A WAVELET KALMAN FILTER WITH PERCEPTUAL MASKING 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, e-mail: {nma, bouchard}@site.uottawa.ca
** Department of Systems and Computer Engineering, Carleton University, 1125 Colonel By Drive,
Ottawa, Ontario, K1S 5B6, Canada, e- mail: Rafik.Goubran@sce.carleton.ca

ABSTRACT

A wavelet Kalman filter combined with masking properties of human auditory systems is proposed for enhancing speech degraded by colored noise. A wavelet domain speech state-space model is first constructed, with a corresponding Kalman filter. A post-filter considering both time and frequency masking properties is then applied. From the calculated masking threshold, the noisy speech spectrum is further enhanced. Simulation results show that the proposed approach has the best performance compared with either classic or recently introduced methods, evaluated using ITU-T P.862 PESQ scores. The proposed approach also has a reduced complexity compared to a time-domain implementation of the Kalman filter with a perceptual post-filter.

1. INTRODUCTION

The Kalman filtering algorithm was first proposed to be applied to speech enhancement by Paliwal and Basu [1], with the speech parameters obtained from the clean speech signal and the noise characteristics obtained from non-speech frames. In [2], a Kalman filter based approach was proposed for colored noise cases. These methods have to detect non-speech frames for the noise covariance estimation. In [3], the authors proposed a Kalman filter approach concatenated with a perceptual post-filter for speech enhancement in colored noise. The noise covariance is estimated recursively using a covariance matching method, and no detection of non-speech frames is needed. Extending the work in [3], this paper presents a wavelet domain Kalman filter for speech enhancement. Since the perceptual post-filter is based on human auditory masking properties and the frequency response of a human auditory system can be mapped into the wavelet domain [4], a wavelet domain speech state space model is constructed so that the Kalman filtering procedure and the post-filter procedure can be done in the same domain. Thus the computational load becomes less than for the time domain Kalman filter as in [3]. The wavelet Kalman filter can get the MMSE or the LMMSE estimate of the speech spectrum in the wavelet domain. In the rest of this paper, Section 2 briefly introduces the wavelet Kalman filter for speech enhancement, Section 3 describes the perceptual post-filter, Section 4 shows the simulation results and Section 5 presents a brief conclusion.

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

2. WAVELET KALMAN FILTER FOR SPEECH ENHANCEMENT

For the problem of enhancing speech degraded by colored noise, a wavelet packet transform is used to construct the state-space model in the wavelet domain.

2.1 Wavelet packet transform and filter banks

For a given sequence of signals $x(i,n)$ at time n and for a fixed resolution level i, the lower resolution signal $x_L(i-1,n)$ is obtained by downsampling the output of a halfband lowpass filter by two, where the impulse response of the filter is $h(n)$:

$$x_L(i-1,n) = \sum_{k=-\infty}^{\infty} h(2n-k)x(i,k) \qquad (1).$$

The wavelet coefficients, as a complement of $x_L(i-1,n)$, are denoted by $x_H(i-1,n)$, and can be computed by first using a highpass filter with an impulse response $g(n)$ and then by downsampling the output of the highpass filter by two. This yields:

$$x_H(i-1,n) = \sum_{k=-\infty}^{\infty} g(2n-k)x(i,k) \qquad (2)$$

with $g(n) = (-1)^n h(L-1-n)$, where L is the length of the filter and must be even. Equations (1) and (2) define the

discrete wavelet transform. In this paper, the filter impulse responses used form an orthonormal set. Thus the reconstruction of the original signal $x(i,n)$ is computed by

$$x(i,n) = \sum_{k=-\infty}^{\infty} h(2k-n)x_L(i-1,k) + \sum_{k=-\infty}^{\infty} g(2k-n)x_H(i-1,k) \quad (3).$$

The discrete wavelet transform can be implemented by an octave-band filter bank [5]. However, the wavelet transform only decomposes the output of the lowpass filter at each resolution level. Then it is difficult to map the frequency response of the human auditory system to the wavelet transform domain. Whereas using a wavelet packet transform the mapping problem can be solved: by using a wavelet packet transform, both the output of the lowpass filter and the highpass filter can be further processed. A wavelet packet tree can be used to represent the operation [5]. For this paper, it is more convenient to describe the wavelet packet transform in an operator form. For simplicity, the wavelet packet transform is denoted by a matrix **T**. An orthogonal wavelet basis is used, so the inverse of matrix **T** equals its transpose, i.e. $\mathbf{T}^{-1} = \mathbf{T}'$.

