Non-Uniform Sub-Band Kalman Filtering for Speech Enhancement

Phu Ngoc Le, Eliathamby Ambikairajah
School of Electrical Engineering and Telecommunications
The University of New South Wales, UNSW Sydney, NSW 2052, Australia
phule@unsw.edu.au, ambi@ee.unsw.edu.au

***Abstract*-In this paper, a novel method for single-channel speech enhancement based on Kalman filtering is proposed. Instead of applying the Kalman algorithm for full-band speech or uniform sub-band speech, speech enhancement is performed by applying the Kalman algorithm to non-uniform sub-band signals obtained from the decomposition of whole-band speech using gammatone filters. Simulation results indicate that in terms of perceptual evaluation speech quality (PESQ), our proposed method shows a relative improvement of 21.4% over the conventional full-band Kalman filtering technique.**

I. Introduction

Speech enhancement aims to extract a clean speech estimate from a noisy observed signal. Given that we live in a natural environment where noise is ubiquitous, speech enhancement is necessary for many real-life systems such as mobile communication, speech recognition and speech verification. However, speech enhancement is a very difficult task due to the non-stationary nature of a speech signal. As a result, it still poses a challenge to researchers despite the availability of various algorithms. Among several approaches used to improve the quality of speech signal, Kalman filters are widely known not only for their low speech distortion but also for their reasonable computation complexity [1], [2]. Recently, it has been reported that by applying Kalman filters to individual sub-band speech signals which were previously decomposed by a filter bank, the effectiveness of speech enhancement is much improved in terms of performance as well as computational complexity [2], [3]. Although this uniform sub-band Kalman filtering method has many advantages compare to full-band Kalman filtering in speech enhancement, problem still remains as reported in [2]. It is the mismatch between the uniform filter bank exploited to decompose the full-band speech signal and the critical band of human auditory system which is non-uniform. The performance of speech enhancement in this method is therefore still limited because the masking property of human auditory system is not fully incorporated into speech enhancement process.

In this paper, a simple but effective method for single-channel speech enhancement is introduced. In this method, speech is firstly decomposed by an array of analysis filters. The individual sub-band speech is then filtered using the Kalman algorithm. The sub-band signals at the output of Kalman filters are then passed through reconstruction filters whose impulse responses are the time reverse of those of the analysis filters. The final enhanced speech is obtained from the synthesis of sub-band signals at the output of the reconstruction filters. The analysis and reconstruction filters in this paper are designed as gammatone filters which are widely recognized as a model of the human auditory system [4], [5]. To evaluate the performance of our method, PESQ scores are used, which are well known for their high correlation with subjective speech quality scores [6]. The relative improvement of our proposed method and the conventional method compared to the noisy speech is also determined in order to evaluate the performance of our method more intuitively.

II. Kalman Filter for Speech Enhancement

On a short-time basis, a speech sequence can be modeled as an autoregressive (AR) process which considers each speech sample as the output of an all-pole linear system driven by an excitation signal:

$$x(n) = \sum_{i=1}^{p} a_i x(n-i) + \omega(n) \qquad (1)$$

where $n = 1,2,...$ is the sample index, a_i is the ith AR coefficient, and the excitation signal $\omega(n)$ is a zero-mean white Gaussian noise process with variance σ_ω^2. The speech signal $s(n)$ is assumed to be contaminated by a zero-mean additive white Gaussian noise $v(n)$ with variance σ_v^2,

$$s(n) = x(n) + v(n). \qquad (2)$$

Let $\mathbf{x}(n) = [x(n)\, x(n-1) ... x(n-p+1)]^T$. To apply Kalman filtering, (1) and (2) are reformulated in the state-space domain as

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

$$s(n) = \mathbf{h}^{\mathrm{T}}\mathbf{x}(n) + v(n) \quad (4)$$

where

$$\mathbf{F} = \begin{bmatrix} a_1 & a_2 & \cdots & a_{p-1} & a_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{p\times p}, \quad \mathbf{g} = \mathbf{h} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}_{p\times 1} \quad (5)$$

