# KALMAN FILTERING OF COLORED NOISE FOR SPEECH ENHANCEMENT

*Dimitrie C. Popescu*

Department of Electrical Engineering
Rutgers University
Piscataway, NJ 08855-0909

*Ilija Zeljković*

AT&T Labs
180 Park Ave.
Florham Park, NJ 07932

**ABSTRACT**

A method for applying Kalman filtering to speech signals corrupted by colored noise is presented. Both speech and colored noise are modeled as autoregressive (AR) processes using speech and silence regions determined by an automatic end-point detector. Due to the non-stationary nature of the speech signal, non-stationary Kalman filter is used. Experiments indicate that non-stationary Kalman filtering outperforms the stationary case, the average SNR improvement increasing from 0.53 dB to 2.3 dB. Even better results are obtained if noise is considered also non-stationary, in addition to being colored, achieving an average of 7.14 dB SNR improvement.

## 1. INTRODUCTION

In many cases background noise is the dominant source of errors in automatic speech recognition (ASR). This is especially true for public phones, and phones located in industrial environments. If not modeled properly, the high intensity noise is often confused for speech by the recognition system.

The best recognition performance in presence of stationary noise is achieved if background and speech models are trained and tested under the same noise conditions. For telephone services this approach is not practical since there are many different environments where calls can originate. Thus, other techniques that better discriminate between noise and speech, or reduce the background noise must be used.

There have been numerous studies [2], [4], [5] dealing with enhancement of speech contaminated by noise. However, most approaches use the standard stationary Gaussian white noise assumption. Colored noise assumption [1] proved to be very useful for speech enhancement [2].

The method used in our work to improve the signal-to-background noise ratio (SNR) for speech signals is Kalman filtering. Colored noise that corrupts the speech signal is modeled from white noise through a shaping filter.

Results indicate that non-stationary Kalman filters lead to a significant gain in SNR, thus visibly improving the quality of the speech signal.

## 2. PROBLEM STATEMENT

Consider the noise free speech signal described by the $p$-th order AR model

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + w(n) \tag{1}$$

where $w(n)$ is zero-mean Gaussian white noise with intensity $Q(n)$. The canonical state-space model is obtained by concatenating $p$ consecutive values of the speech signal $s$, denoting them by $\mathbf{x}_1(n) = [s(n-p+1)\ s(n-p+2)\ \dots\ s(n)]^T$ and writing corresponding equations in matrix form

$$\mathbf{x}_1(n) = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \\ a_p & a_{p-1} & \cdots & a_1 \end{bmatrix} \mathbf{x}_1(n-1) + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} w(n)$$

$$s(n) = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \mathbf{x}_1(n) \tag{2}$$

which can be rewritten as

$$\mathbf{x}_1(n) = A_s \mathbf{x}_1(n) + G_s w(n)$$

$$s(n) = C_s \mathbf{x}_1(n) \tag{3}$$

The speech signal $s(n)$ is contaminated by zero-mean additive Gaussian noise $v(n)$ which is colored, but independent of $w(n)$

$$y(n) = s(n) + v(n) \tag{4}$$

Colored noise will be modeled by the same type of AR equations, but of lower order $m$

$$v(n) = \sum_{l=1}^{m} b_l v(n-l) + u(n) \tag{5}$$

where $u(n)$ is zero-mean Gaussian white noise with intensity $R(n)$, not correlated with $w(n)$. The canonical state-space model for colored noise is obtained similarly, by concatenating $m$ consecutive values of $v$ and denoting them by $\mathbf{x}_2(n) = [v(n-m+1)\ v(n-m+2)\ \dots\ v(n)]^T$.

$$\mathbf{x}_2(n) = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \\ b_m & b_{m-1} & \cdots & b_1 \end{bmatrix} \mathbf{x}_2(n-1) + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u(n)$$

$$v(n) = \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \mathbf{x}_2(n) \tag{6}$$

which can be rewritten as

$$\mathbf{x}_2(n) = A_n \mathbf{x}_2(n) + G_n u(n)$$

$$v(n) = C_n \mathbf{x}_2(n) \tag{7}$$

In conclusion, the process is modeled by the following state-space equations

$$\mathbf{x}_1(n) = A_s \mathbf{x}_1(n-1) + G_s w(n) \tag{8}$$

$$\mathbf{x}_2(n) = A_n \mathbf{x}_2(n-1) + G_n u(n) \tag{9}$$

$$y(n) = C_s \mathbf{x}_1(n) + C_n \mathbf{x}_2(n) \tag{10}$$

By adjoining states in (8) and (9) the augmented system will have the form

$$\mathbf{x}(n) = A_a \mathbf{x}(n-1) + G_a \nu(n-1)$$
$$y(n) = C_a \mathbf{x}(n) \tag{11}$$