2.2 State-Space model

The speech signal is processed on a block by block basis, and the length of each block equals the LPC order p (see below). Since the length of the data to be transformed by wavelet packets usually is a power of 2, in the simulations the frame length was set to be F =128 and the LPC order p was set to be 8. There are thus $B = F/p$ =16 blocks within a frame, the block index is denoted by *N*, and the frame index is denoted by *w*.

A clean speech signal *s*(*n*) can be described by a LPC model:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \quad (4)$$

where p is the LPC order, $u(n)$ is a zero mean, white Gaussian process with variance σ_u^2, and a_i is the *i*-th AR parameter. The *n*-th sample of the noisy speech signal $y(n)$ is described as:

$$y(n) = s(n) + v(n) \quad (5)$$

where $v(n)$ is a colored measurement noise process. The *N-th* block of clean speech data is written as:

$$\mathbf{S}_N = [s((N-1)\cdot p+1), s((N-1)\cdot p+2), \cdots, s((N-1)\cdot p+p)]'$$

where ' denotes a transpose operation. From (5), the relationship between $\mathbf{S}_N$ and $\mathbf{S}_{N+1}$ is as follows:

$$\mathbf{S}_{N+1} = \mathbf{F}\mathbf{S}_N + \mathbf{G}\mathbf{U}_{N+1} \quad (6)$$

where

$$\mathbf{U}_N = [u((N-1)\cdot p+1), u((N-1)\cdot p+2), \cdots, u((N-1)\cdot p+p)]'$$

and the transition matrix **F** is computed by the following steps. First, a matrix $\mathbf{F}_1$ is written as:

$$\mathbf{F}_1 = \begin{bmatrix} a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \\ 0 & a_p & a_{p-1} & \cdots & a_2 \\ 0 & 0 & a_p & \cdots & a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_p \end{bmatrix}$$

The first row of the matrix **F** equals the first row of the matrix $\mathbf{F}_1$. From the second row to the *p*-th row, the elements of the *i*-th row of matrix **F** are computed by:

$$(\mathbf{F})_i = \sum_{k=0}^{i-1} (\mathbf{F}_1)_{i-k} \cdot a_k \ , \quad (7)$$

where $a_0 = 1$, $(\mathbf{F})_i$ and $(\mathbf{F}_1)_{i-k}$ represent the *i*-th row of matrix **F** and the (i-*k*)-th row of matrix $\mathbf{F}_1$, respectively. The matrix **G** is computed by applying the same scheme as in (7) to an identity matrix of size $p \times p$.

Based on the above block state space model, the following frame by frame state space model is obtained. The adjacent frames have an overlap of $B-1$ = 15 blocks. Let $\mathbf{X}_w$ denote the clean speech data of frame *w*:

$$\mathbf{X}_w = [\mathbf{S}'_{(w-1)+1}, \mathbf{S}'_{(w-1)+2}, \cdots \mathbf{S}'_{(w-1)+B}]'$$

$$\mathbf{X}_{w+1} = \mathbf{A}\mathbf{X}_w + \mathbf{\Gamma}\vec{\mathbf{U}}_{w+1} \quad (8)$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{F} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{F} \end{bmatrix}, \qquad \mathbf{\Gamma} = \begin{bmatrix} \mathbf{G} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{G} \end{bmatrix},$$

$$\vec{\mathbf{U}}_w = [\mathbf{U}'_{(w-1)+1}, \mathbf{U}'_{(w-1)+2}, \cdots \mathbf{U}'_{(w-1)+B}]'.$$