By applying the Kalman algorithm, the optimal estimate of $\mathbf{x}(n)$ can be obtained as

$$\hat{\mathbf{x}}(n) = \mathbf{F}\hat{\mathbf{x}}(n-1) + \mathbf{k}(n)\left[s(n) - \mathbf{h}^{T}\mathbf{F}\hat{\mathbf{x}}(n-1)\right] \quad (6)$$

where

$$\mathbf{k}(\mathbf{n}) = \mathbf{P}(n \mid n-1)\,\mathbf{h}\,[R + \mathbf{h}^{T}\mathbf{P}(n \mid n-1)\mathbf{h}]^{-1} \quad (7)$$

$$\mathbf{P}(n \mid n-1) = \mathbf{F}\mathbf{P}(n-1)\mathbf{F}^{T} + \mathbf{g}Q\mathbf{g}^{T} \quad (8)$$

$$\mathbf{P}(n) = \left[\mathbf{I} - \mathbf{k}(n)\mathbf{h}^{T}\right]\mathbf{P}(n \mid n-1) \quad (9)$$

here $\hat{\mathbf{x}}(n)$ is the estimate of $\mathbf{x}(n)$, $\mathbf{k}(n)$ is the Kalman gain, $\mathbf{P}(n \mid n-1) = \mathbf{E}\left\{\left[\mathbf{x}(n) - \mathbf{F}\hat{\mathbf{x}}(n-1)\right]\left[\mathbf{x}(n) - \mathbf{F}\hat{\mathbf{x}}(n-1)\right]^{T}\right\}$ is the prediction-error covariance matrix where $\mathbf{E}$ is the expectation operator, $\mathbf{P}(n) = \mathbf{E}\left\{\left[\mathbf{x}(n) - \hat{\mathbf{x}}(n)\right]\left[\mathbf{x}(n) - \hat{\mathbf{x}}(n)\right]^{T}\right\}$ is the estimation error covariance matrix, $R = \sigma_v^2$ is the measurement-noise variance, and $Q = \sigma_\omega^2$ is the driving-noise variance. A speech sample at time instant n can be estimated as $\hat{x}(n) = \mathbf{h}^{\mathrm{T}}\hat{\mathbf{x}}(n)$.

III. Proposed Approach

By filtering the speech signal $s(n)$ using a bank of M analysis filters, assuming that the mth analysis filter has the impulse response $h_m(n)$, the sub-band speech on mth band can be obtained as

$$s_m(n) = s(n) * h_m(n) \quad (10)$$

The Kalman algorithm is then applied to each sub-band speech $s_m(n)$. The outputs of Kalman filters $y_m(n)$ are then passed through reconstruction filters which are the time reverse of the analysis filters. The final enhanced speech signal is obtained from the synthesis of sub-band signals $u_m(n)$.

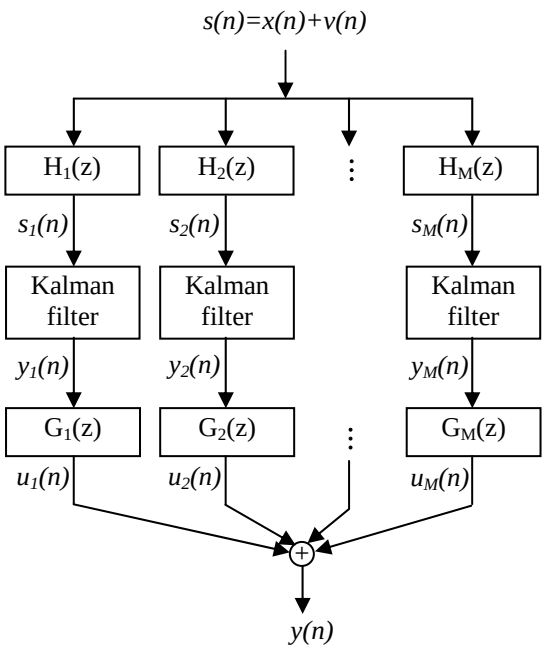

Fig. 1. Sub-band speech enhancement system.