where $\mathbf{x}(n) = [\mathbf{x}_1(n)\ \mathbf{x}_2(n)]^T$, $\nu(n) = [w(n)\ u(n)]^T$, and the augmented matrices are

$$A_a = \begin{bmatrix} A_s & 0 \\ 0 & A_n \end{bmatrix}, \quad G_a = \begin{bmatrix} G_s & 0 \\ 0 & G_n \end{bmatrix}$$

$$C_a = [C_s\ C_n]$$

Relations (11) describe a linear system driven by white noise $\nu$ with intensity matrix $Q_a(n) = \begin{bmatrix} Q(n) & 0 \\ 0 & R(n) \end{bmatrix}$ but with no measurement noise; by including colored noise model as an additional state in the state-space model, noise that affected the output $y$ has been moved into the state $\mathbf{x}$.

## 3. THE KALMAN FILTER

The equations of the non-stationary Kalman filter for system (11) which has "perfect measurements" are [3]

$$\begin{aligned} &\hat{\mathbf{x}}(n) = \hat{A}(n-1)\hat{\mathbf{x}}(n-1) + K(n-1)y(n) \\ &K(n) = A_a P(n) C_a^T [C_a P(n) C_a^T]^{-1} \\ &\hat{A}(n) = A_a - K(n) C_a \\ &P(n+1) = \hat{A}(n) P(n) A_a^T + G_a Q_a(n) G_a^T \end{aligned} \tag{12}$$

where $\hat{\mathbf{x}}(n)$ is the optimal estimate of $\mathbf{x}(n)$, $K(n)$ is the filter gain, and $P(n)$ is the covariance of the error between the actual $\mathbf{x}(n)$ and its estimate $\hat{\mathbf{x}}(n)$.

Since only an estimate of the noise-free speech signal $s$ is needed, the output equation of the Kalman filter will be

$$\hat{s}(n) = [C_s\ 0_m] \begin{bmatrix} \hat{\mathbf{x}}_1(n) \\ \hat{\mathbf{x}}_2(n) \end{bmatrix} \tag{13}$$

where $0_m$ is a row vector of dimension $m$ having all entries zero, and $\hat{s}(n)$ is the optimal estimate of the speech signal.

The state-space model allows also estimation of the colored noise that corrupts the speech signal. For this, a second output equation needs to be added to the Kalman filter

$$\hat{v}(n) = [0_p\ C_n] \begin{bmatrix} \hat{\mathbf{x}}_1(n) \\ \hat{\mathbf{x}}_2(n) \end{bmatrix} \tag{14}$$

with $0_p$ a row vector with of dimension $p$ with all entries zero.

It must be noted here that Kalman filter offers optimal estimate when the system parameters are known, so that it is important that system matrices $A_a$, $G_a$, $C_a$, and especially noise intensity $Q_a(n)$, be modeled as accurate as possible.

## 4. PARAMETER ESTIMATION

The speech signal is modeled using linear predictive analysis (LPC), which is the predominant technique for estimating basic speech parameters [6]. The idea behind LPC, is that speech samples can be expressed as a linear combination of past speech samples, so that the speech signal can be modeled as an AR process (1). The importance of linear prediction lies in the accuracy with which the basic model applies to speech, which is - as it has already been mentioned - of crucial importance for Kalman filtering.

An automatic speech end-point detector is used for discrimination of speech vs. silence [6] in the noisy speech signal; models for speech/noise are obtained from speech/silence regions indicated by the end-point detector. However, the filtering algorithm is not depending on the end-point information.

In order to fully utilize the power of Kalman filtering - which is non-stationary filtering as opposed to Wiener filtering which is stationary - the most important problem is to accurately detect the non-stationary noise, especially impulsive noise [1].

The total intensity of the speech signal contaminated with uncorrelated noise is $P = Q + R$, where $Q$ is the intensity of the clean speech signal, and $R$ is noise intensity.

