PERCEPTUAL KALMAN FILTERING FOR SPEECH ENHANCEMENT IN COLORED NOISE

Ning Ma, Martin Bouchard*, and Rafik. A. Goubran***

* School of Information Technology and Engineering, University of Ottawa,
800 King Edward, Ottawa (Ontario), K1N 6N5, Canada, email: bouchard@site.uottawa.ca
** Department of Systems and Computer Engineering, Carleton University, 1125 Colonel By Drive,
Ottawa, Ontario, K1S 5B6, Canada, email: Rafik.Goubran@sce.carleton.ca

ABSTRACT

A new method for speech enhancement in colored noise is proposed in this paper. A Kalman filter concatenated with a post-filter based on masking properties of human auditory systems is proposed for the problem. A recursive approach to compute the noise covariance matrix is used for estimating the colored noise statistics. In the post-filter, both time domain masking properties and frequency domain masking properties are taken into account. From the calculated masking level, the noisy speech spectrum is adjusted accordingly. Simulation results show that the proposed approach has the best performance compared with other recent methods, evaluated with PESQ scores.

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 [1-2] assume that the environmental noise is white noise. When used in colored noise environments, the methods will produce a weaker performance.

In [3], signal subspace approaches for speech enhancement with colored noise were proposed. In [4], a Kalman filter based approach was proposed for colored noise cases. These methods have to detect non-speech frames for the noise covariance estimation. This paper proposes a new Kalman filter based method combined with a post-filter using masking properties of human auditory systems. The noise covariance is estimated recursively using a covariance matching method, and no detection of non-speech frames is needed. Section 2 introduces the perceptual Kalman filter method. Both time domain masking and frequency domain masking properties are used in the approach. Section 3 illustrates the recursive estimation of the noise covariance matrix in the Kalman filter. The estimation is computed by matching the theoretical and estimated values of the noise covariance. Section 4 shows the simulation results. Conclusions are drawn in Section 5.

This work was supported by the National Science and Engineering Research Council (NSERC), Canada, and the National Capital Institute of Telecommunications (NCIT), Canada

2. PERCEPTUAL KALMAN FILTER

In a colored noise environment, a mixed Kalman filter is proposed. A post-filter based on human auditory system properties is concatenated with the Kalman filter.

2.1 Noisy speech model and mixed Kalman filter

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

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

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

The n-th sample of the noisy speech signal $y(n)$ is:

$$y(n) = s(n) + v(n) \quad (2)$$

where $v(n)$ is a colored measurement noise process, with covariance matrix $\mathbf{R}$. Using a mixed Kalman filter, the state-space model is expressed as

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

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

where

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

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

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

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

$$\mathbf{G} = [0\ 0\cdots 1]^T \quad (9)$$

and $\mathbf{H}$ is a p-th order identity matrix.

Thus the Kalman filter estimation and updating equations are as follows:

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

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

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

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

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

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

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 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 $y(1), \ldots, y(n-1)$, $\mathbf{P}(n \mid n)$ 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.

2.2 Post-filter based on masking properties of auditory systems

In the post-filter, the Kalman filtered speech signal is processed on frame-by-frame basis. After each noisy speech frame is passed through the Kalman filter, the whole frame of the filtered signal $\hat{\mathbf{s}}(n) = [\hat{s}(n-L+1)\cdots\hat{s}(n)]$ is input into the post-filter, where L is the frame length. Both time domain forward masking effects and frequency domain simultaneous masking properties are considered in the proposed system. 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 [5]. The total masking level is determined by integrating the frequency-domain simultaneous masking effect and the time-domain forward masking effect [5]. The level depends not only on the current frame but also on the previous frames. The total masking level of the i^{th} critical band ($i = 1, 2, \cdots, 18$) at time t is:

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

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 [6], Δt is the time difference between two frames, $\tau(i)$ is the maximum decay time constant in each critical band [5], and Q is the total loudness level computed as in [5].

The post-filter performs thresholding on the 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 $T(\omega_j)$ ($j = 1, 2, \cdots, 256$).