The analysis filter bank in this paper is based on the gammatone filters whose frequency responses are matched to the critical bands [4]. Practically, gammatone filters can be implemented using FIR or IIR filters. In this paper, FIR filters were employed in order to implement linear phase filters with identical delay in each critical band [6]. The analysis filter for the mth sub-band is obtained by the following expression,

$$h_m(n) = a_m(nT)^{N-1}e^{-2\pi bBW_m nT}\cos(2\pi f_{cm}nT) \quad (11)$$

where f_{cm} is the center frequency of the mth sub-band, T is the sampling period, n is the discrete time sampling index, BW_m is the bandwidth of the mth filter, the constant $b = 1.65$ and values for a_m were selected for each filter such that the filter gain was normalized to $0dB$. The magnitude response of 18 gammatone filters in the frequency range from 0 to 4 kHz is showed in Fig. 2.

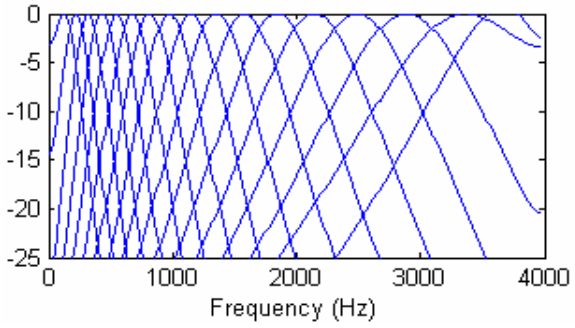

Fig. 2. Magnitude response of gammatone filters.

To achieve perfect reconstruction, the reconstruction filters $g_m(n)$ are designed as the time reverse of the analysis filter, $g_m(n) = h_m(-n)$.

IV. Simulation Results

In order to evaluate the performance of our proposed method, extensive simulations have been carried out to achieve speech enhancement. In these simulations, the 8 kHz sampling rate input speech samples were taken from the EBU SQAM dataset. This dataset consists of speech samples spoken by 6 speakers (three males and three females). The speech enhancement performance using the proposed method was also confirmed by employing eleven types of common background noises that are listed in Table 2. These noise signals were taken from NOISEX-92 database. In our simulations, the noise variance was adjusted to obtain input noisy speech signal SNRs ranging from -5 dB to 10 dB. This adjustment is expressed in the following equation:

$$s(n) = x(n) + v(n) \times \sqrt{\frac{\sigma_s^2}{\sigma_v^2} \times \frac{1}{10^{SNR/10}}} \qquad (12)$$

where σ_s^2 and σ_v^2 are the power of clean speech and noise respectively, and SNR is the signal to noise ratio in dB of the input noisy speech signal.

In this paper, speech enhancement was implemented within frames of 128 samples corresponding to a duration of 16 ms. The speech AR prediction order p was set to 10. The simulations in this paper were initially performed on an ideal case where the AR coefficients are estimated from clean speech and noise power is perfectly estimated. This was done to judge the performance of an ideal Kalman filter. Later they were replaced with AR estimates obtained from pre-enhanced speech using spectral subtraction. The performance indices used were ITU-T P.862 PESQ scores which are highly correlated with subjective speech quality scores. In the P.862 standard, PESQ scores range from 4.5 (the highest quality of speech) down to -0.5. Fig. 3 shows the PESQ scores of enhanced speech obtained from our proposed method in the same plot with those of noisy speech and full-band Kalman-filter-enhanced speech. It is clear that in terms of PESQ scores, the proposed method out-performs the conventional Kalman filter, especially at the low SNR regions.

In order to have a more intuitive comparison, the measure of relative PESQ improvement introduced in [4] was also employed in this paper. This measure is described as follows

$$\delta(\%) = \frac{PESQ_{proc} - PESQ_{ref}}{PESQ_{ref}} \times 100$$

Here $PESQ_{ref}$ and $PESQ_{proc}$ are the $PESQ$ scores of the noisy speech and enhanced speech, respectively, referenced to the clean speech.