The power of $y(n)$ is smoothed by a low-pass filter:

$$P(n) = (1-\lambda)P(n-1) + \lambda y(n)^2 \tag{15}$$

with the time constant defined by $\lambda$. The higher the value of $\lambda$, the longer the period over which the power is averaged.

Conventionally, the intensity of stationary noise is independent of time, i.e. $R(n) = R,\ \forall n$, and it is estimated from the portion of the signal assumed to be noise (beginning of the utterance), or declared as background signal by a speech end-pointer.

In addition to the stationary noise estimates from the portion of the signal declared as noise, a new approach is used to estimate time-varying noise intensity based on distinct properties of speech and noise signals:

- Sudden changes in signal intensity indicate a beginning or an end of a impulsive noise. The rate of change for the intensity of the speech signal is limited by the inertia of human speech production system. It is widely accepted that the speech signal remains stationary (produced by unchanged vocal tract) within 5 to 10 ms segments, thus any any quicker change in the speech signal can be attributed to impulsive noise. We estimate the instantaneous noise intensity $R(n)$ to be proportional to the rate of signal intensity change. The change is measured as a difference between slightly and heavily smoothed variance of the signal.
- Noise signal tends to be dominated by high frequency components and is much less autocorrelated (is more white) than the speech signal. The degree of noisiness is computed based on first autocorrelation coefficient and the position of the dominant pole of the low order (2 or 3) LPC model of the signal on very short segments (around 5 ms). Other methods like zero-crossing, reflection coefficients, and signal intensity can also be used.

The noise intensity is made proportional to the degree of noisiness. The duration of the impulse is determined either by following the change of parameters (intensity, spectrum), or by assumption that the impulse noise is short (20-50 ms). In either case the sudden

[1] that is to accurately estimate time-varying noise intensity

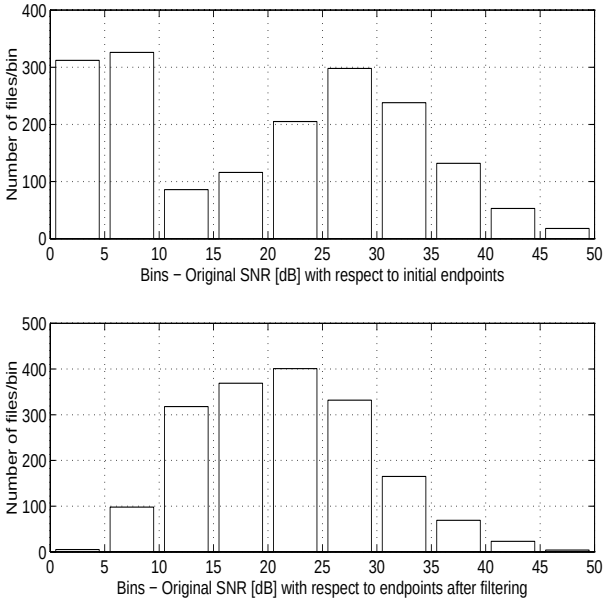


Figure 1: Original SNR distribution before and after non-stationary Kalman filtering of non-stationary colored noise

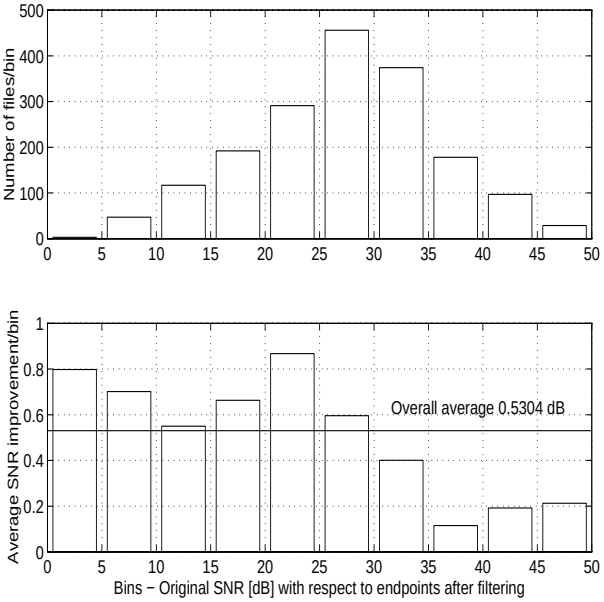


Figure 2: SNR improvement for stationary Kalman filtering of colored noise

change of contaminated speech intensity is immediately reflected in noise intensity. In later case, the rapid change in the signal intensity is immediately passed to noise intensity which is than decaying slowly with a prescribed rate.