(2) Performing the thresholding on the Kalman filtered speech spectrum $\hat{S}(\omega_j)$:

$$\left|\tilde{\hat{S}}(\omega_j)\right| = \begin{cases} \alpha\left|\hat{S}(\omega_j)\right|, & \text{if } \left|\hat{S}(\omega_j)\right|^2 < T(\omega_j); \\ \left|\hat{S}(\omega_j)\right| \times \dfrac{T(\omega_j)}{T_{\max}}, & \text{otherwise.} \end{cases} \quad (17)$$

where α is a constant and $0 < \alpha < 1$ (in our simulations $\alpha = 0.8$) and $T_{\max}$ is the maximum value of $T(\omega_j)$ ($j = 1, 2, \cdots, 256$). $\hat{S}(\omega_j)$ is computed by a 256-points FFT, inserting (256-L) zeros at the beginning of $\hat{\mathbf{s}}(n)$.

(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 improved frame of speech signal.

From (17), the speech spectrum is reshaped using the masking threshold. When the frequency component of the noisy speech signal at a certain frequency point is smaller than the threshold, the noise component at that frequency may be masked. If the masking threshold was an ideal one, we could keep the component at that frequency unchanged. However, the masking threshold is computed from a noisy speech signal. To reduce the effect of the noise, the amplitude component is decreased by some constant factor α. When the energy of the noisy speech at a certain frequency is greater than the masking threshold, the noise may not be masked. Then a factor computed from the threshold is used to decrease the amplitude and make the noise component smaller. A larger threshold represents a larger speech energy at that frequency, thus the amplitude is less changed, while a smaller threshold represents a smaller speech energy, and the amplitude is then decreased more, to mask the noise.

3. RECURSIVE COVARIANCE ESTIMATION

In the previous section, the speech model noise is assumed to be a zero mean white Gaussian noise and only the variance is required to be estimated, while the measurement is assumed to be colored noise. The noise statistics computation is based on a covariance matching

method. The measurement noise covariance matrix is found by using the theoretical and estimated covariance of the innovation process, while the model noise variance is found by using the theoretical and estimated variance of the residual process. The recursive procedure for the noise statistics estimation is presented in the sub-sections below.

3.1 Estimation of Model Noise Process Statistics

The estimation of the model noise process statistics is derived under the assumption of a constant mean and variance over N samples of $u(n), u(n-1), \cdots, u(n-N+1)$ [1]. From (3), the model noise process is given by the following equation:

$$u(n) = \mathbf{G}^T [\mathbf{x}(n) - \mathbf{F}\mathbf{x}(n-1)] \tag{18}$$

Although $\mathbf{x}(n)$ and $\mathbf{x}(n\text{-}1)$ are unknown, $u(n)$ can be represented by its approximation $\beta(n)$ as follows:

$$\beta(n) = \mathbf{G}^T [\mathbf{x}(n \mid n) - \mathbf{x}(n \mid n-1)] \tag{19}$$

Then by computing the mean and variance of $\beta(n)$, the unbiased estimate of the variance of $u(n)$ is obtained by:

$$\begin{aligned}\hat{\sigma}_u^2 = \frac{1}{N-1}\sum_{i=0}^{N-1} \Big\{ & [\beta(n-i) - \hat{\bar{\beta}}(n)]^2 \\ & - \frac{N-1}{N}\mathbf{G}^T\mathbf{F}\mathbf{P}(n-i-1 \mid n-i-1)\mathbf{F}^T\mathbf{G} \\ & - \frac{N-1}{N}\mathbf{G}^T\mathbf{P}(n-i \mid n-i)\mathbf{G} \\ & + 2\frac{N-1}{N}\mathbf{G}^T\mathbf{P}(n-i \mid n-i-1)[\mathbf{I} - \mathbf{K}(n-i)\mathbf{G}^T]^T\mathbf{G}\Big\}\end{aligned} \tag{20}$$

where $\hat{\bar{\beta}}(n)$ is the mean of the last N measurements $\beta(n), \beta(n-1), \cdots, \beta(n-N+1)$. In the simulation, N is chosen to equal 80 samples.

3.2 Estimation of Measurement Colored Noise Statistics

The estimation is derived under the assumption that the colored noise is wide sense stationary. From (4), the colored noise vector is represented by:

$$\mathbf{v}(n) = \mathbf{y}(n) - \mathbf{x}(n) \tag{21}$$

The true state vector $\mathbf{x}(n)$ is unknown, so $\boldsymbol{v}(n)$ can not be determined, but the innovation process $\mathbf{e}(n)$ can approximate the noise process, as (10) shows. From (4) and (10), $\mathbf{e}(n)$ can be written as:

$$\mathbf{e}(n) = \mathbf{x}(n) - \hat{\mathbf{x}}(n \mid n-1) + \mathbf{v}(n) = \tilde{\mathbf{x}}(n \mid n-1) + \mathbf{v}(n) \tag{22}$$

where $\tilde{\mathbf{x}}(n \mid n-1)$ is the *a priori* error vector and its covariance matrix is $\mathbf{P}(n \mid n-1)$.