The covariance matrix associated with $\vec{\mathbf{U}}_w$ is **Q**. The observation data is represented by

$$\vec{\mathbf{Y}}_w = \mathbf{X}_w + \vec{\mathbf{V}}_w \quad (9)$$

where

$$\vec{\mathbf{Y}}_w = [\mathbf{Y}'_{(w-1)+1}, \mathbf{Y}'_{(w-1)+2}, \cdots \mathbf{Y}'_{(w-1)+B}]',$$

$$\vec{\mathbf{V}}_w = [\mathbf{V}'_{(w-1)+1}, \mathbf{V}'_{(w-1)+2}, \cdots \mathbf{V}'_{(w-1)+B}]',$$

$$\mathbf{Y}_N = [y((N-1)\cdot p+1), y((N-1)\cdot p+2), \cdots, y((N-1)\cdot p+p)]'$$

and

$$\mathbf{V}_N = [v((N-1)\cdot p+1), v((N-1)\cdot p+2), \cdots, v((N-1)\cdot p+p)]'$$

The covariance matrix associated with $\vec{\mathbf{V}}_w$ is **R**. Eq. (8) and (9) form the frame by frame state space model in the

time domain. To transform it into the wavelet domain, the wavelet packet transform is applied on both sides of (8):

$$\mathbf{TX}_{w+1} = \mathbf{TAX}_w + \mathbf{T\Gamma}\vec{\mathbf{U}}_{w+1} = \mathbf{TAT'TX}_w + \mathbf{T\Gamma}\vec{\mathbf{U}}_{w+1} \quad (10).$$

Let $\vec{\mathbf{X}}_w = \mathbf{TX}_w$ which is the wavelet packet transform of the clean speech data, $\underline{\mathbf{A}} = \mathbf{TAT'}$ and $\underline{\mathbf{\Gamma}} = \mathbf{T\Gamma}$, then in the wavelet domain, the state space model is represented by:

$$\vec{\mathbf{X}}_{w+1} = \underline{\mathbf{A}}\vec{\mathbf{X}}_w + \underline{\mathbf{\Gamma}}\vec{\mathbf{U}}_{w+1} \quad (11)$$

$$\vec{\mathbf{Y}}_w = \mathbf{T'}\vec{\mathbf{X}}_w + \vec{\mathbf{V}}_w \quad (12).$$

The following Kalman filtering algorithm can be used for this model:

$$\mathbf{e}(w) = \vec{\mathbf{Y}}_w - \mathbf{T'}\vec{\mathbf{X}}_{w|w-1} \quad (13)$$

$$\mathbf{K}(w) = \mathbf{P}(w \mid w-1) \times \mathbf{T}(\mathbf{T'P}(w \mid w-1)\mathbf{T} + \mathbf{R})^{-1} \quad (14)$$

$$\vec{\hat{\mathbf{X}}}_{w|w} = \vec{\hat{\mathbf{X}}}_{w|w-1} + \mathbf{K}(w) \times \mathbf{e}(w) \quad (15)$$

$$\mathbf{P}(w \mid w) = (\mathbf{I} - \mathbf{K}(w)\mathbf{T'}) \times \mathbf{P}(w \mid w-1) \quad (16)$$

$$\vec{\hat{\mathbf{X}}}_{w+1|w} = \underline{\mathbf{A}}\vec{\hat{\mathbf{X}}}_{w|w} \quad (17)$$

$$\mathbf{P}(w+1 \mid w) = \underline{\mathbf{A}}\mathbf{P}(w \mid w)\underline{\mathbf{A}}' + \underline{\mathbf{\Gamma}}\mathbf{Q}\underline{\mathbf{\Gamma}}' \quad (18)$$

where $\mathbf{e}(w)$ is the innovation vector, $\mathbf{K}(w)$ is the Kalman gain, $\vec{\hat{\mathbf{X}}}_{w|w}$ represents the filtered estimate of state vector $\vec{\mathbf{X}}_w$, $\vec{\hat{\mathbf{X}}}_{w|w-1}$ is the minimum mean-square estimate of the state vector $\vec{\mathbf{X}}_w$, $\mathbf{P}(w \mid w)$ is the filtered state error covariance matrix, and $\mathbf{P}(w \mid w-1)$ is the *a priori* error covariance matrix. The estimation of the covariance matrices **Q** and **R** can be derived in the time domain [3].

