# A KALMAN FILTER WITH A PERCEPTUAL POST-FILTER TO ENHANCE SPEECH DEGRADED BY COLORED NOISE

**Ning Ma [1], Martin Bouchard [1], and Rafik A. Goubran [2]**
[1] School of Information Technology and Engineering, University of Ottawa, 800 King Edward,
Ottawa, Ontario, Canada, K1N 6N5 email: {nma, bouchard} @site.uottawa.ca
[2] Department of Systems and Computer Engineering, Carleton University, 1125 Colonel By Drive,
Ottawa, Ontario, K1S 5B6, Canada, email: Rafik.Goubran@sce.carleton.ca

## 1. INTRODUCTION

Speech enhancement algorithms have been employed successfully in many areas such as VoIP, automatic speech recognition and speaker verification. Some of the methods assume that the environmental noise is white noise. However, when used in colored noise environments, those methods will produce a weaker performance. Approaches for colored noise have also been previously proposed, however those previous methods have to detect non-speech frames for the noise covariance estimation. This paper proposes a method for colored noise speech enhancement based on a Kalman filter combined with a post-filter using masking properties of human auditory systems. No detection of non-speech frames is needed in the proposed method.

## 2. THE PERCEPTUAL KALMAN FILTER ALGORITHM

A clean speech signal $s(n)$ can be modeled as:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{1}$$

where $s(n)$ is the $n$-th sample of the clean speech signal, $u(n)$ is a zero mean white Gaussian process with variance $\sigma_u^2$ and $a_i$ is the $i$-th autoregressive (AR) model parameter. The $n$-th sample of the noisy speech signal $y(n)$ is:

$$y(n) = s(n) + v(n) \tag{2}$$

where $v(n)$ is a colored measurement noise process with covariance matrix **R**. Using a vector Kalman filter as in [1], a state-space model can be expressed as

$$\mathbf{x}(n) = \mathbf{F}\mathbf{x}(n-1) + \mathbf{G}u(n) \tag{3}$$

$$\mathbf{y}(n) = \mathbf{H}\mathbf{x}(n) + \mathbf{v}(n) \tag{4}$$

where

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix} \tag{5}$$

$$\mathbf{x}(n) = [s(n-p+1) \cdots s(n)]^T \tag{6}$$

$$\mathbf{y}(n) = [y(n-p+1) \cdots y(n)]^T \tag{7}$$

$$\mathbf{v}(n) = [v(n-p+1) \cdots v(n)]^T \tag{8}$$

$$\mathbf{G} = [0\ 0 \cdots 1]^T \tag{9}$$

and **H** is a $p$-th order identity matrix. Thus the Kalman filter estimation and updating equations are as follows:

$$\mathbf{e}(n) = \mathbf{y}(n) - \hat{\mathbf{x}}(n \mid n-1) \tag{10}$$

$$\mathbf{K}(n) = \mathbf{P}(n \mid n-1) \times (\mathbf{P}(n \mid n-1) + \mathbf{R})^{-1} \tag{11}$$

$$\hat{\mathbf{x}}(n \mid n) = \hat{\mathbf{x}}(n \mid n-1) + \mathbf{K}(n) \times \mathbf{e}(n) \tag{12}$$

$$\mathbf{P}(n \mid n) = (\mathbf{I} - \mathbf{K}(n)) \times \mathbf{P}(n \mid n-1) \tag{13}$$

$$\hat{\mathbf{x}}(n+1 \mid n) = \mathbf{F}\hat{\mathbf{x}}(n \mid n) \tag{14}$$

$$\mathbf{P}(n+1 \mid n) = \mathbf{F}\mathbf{P}(n \mid n)\mathbf{F}^T + \mathbf{G}\mathbf{G}^T \sigma_u^2 \tag{15}$$

where $\mathbf{e}(n)$ is the innovation vector, $\mathbf{K}(n)$ is the Kalman gain, $\hat{\mathbf{x}}(n \mid n)$ represents the filtered estimate of the state vector $\mathbf{x}(n)$, $\hat{\mathbf{x}}(n \mid n-1)$ is the minimum mean-square estimate of the state vector $\mathbf{x}(n)$ given the past observations $y(1)$, …, $y(n\text{-}1)$, $\mathbf{P}(n \mid n)$ is the filtered state error covariance matrix; and $\mathbf{P}(n \mid n-1)$ is the *a priori* error covariance matrix. The last element of $\hat{\mathbf{x}}(n \mid n)$, $\hat{s}(n)$, is the output of the Kalman filter. The computation of the model noise process statistics (variance $\sigma_u^2$ ) and the measurement colored noise statistics (covariance matrix **R**) can be done with a covariance matching method, as in [1].

The Kalman filtered speech signal is then processed by a post-filter, on frame-by-frame basis, with the frame length defined as $L$. Both time domain forward masking effects and frequency domain simultaneous masking properties are considered in the proposed post-filter. The psychoacoustic time domain forward masking effects are modeled as a psychoacoustic specific loudness versus critical-band rate and time. The total loudness $Q$, defined as the sum of the output specific loudness in all critical bands, is used as an estimate of the time domain forward masking level [2]. The total masking level is determined by integrating the frequency-domain simultaneous masking effect and the time-domain forward masking effect [2]. The total masking level of the $i^{\text{th}}$ critical band ( $i = 1, 2, \cdots, 18$ ) at time $t$ is:

$$M_t(i) = \max\{M_s(i), M_t(i)^* \cdot \exp^{-\Delta t/(\tau(i)\cdot Q)}\} \qquad (16)$$

where $M_t(i)$ and $M_t(i)^*$ are the total masking levels of the current frame and the previous frame, respectively; $M_s(i)$ is the masking level computed from the simultaneous frequency domain masking model [3], $\Delta t$ is the time difference between two frames, $\tau(i)$ is the maximum decay time constant in each critical band [2], and $Q$ is the total loudness level computed as in [2]. The post-filter performs thresholding on its input signal based on the computed total masking level $M_t(i)$. To perform the thresholding, the following procedure is used:

(1) Mapping the total masking level $M_t(i)$ in each critical band to frequency domain (FFT bins) to obtain $T(\omega_j)$ ( $j = 0, 1, \cdots, 255$ ).

