Single Channel Speech Enhancement Using Subband Iterative Kalman Filter

Sujan Kumar Roy*, Wei-Ping Zhu* and Benoit Champagne†
*Department of Electrical and Computer Engineering
Concordia University, Montreal, Quebec, Canada H3G 1M8
Emails: su_ro@encs.concordia.ca, weiping@ece.concordia.ca

†Department of Electrical and Computer Engineering
McGill University, Montreal, Quebec, Canada H3A 0E9
Email: benoit.champagne@mcgill.ca

***Abstract*—In this paper, we propose a single channel speech enhancement algorithm using a subband iterative Kalman filter. A wavelet filterbank is first used to decompose the noise corrupted speech into a number of subbands. To achieve the best trade-off among the noise reduction, speech intelligibility and computational complexity, a partial reconstruction scheme based on consecutive mean squared error is proposed to synthesize the low-frequency (LF) and high-frequency (HF) subbands. An iterative Kalman filter is then applied to the partially reconstructed HF subband speech. Finally, the enhanced HF subband speech is combined with the partially reconstructed LF subband speech to reconstruct the fullband enhanced speech. Experimental results show that the proposed subband iterative Kalman filter based algorithm is capable of reducing adverse environmental noises for a wide range of input SNRs. The overall performance of our method in terms of segmental SNR, perceptual evaluation of speech quality (PESQ) and computational cost is superior to several existing Kalman filter based algorithms.**

***Keywords*—Speech enhancement, Kalman filter, wavelet filterbank, subband decomposition, partial reconstruction.**

I. Introduction

The main objective of speech enhancement (SE) is to eliminate or reduce disturbing noises from a degraded speech. SE has been widely used as a front end tool for speech recognition, telecommunication etc. Various SE methods have been introduced in the literature, including spectral subtraction (SS) [1], Wiener filter (WF) [2], and Kalman filter [3]-[6]. The SS[1] and WF [2] methods have been widely used due to their simplicity of implementation. However, these methods suffer from the so-called *musical* noise that is introduced in the enhanced speech.

Kalman filtering (KF) based on the minimum mean squared error (MMSE) criterion have been used in SE. A subband modulation KF based SE technique was proposed in [3], where the noisy speech is decomposed into a number of subbands followed by KF of each subband separately. However, the application of KF to each subband increases the computational complexity. Gibson *et al.* in [4] have proposed an iterative KF based SE method, in which the noise variance is estimated during silent periods, which implies that a voice activity detector is needed. In [5]-[6], iterative KF for SE using overlapped frames has been introduced. These methods however need access to the clean speech and the additive noise signals for parameter estimation.

In this paper, we propose a subband iterative KF method for SE. The noisy speech is decomposed first into a set of subbands using a wavelet filterbank. A consecutive mean squared error (CMSE) based synthesis method is proposed to undertake a partial reconstruction of the decomposed subbands into HF and LF subband speeches. Then, an iterative KF is applied to the partially reconstructed HF subband, while keeping the partially reconstructed LF subband unchanged, since it mainly contains the intelligible speech components. Finally, the enhanced speech of the HF subband produced by the iterative KF is combined with the LF subband to reconstruct the fullband enhanced speech.

II. Proposed Method

The noisy speech $y(n)$ captured by a single microphone can be written as

$$y(n) = s(n) + v(n) \tag{1}$$

where $s(n)$ and $v(n)$ represent the clean speech and the additive noise, respectively, at time n. The overall block-diagram for the proposed method is shown in Fig. 1, and the constituent modules are explained in the following subsections.

A. Wavelet Filterbank

A simple two-channel filterbank normally decomposes an input signal into two parts: low-frequency and high-frequency subbands. Each of the two subbands can be further divided by using the same two-channel filterbank. One can continue this two-band division for several levels to implement a wavelet packet tree decomposition, which provides more detailed analysis of a non-stationary signal. It is important to note that the wavelet packet coefficients at each subband can be reconstructed independently by using the wavelet packet reconstruction algorithm [7]. In addition, the length of the reconstructed subband signals in samples is equal to that of the given signal (at the same sampling rate). In the proposed speech enhancement algorithm (SEA), 16 reconstructed subbands, denoted as $y_i(n)$, $i = 1, 2, \ldots, 16$ are obtained using a 4-level wavelet packet tree decomposition. Note that the lowest subband index $i = 1$ denotes the highest frequency subband in this algorithm. Fig. 2 shows a noisy speech and the corresponding 16 reconstructed subbands, where the subband