3. POST-FILTER BASED ON MASKING PROPERTIES OF AUDITORY SYSTEMS

To perform the post-filter masking and thresholding, the following procedure is used:

(1) Using the wavelet domain clean speech estimates $\vec{\hat{\mathbf{X}}}_{w|w}$ to compute the total masking level M_t in each critical band using the procedure in [3]

(2) Decomposing each critical band masking level M_t into the corresponding wavelet packet coefficients $T(\omega_j)$ ($j = 0, \cdots, 127$). In the simulation, a full wavelet packet tree was used and each coefficient $T(\omega_j)$ corresponds to a frequency bandwidth of 31.25 Hz (with a sampling frequency of 8000 Hz).

(3) Estimating the power spectrum density (PSD) of the filtered state error in the wavelet packet domain, from the last 20 filtered state error covariance matrices

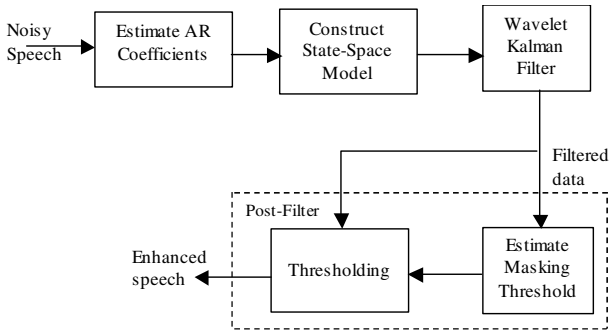

Fig. 1 Diagram of the Proposed System

$\mathbf{T'P}(w-i \mid w-i-1)\mathbf{T}$ ($i = 0, \cdots, 19$, from (14)). This produces an average covariance matrix $\overline{\mathbf{P}}$. Then a covariance vector is constructed from the matrix $\overline{\mathbf{P}}$, by averaging the elements of $\overline{\mathbf{P}}$ in each of its diagonals. The PSD in the wavelet domain $P_e(\omega_i)$ is estimated by using a wavelet packet transform on the covariance vector. $P_e(\omega_i)$ is then the absolute value of the transform coefficients. A thresholding is then performed on the components $\hat{X}(\omega_i)$ of the input wavelet packet transformed speech spectrum $\vec{\hat{\mathbf{X}}}_{w|w}$:

$$\left|\tilde{\hat{X}}(\omega_i)\right| = \begin{cases} \left|\hat{X}(\omega_i)\right| \times \alpha^{P_e(\omega_i)/T(\omega_i)} & P_e(\omega_i) < T(\omega_i) \\ \left|\hat{X}(\omega_i)\right| \times \alpha \times (1 + \alpha^{P_e(\omega_i)/T(\omega_i)}) & \text{otherwise} \end{cases} \quad (19)$$

where $i = 0,1, \cdots 127$, and α is a tonality coefficient [3] ($0 \le \alpha \le 1$, 0 is for a purely white noise signal and 1 is for a purely tonal signal).

(4) Using $\left|\tilde{\hat{X}}(\omega_i)\right|$ and the sign of $\hat{X}(\omega_i)$, and then doing an inverse wavelet packet transform to obtain an enhanced frame of speech. The last block of data in the frame is then averaged with the second last block of data from the previous frame, to produce the final block of enhanced speech.

When the post-filter is designed, a tradeoff between signal distortion and residual noise is needed. Both the masking properties and the characteristics of the speech frame are taken into account in the thresholding procedure (19). The smaller the α is, i.e. the speech frame is more like a white noise, the more the Kalman filtered signal power is reduced. When the error power density $P_e(\omega_i)$ is smaller than the masking threshold $T(\omega_i)$, most of the power of the Kalman filtered signal can be kept to reduce distortion. The smaller the $P_e(\omega_i)$ is, the more the