Table 1 shows the *PESQ* improvement in relative measure δ of six different speech signals contaminated by car noise using different enhancement algorithms. The values of *PESQ* improvement of our proposed NSK (Non-uniform sub-band Kalman filter) method are always much larger than those of the conventional FK (full-band Kalman filter) method. From this table, it's also likely that the enhancement performance of our proposed method for female speech is better than that of male speech.

Table 2 shows the *PESQ* improvement in relative measure δ of English female speech contaminated by eleven different kinds of noise using different enhancement algorithms. Similar to the previous case, our proposed method outperforms the conventional method.

It is more intuitive to describe the PESQ improvement using a bar diagram as in Fig. 4. This bar diagram shows the result of speech enhancement for German female speech contaminated by street noise at various values of SNR using our proposed algorithm and conventional algorithms.

TABLE I
PESQ Improvement δ (%) of Various Speech Files Contaminated by Car Noise at 0 dB SNR

Speech	FK (%)	NSK (%)
English Male	36.84	**54.16**
English Female	33.11	**58.80**
French Male	22.85	**55.59**
French Female	33.87	**72.14**
German Male	28.19	**59.43**
German Female	65.07	**134.70**

TABLE 2
PESQ Improvement δ (%) of English Female Speech Signal Contaminated by Different Kinds of Noise at 0 dB SNR

Noise	FK (%)	NSK (%)
Airport noise	40.8327	**68.37**
Babble noise	48.4252	**100.39**
Car noise	37.0567	**80.85**
Exhibition noise	136.1446	**174.83**
Factory noise	61.1366	**91.75**
Pink noise	69.7318	**98.75**
Restaurant noise	129.6937	**181.49**
Street noise	60.8916	**108.21**
Subway noise	40.8327	**68.37**
Train noise	35.2545	**56.27**
White noise	63.6520	**79.33**

A total of 264 combinations from six speech signals, eleven noisy signals at four SNR values which are -5 dB, 0 dB, 5 dB, and 10 dB were used to test the algorithm. It is derived that the average improvement in terms of PESQ our proposed method is 0.25 higher when compared to that of conventional method. In terms of relative measure δ, the performance of our proposed method is 21.4% better than that of the full-band Kalman filter method.