The covariance matrix of $\mathbf{e}(n)$ is represented by:

$$\mathbf{C}_e(n) = \mathbf{P}(n \mid n-1) + \mathbf{R} \tag{23}$$

where $\mathbf{v}(n)$ and $\tilde{\mathbf{x}}(n \mid n-1)$ are assumed uncorrelated and $\mathbf{R}$ is the covariance matrix of the measured colored noise.

The current unbiased estimate of $\mathbf{C}_e(n)$ is computed as follows:

$$\hat{\mathbf{C}}_e(n) = \frac{n-1}{n}\hat{\mathbf{C}}_e(n-1) + \frac{1}{n}[\mathbf{e}(n) - \bar{e}(n)] \cdot [\mathbf{e}(n) - \bar{e}(n)]^T \tag{24}$$

where $\bar{e}(n)$ is the mean value of the innovation process calculated from all past values of $\mathbf{e}(n)$, $\mathbf{e}(n\text{-}1)$, …, $\mathbf{e}(1)$. $\hat{\mathbf{C}}_e(n)$ is actually the mean value of $\mathbf{C}_e(n), \cdots, \mathbf{C}_e(1)$. From (23), the mean value of $\mathbf{C}_e(n)$ can also be computed by the following equation:

$$\bar{\mathbf{C}}_e(n) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}(i \mid i-1) + \bar{\mathbf{R}} \tag{25}$$

where $\bar{\mathbf{R}}$ is the mean of $\mathbf{R}$ (or the unbiased estimate of $\mathbf{R}$). Matching (24) and (25), the unbiased estimate of $\mathbf{R}$ is then represented as:

$$\bar{\mathbf{R}} = \hat{\mathbf{C}}_e(n) - \frac{1}{n}\sum_{i=1}^{n}\mathbf{P}(i \mid i-1) \tag{26}.$$

4. 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 was used as the noise source. The colored noise $v(n)$ was obtained by running a white noise signal $w(n)$ through a 8th order AR filter as follows:

$$\begin{aligned} v(n) = & -0.0851v(n-1) + 0.19126v(n-2) \\ & + 0.0458v(n-3) + 0.0229v(n-4) \\ & + 0.1097v(n-5) + 0.1553v(n-6) \\ & - 0.132v(n-7) - 0.76v(n-8) + w(n) \end{aligned} \tag{27}$$

All of the speech files were taken from the ITU-T Supplement P.23 speech database. The sampling frequency used was 8000 Hz, and the input signals were normalized so that the amplitude is in the interval [-1, 1]. 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 (i.e. to compute correlation values) was 160 samples, which includes the current noisy frame and one previous enhanced frame. The performance index used is 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 different algorithms for colored noise and various noisy speech signal-to-noise ratios (SNR) are shown in Fig. 1.