## 5. EXPERIMENTAL RESULTS

The performance criterion used for evaluating the proposed Kalman filtering algorithms is the SNR defined by:

$$\text{SNR} = 10\log_{10}\frac{\frac{1}{N_s}\sum_{n=1}^{N_s} s^2(n)}{\frac{1}{N_n}\sum_{n=1}^{N_n} v^2(n)}[\text{dB}] \qquad (16)$$

with speech $s$, noise $v$, and their corresponding lengths $N_s$, $N_n$, determined using the automatic end-point detector.

In addition to the improvement in SNR, which will be discussed below, Kalman filtering has also a beneficial effect on end-point detection. For very noisy signals, in which speech and noise intensities are of the same order, the end-pointer very often fails to detect accurate end-points, part of speech being labeled as noise. On the other hand, impulsive noise (bangs or clicks), which has dynamics similar to short speech segments, may be labeled as speech by the end-pointer. Also, very often noise at the beginning and at the end of the speech, as well as in between words or group of words is labeled as speech, thus falsely increasing the SNR. Figure 1 shows SNR distribution for 1800 testing files, before and after non-stationary Kalman filtering of non-stationary colored noise. It is easily noticed that the number of files with SNR greater than 25 dB is almost double when this is calculated with respect to initial end-points than when it is calculated with respect to end-points determined after filtering. Also, for the majority of very noisy files (0 to 5 dB) the end-pointer labels speech as noise before filtering, while after filtering end-points are more accurate.

For stationary Kalman filtering the SNR improvements are comparable to those in [2], even though their SNR definition is slightly different. The average SNR improvement is 0.54 dB, the maximum being as much as 0.85 dB. Figure 2 presents the distribution of the SNR improvement for stationary Kalman filtering, with respect to the initial SNR of the signal. The same observation as in [2] can be made about the order of the system: SNR improvements increase with increasing the order of both the speech model, and the colored noise model. However, all experiments (stationary and non-stationary) have used the same orders $p = 10$ for speech and $m = 5$ for colored noise. The use of non-stationary Kalman filtering improves results from stationary case. Even when considering that noise is white and stationary, but speech is non-stationary, results are better as it can be observed from figure 3. Results become even better when stationary colored noise model is used, figure 4.

However, the best results are obtained when both speech and colored noise, are considered non-stationary, with intensities estimated according to the procedures of the previous section; average improvement is 7.1 dB, with more than 8 dB for signals with initial SNR $\in [15, 25]$ dB, figure 5.

All the above mentioned results are contained in Table 1 which gives an overview of average SNR improvements that can be obtained with Kalman filtering.

| Speech int. $Q$ | Noise int. $R$ | Noise type | Avg. SNR impr. [dB] |
|---|---|---|---|
| stat. | stat. | colored | 0.53 |
| non-stat. | stat. | white | 1.87 |
| non-stat. | stat. | colored | 2.3 |
| non-stat. | non-stat. | colored | 7.14 |