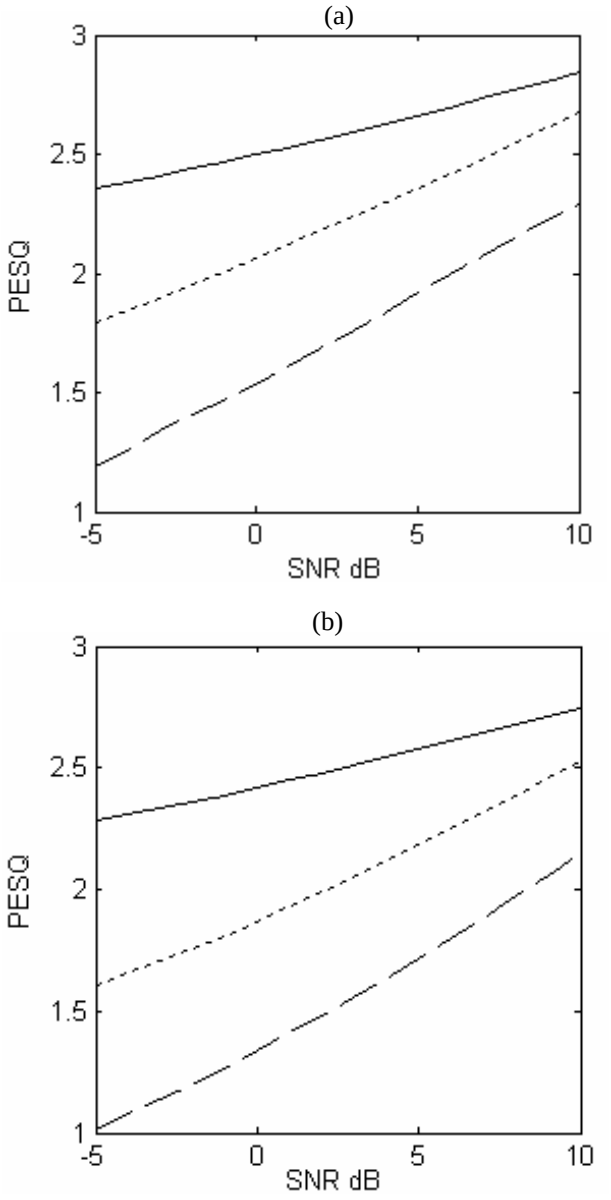

Fig. 3. PESQ scores obtained from noisy speech (dashed line) and enhanced speech using NSK (solid line) and FK (dotted line) algorithms. (a) German female speech contaminated by car noise (b) German female speech contaminated by airport noise.

In order to evaluate the performance of our method in the real system, speech enhancement with AR parameters estimated from the speech signal enhanced by spectrum subtraction algorithm [9] is also simulated in this paper. The results of this simulation are shown in Fig. 5. It reveals that our proposed method outperforms the conventional full band Kalman filtering method not only in the ideal case in which the AR parameters of speech signal is estimated from clean speech signal but also in the real case in which the AR parameters of speech signal is estimated from the enhanced speech signal.

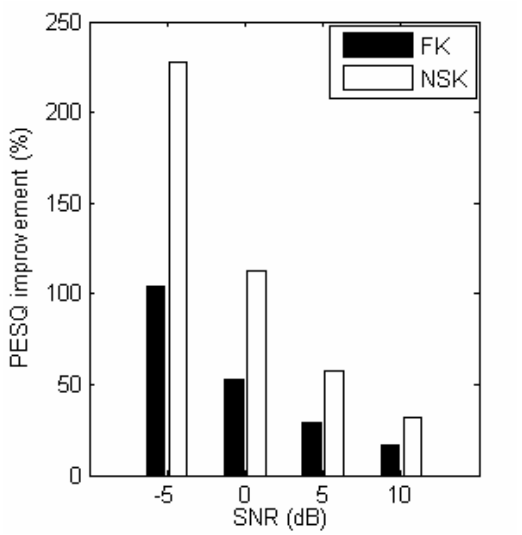

Fig. 4. Comparison of PESQ improvement δ (%) of German female speech file contaminated by street noise.

V. Conclusion and Furure Work

Rather than processing the whole frequency band of the speech signal, we have applied the Kalman filter algorithm to non-uniform sub-band speech signals obtained from the decomposition of full-band speech using gammatone filters. In this manner, a novel approach for speech enhancement has been presented. Simulation results have shown that the performance of our proposed method is better than that of the conventional Kalman filter method. The effectiveness of our approach could be further improved by exploiting characteristics of the human auditory system such as masking property [7], [8]. The masking threshold taking both temporal and simultaneous masking into account would be conveniently determined in the time domain and in individual sub-band signals [4], [10]. By extending our algorithm in that way, the performance of our proposed method is expected to be further improved and better than those of conventional Kalman algorithms not only full band but also uniform subband speech enhancement.

This paper is expected to be extended by fully incorporating the masking property of human auditory system into enhancement process. The performance of our proposed method then is compared to that of full band and uniform sub-band Kalman filtering speech enhancement.