(2) From previously computed filtered state error covariance matrices ( $\mathbf{P}(n-i \mid n-i)$ $i = 0,1,2,\cdots$ ), estimate a covariance function for the filtered state error signal. The power spectrum density (PSD) $P_e(\omega_i)$ of the filtered state error signal is then computed by a 256-points FFT of the covariance vector, i.e., for $\omega_j = 2\pi/256 \cdot j$ ( $j = 0,\cdots,255$ ). Then the thresholding is performed on the speech spectrum:

$$\left|\tilde{\hat{S}}(\omega_i)\right| = \begin{cases} \left|\hat{S}(\omega_i)\right| \times \alpha^{P_e(\omega_i)/T(\omega_i)} & \text{if } P_e(\omega_i) < T(\omega_i) \\ \left|\hat{S}(\omega_i)\right| \times \alpha \times (1 + \alpha^{P_e(\omega_i)/T(\omega_i)}) & \text{otherwise} \end{cases} \qquad (17),$$

where $\alpha$ ( $0 < \alpha < 1$ ) is a tonality coefficient [3].

(3) Doing an IFFT using $\left|\tilde{\hat{S}}(\omega_j)\right|$ and the phase of $\hat{S}(\omega_j)$, keeping the last $L$ values of the size-256 IFFT outputs to obtain the frame of improved speech signal.

From (17), the masking properties and the characteristics of the speech frame are taken into account. If $P_e(\omega_i) < T(\omega_i)$, most of the power in the Kalman filtered signal can be kept, to reduce distortion. If $P_e(\omega_i) > T(\omega_i)$, whether to increase the power of the Kalman filtered signal or to reduce it depends on the value of $\alpha$ and $P_e(\omega_i)$. For $\alpha > (\sqrt{5} - 1)/2$ (tone-like frame) and $P_e(\omega_i) < T(\omega_i)\ln(1/\alpha - 1)/\ln\alpha$, the power of the Kalman filtered signal is increased. In any other cases, it is reduced.

## 3. SIMULATION RESULTS

Four different speech sentences of 4.75 seconds spoken by 2 females and 2 males were used in the simulations. A colored noise $v(n)$ was used as the noise source, and it was obtained by running a white noise signal through a 8th order AR filter. The frame size was 80 samples ($L$=80), i.e. 10 ms frames. The AR prediction order $p$ was set to be 10, and the AR coefficients were updated for every frame. The data length used to compute the AR parameters was 160 samples (the current noisy frame and one previous enhanced frame). The performance index used was ITU-T P.862 PESQ scores, in order to have a close match with subjective speech quality scores. In the P.862 standard, the lowest PESQ score is –0.5 and the highest score is 4.5. High scores stand for good speech quality. The PESQ scores obtained by using the proposed approach and other recent algorithms [3,4] for colored noise under various noisy speech signal-to-noise ratios (SNR) are shown in Table 1. The simulation results show that in the view of the PESQ scores, the new proposed method has the best performance for any input noisy speech SNR value.

## 4. CONCLUSION

In this paper, a total masking threshold including frequency domain simultaneous masking effects and time domain forward masking effects was applied as a post-filter to a Kalman filtered signal, to further enhance it in a perceptual sense. A thresholding procedure suitable for colored noise based on the computed masking level was proposed. Simulation results have shown that the new method leads to very promising results. No speech versus noise detection (i.e. VAD) is required in the proposed method.

## REFERENCES

[1] Ma, N., Bouchard M. and Goubran, R. (2004). Perceptual Kalman filtering for speech enhancement in colored noise, Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)

[2] Huang, Y.-H. and Chiueh, T.-D. (2002). A New Audio Coding Scheme Using a Forward Masking Model and Perceptually Weighted Vector Quantization, IEEE Trans. Speech and Audio Proc., vol.10, 325-335.

[3] Virag, N. (1999). Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System, IEEE Trans. Speech Audio Proc., vol. 7, 126-137

[4] Popescu, D. C. and Zeljkovic, I. (1998). Kalman Filtering of Colored Noise for Speech Enhancement, Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP), vol.2, 997-1000.

Table 1. PESQ scores obtained by different algorithms

| | PESQ Scores | | | | |
|---|---|---|---|---|---|
| Input SNR (dB) | Original Noisy Speech | Spectral Subtraction | Method from [3] | Method from [4] | Proposed Method |
| -5 | 1.293 | 1.266 | 1.296 | 1.371 | **1.577** |
| 0 | 1.605 | 1.636 | 1.673 | 1.692 | **1.906** |
| 5 | 1.882 | 1.950 | 1.981 | 1.977 | **2.227** |
| 10 | 2.182 | 2.259 | 2.304 | 2.292 | **2.549** |
| 15 | 2.502 | 2.585 | 2.627 | 2.603 | **2.873** |