Kalman filtered signal energy is kept. On the other hand, if $P_e(\omega_i)$ is larger than $T(\omega_i)$, whether to increase the power of the Kalman filtered signal or to reduce it depends on the value of α and $P_e(\omega_i)$. From (19), if $P_e(\omega_i) > T(\omega_i)$ and $\alpha(1+\alpha^{P_e(\omega_i)/T(\omega_i)}) > 1$, then the following two solutions can be derived: $\alpha > \frac{\sqrt{5}-1}{2}$ and $P_e(\omega_i) < \frac{T\ln(1/\alpha - 1)}{\ln\alpha}$. This case represents a speech frame which is more tone-like and the error estimate may not be too large. To mask the error estimate, the power of the Kalman filtered signal is increased. In other cases, the power of the Kalman filtered signal is reduced. If $\alpha > \frac{\sqrt{5}-1}{2}$ and $P_e(\omega_i) > \frac{T\ln(1/\alpha - 1)}{\ln\alpha}$, the error estimate may be too large. If the signal power is increased, it may incur a large distortion. Thus in this case, the signal power is reduced. The larger the $P_e(\omega_i)$ is, the more the signal power is reduced. If $\alpha < \frac{\sqrt{5}-1}{2}$, the speech frame is assumed to be more like a noise frame. The larger the $P_e(\omega_i)$ is, the more the signal power is reduced.

4. SIMULATION RESULTS

Six different speech sentences of 4 seconds each spoken by three females and three males were used in our simulations. The noise signals used were babble noise and street noise. All of the speech and noise 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 AR prediction order *p* was set to be 8, 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 136 samples, which includes the current noisy speech frame and one previous enhanced speech block. DB4 wavelets were used and the decomposition level of the data was 7. The performance index used was ITU-T P.862 PESQ scores, in order to have a close match with subjective speech quality scores. High scores stand for good speech quality. The PESQ scores are shown in Table 1, obtained by comparing the proposed method with classical and recently introduced speech enhancement algorithms for colored noise, under various input signal-to-noise ratios (SNR). The simulation results show that in the view of PESQ scores, the new proposed method has the best performance under any input SNR value. The slight improvement over our previous time domain perceptual Kalman filtering method in [3] is due to improvements in the post-filter thresholding. The number of required multiplies for the Kalman filtering and the noise covariance matrices estimation in each frame was estimated to be reduced from 278,528 in [3] to 34,816 multiplies in the proposed approach. Moreover, the latter does not require a FFT on the input data of the post-filter.

input speech SNR → algorithm ↓	-5 dB	0 dB	5 dB	10 dB	15 dB
original noisy speech	1.17	1.50	1.78	2.09	2.42
spectral subtract. (SS)	1.14	1.53	1.85	2.17	2.53
perceptual SS [6]	1.17	1.57	1.89	2.22	2.56
Kalman filtering [2]	1.26	1.59	1.89	2.24	2.55
KLT approach [7]	1.14	1.59	1.93	2.29	2.61
perceptual Kalman [3]	1.37	1.75	2.03	2.39	2.71
proposed approach	1.47	1.91	2.14	2.49	2.83

Table 1. Average PESQ scores obtained by different methods.

5. CONCLUSION

This paper proposed a wavelet Kalman filter concatenated with a perceptual post-filter for speech enhancement. The multi-resolution property of human auditory systems can be mapped very well to the wavelet packet transform domain. By constructing the state-space model in the wavelet domain, the output of the wavelet Kalman filtering algorithm is directly used by the post-filter without any transform, reducing the complexity compared to a pure time domain Kalman filter approach. The proposed method produced a better PESQ score performance than other speech enhancement algorithms for colored noise, and it does not require the detection of noise frames (i.e. a voice activity detector), unlike some other algorithms.

6. REFERENCES

[1] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *Proc. of IEEE ICASSP'87*, vol. 12, pp. 297-300, April 1987.

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

[3] N. Ma, M. Bouchard and R. A. Goubran, "Perceptual Kalman Filtering for Speech Enhancement in Colored Noise", *Proc. of IEEE ICASSP'2004*, Montreal, Canada, May 2004.

[4] Y. Huang, and T. Chiueh, "A New Audio Coding Scheme Using a Forward Masking Model and Perceptually Weighted Vector Quantization", *IEEE Trans. Speech Audio Processing*, vol. 10, pp.325-33, July 2002.

[5] C.K.Chui, G. Chen, *Kalman Filtering with Real-Time Applications*, 3rd Edition, Springer-Verlag, 1999

[6] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System", *IEEE Trans. Speech Audio Processing*, vol. 7, pp.126-137, Mar. 1999

[7] U. Mittal and N. Phamdo, "Signal/Noise KLT Based Approach for Enhancing Speech Degraded by Colored Noise", *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 159-167, Mar. 2000.