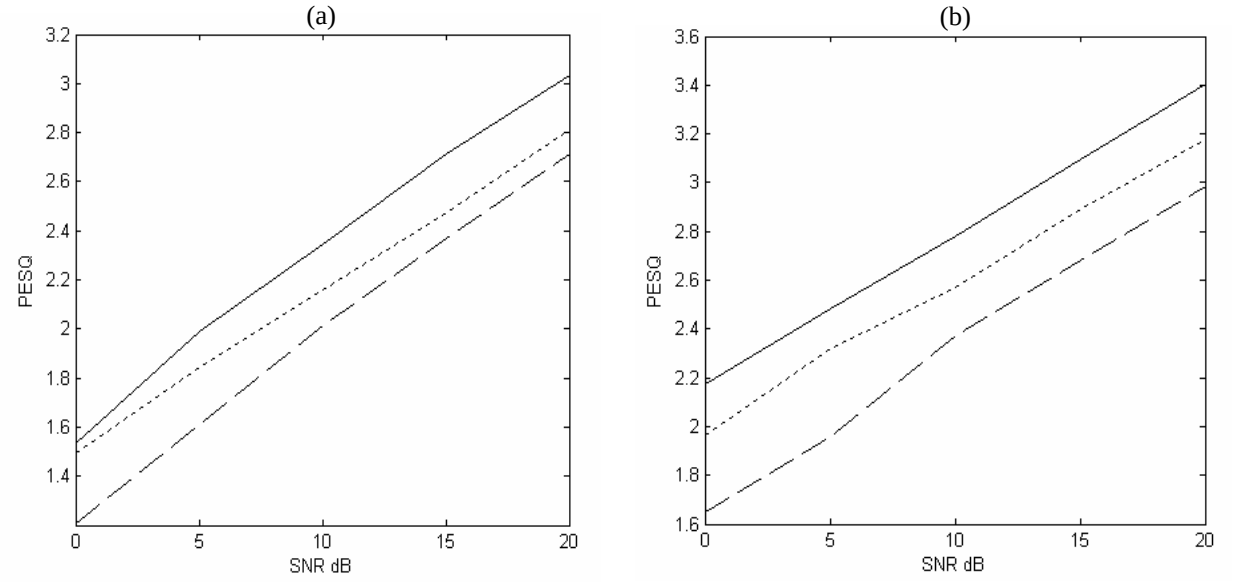

Fig. 5. PESQ scores obtained from noisy speech (dashed line) and enhanced speech using NSK (solid line) and FK (dotted line) algorithms. (a) German female speech contaminated by car noise (b) German male speech contaminated by F16 noise.

REFERENCES

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

[2] Chang Huai You, Soo Ngee Koh, Susanto Rahardja, "Kalman Filtering Speech Enhancement Incorporating Masking Properties For Mobile Communication In a Car Environment", *Proceedings of IEEE ICME 2004*, vol. 2, pp. 1343- 1346.

[3] Wen-Rong Wu, Po-Cheng Chen, "Subband Kalman Filtering for Speech Enhancement", *IEEE Trans. On Circuits and Systems-II: Analog and Digital Signal Processing*, vol. 45, No. 8, pp. 1072-1083, Agust 1998.

[4] Teddy Surya Gunawan, Eliathamby Ambikairajah. "Speech Enhancement using Temporal Masking and Fractional Bark Gammatone Filters", *Proceedings of the 10th Australian International Conference on Speech Science & Technology*, Macquarie University, Sydney, December 2004, pp. 420-425.

[5] Eliathamby Ambikairajah, Julien Epps, Lee Lin, "Wideband Speech And Audio Coding Using Gammatone Filter Banks", *Proceedings of IEEE ICASSP 2001*, vol. 2, pp. 773-776.

[6] Ning Ma, Martin Bouchard, and Rafik. A. Goubran, "Perceptual Kalman Filtering for Speech Enhancement in Colored Noise", *Proceedings of IEEE ICASSP 2004*, vol. 1, pp. I- 717-20.

[7] Brian C. J. Moore, "An Introduction to the Psychology of Hearing", *Academic Press 1997*, ISBN 0-12-505627-3.

[8] E. Zwicker, H. Fastl, "Psychoacoustics, Facts and Models", *Springer-Verlag 1990*, ISBN 3-540-52600-5.

[9] Berouti, M., Schwartz, M., and Makhoul, J. "Enhancement of speech corrupted by acoustic noise". *Proceedings of IEEE Int. Conf. Acoustic, Speech, Signal Processing* 1979, pp 208-211.

[10] James D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria", *IEEE Jounal on Selected Areas in Communications* vol. 6. No. 2, Feb. 1988, pp. 314-323.