Table 1: Average SNR improvement after Kalman filtering

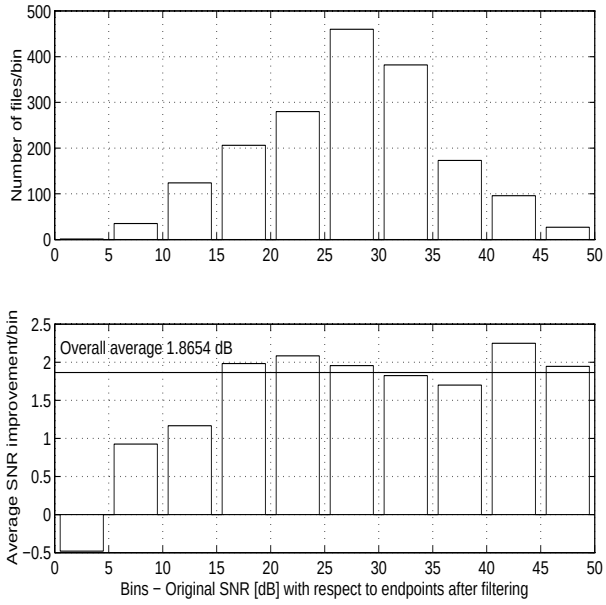


Figure 3: SNR improvement for non-stationary Kalman filtering of white noise

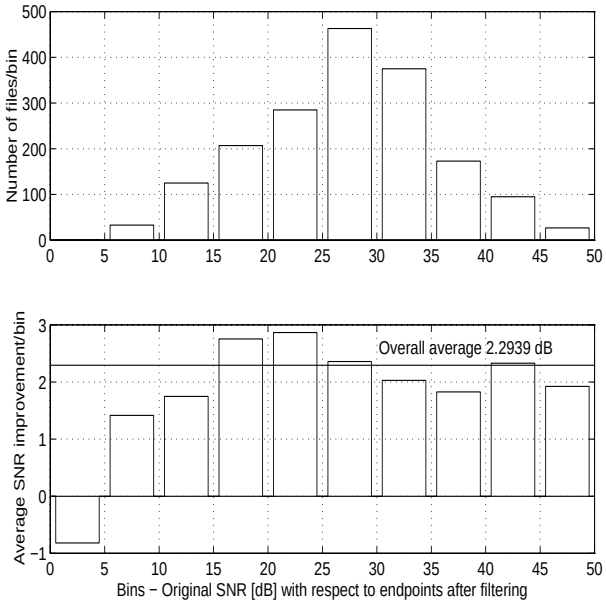


Figure 4: SNR improvement for non-stationary Kalman filtering of stationary colored noise

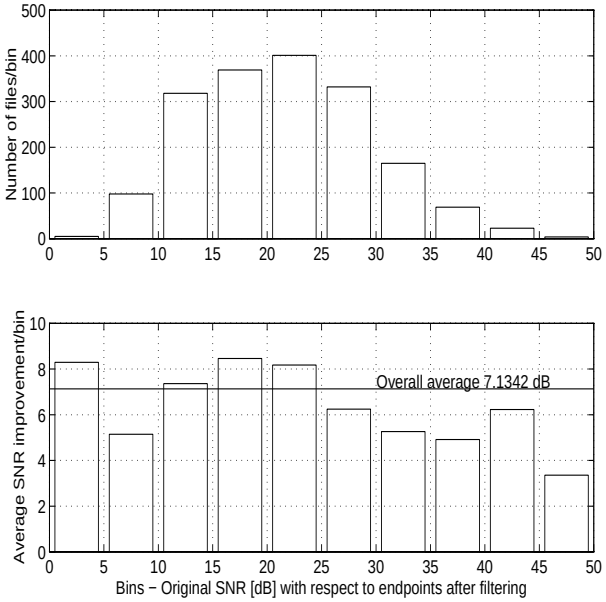


Figure 5: SNR improvement for non-stationary Kalman filtering of non-stationary colored noise

## 6. CONCLUSIONS

The paper shows that non-stationary Kalman filtering is a good tool for improving the SNR for speech signals corrupted by non-stationary colored noise. As it has already been mentioned, best results are achieved when both speech, and colored noise are modeled as non-stationary signals, SNR being improved with an average of 7.1 dB.

The paper also shows that end-point detection is improved; the end-point detector applied on the filtered signal gives more accurate results than on the original signal.

## 7. REFERENCES

[1] Bryson A. E., Johansen D. E., "Linear Filtering for Time-Varying Systems Using Measurements Containing Colored Noise", *IEEE Trans. Automatic Control*, January 1965, pp. 4-10.

[2] Gibson J. D., Koo B., Gray S. D., "Filtering of Colored Noise for Speech Enhancement and Coding", *IEEE Trans. Signal Processing*, vol. 39, No. 8, Aug. 1991, pp. 1732-1742.

[3] Kalman R. E., "A New Approach to Linear Filtering and Prediciton Problems", *Trans. ASME, J. Basic. Eng.*, Series 82D, Mar. 1960, pp. 35-45.

[4] Lim J. S., Oppenheim A. V., "All-pole modelling of degraded speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp.197-210, June 1978.

[5] Paliwal K. K., Basu A., "A speech enhancement method based on Kalman filtering", *Proc. IEEE Conf. Acoust., Speech, Signal Processing*, Dallas, TX, April 1987, pp. 177-180.

[6] Rabiner L. R., Schafer R.W., *Digital Processing of Speech Signals*, Prentice-Hall 1978.