The simulation results show that in the view of the

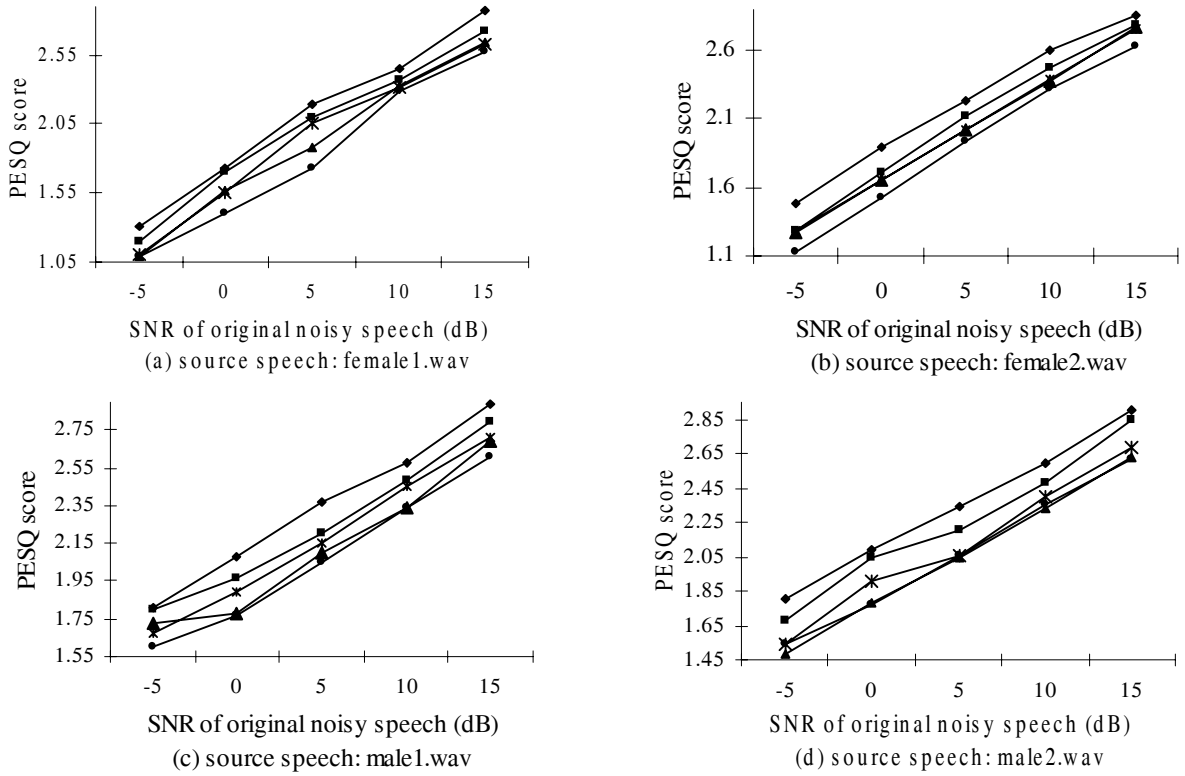

Fig. 1. PESQ scores obtained by different methods, for two female speech files (upper figures) and two male speech files (lower figures), with colored noise signal. ◆: proposed approach, ■: approach based on Kalman filtering with simultaneous masking properties [2], ▲: Kalman filtering approach [1], *: spectral subtraction approach with simultaneous masking properties [6], •: noisy speech before enhancement.

PESQ scores the new proposed method has the best performance for any input noisy speech SNR value.

5. CONCLUSIONS

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 enhanced it in a perceptual sense. A thresholding procedure suitable for colored noise based on the computed masking level was proposed. A recursive computation method for estimating the noise covariance matrix was proposed for the enhancement of speech signals in colored noise. The estimation of the noise covariance matrix was made during the parameter updating step of the Kalman filtering algorithm. Simulation results have shown that the idea leads to very promising results. No speech versus noise detection (i.e. VAD) is required in the proposed method.

6. REFERENCES

[1] M. Gabrea, "Adaptive Kalman Filtering-Based Speech Enhancement Algorithm", in *Proc. of Canadian Conference on Electrical and Computer Engineering 2001*, Vol. 1, pp.521-526, Fredericton, New-Brunswick, 2001.

[2] N. Ma, M. Bouchard and R. A. Goubran, "A Perceptual Kalman Filtering-Based Approach for Speech Enhancement", *Proceedings of IEEE ISSPA 2003*, Vol. 1, pp. 373-376, Paris, France, July 2003

[3] Udar Mittal and Nam Phamdo, "Signal/Noise KLT Based Approach for Enhancing Speech Degraded by Colored Noise", *IEEE Trans. Speech Audio Processing*, Vol. 8, No. 2, pp. 159-167, Mar. 2000.

[4] D. C. Popescu, I. Zeljkovic, "Kalman Filtering of Colored Noise for Speech Enhancement", in *Proc. of IEEE ICASSP'98*, vol.2, pp. 997-1000, Seattle, USA, May 1998.

[5] Y.-H. Huang, and T.-D. Chiueh, "A New Audio Coding Scheme Using a Forward Masking Model and Perceptually Weighted Vector Quantization", *IEEE Trans. Speech and Audio Processing*, Vol.10, No.5, pp.325-335, Jul. 2002

[6] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System", *IEEE Trans. Speech Audio Processing*, Vol. 7, No. 2, pp.126-137, Mar. 1999