweden, pp. 1680–1684, 1994.

[2] B.D.O. Anderson and J.B. Moore, *Optimal Filtering*, Prentice-Hall Inc., Englewood Cliffs, NJ, 1979.

[3] J.D. Gibson, B. Koo and S.D. Gray, "Filtering of colored noise for speech enhancement and coding", *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 39, no. 8, pp. 1732–1742, August 1991.

[4] T. Söderström and P. Stoica, *System Identification*, Prentice Hall, Englewood Cliffs, NJ, 1988.

[5] P. Händel, "Low-distortion spectral subtraction for speech enhancement", *4th European Conference on Speech Communication and Technology*, Madrid, Spain, vol. 2, pp. 1549–1552, 18-21 September 1995.