978-1-4799-5341-7/16/$31.00 ©2016 IEEE

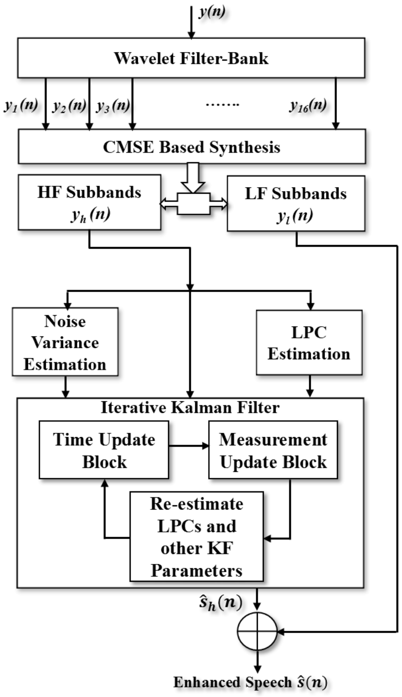

Fig. 1. Block-diagram of the proposed method.

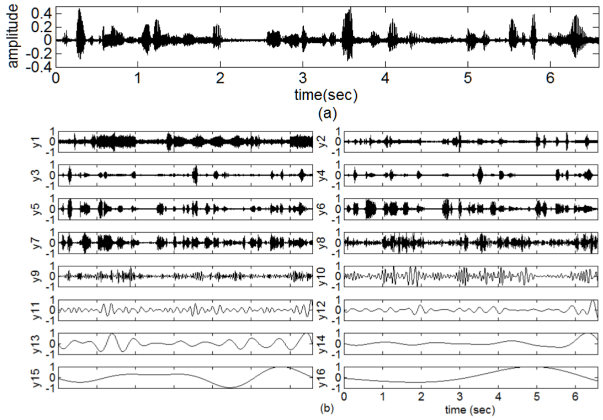

Fig. 2. (a) Speech (combined 2 NOIZEUS speech sentences) corrupted by babble noise (SNR=10dB), (b) the corresponding 16 reconstructed subbands.

signals are normalized with respect to their largest amplitudes. In this work, we aim to synthesize the 16 subbands into two major bands, HF and LF components. We will then apply KF to the HF band for noise reduction.

B. CMSE Based Synthesis

Here, we use the mean square error between two consecutive normalized subbands, called consecutive mean square error (CMSE), to decide what subbands are reconstructed into the HF band for Kalman filtering. For each subband of L samples, the CMSE is defined as

$$E_k = CMSE(y_k(n), y_{k+1}(n)) = \frac{1}{L}\sum_{n=1}^{L}(y_k(n) - y_{k+1}(n))^2 \tag{2}$$

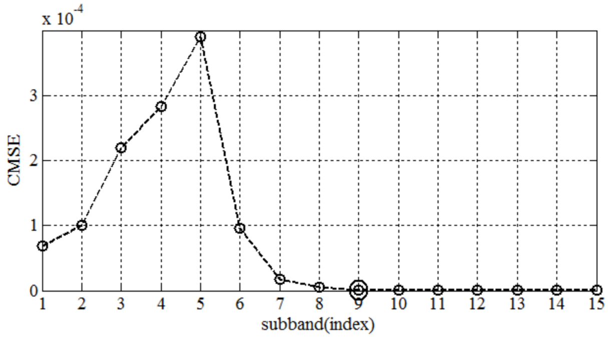

Fig. 3. The CMSE values corresponding to the normalized subbands in Fig. 2. The double circle indicates the j_s.

where $k = 1, 2, \ldots, 15$ is the subband index, N is the number of the subband speech samples. Our idea is to find $k = j_s$, the index of the last HF subband, such that no significant difference between two consecuative CMSE values, namely E_{j_s} and E_{j_s+1}, is observed. Specifically, we compute E_k and E_{k+1} for $k = 1, 2, \ldots, 15$ until their difference is sufficiently small. Then such a value of k is denoted as j_s. This empirical criterion is obtained from extensive experiments. Once the value of j_s is identified, the partially reconstructed HF and LF subband speech signals are respectively given by

$$y_h(n) = \sum_{i=1}^{j_s} y_i(n) \tag{3}$$

$$y_l(n) = \sum_{i=j_s+1}^{16} y_i(n) \tag{4}$$

Fig. 3 shows the CMSE values for the 16 reconstructed subbands of the noisy speech $y(n)$ shown in Fig. 2. From Figs. 2 and 3, it is clearly observed that for this particular speech segment, the 9^{th} subband is the last subband to be used for the partial reconstruction of the HF band. In general, the value of j_s depends on the input speech samples, the noise types, and the input SNR.

C. Proposed Subband Iterative Kalman Filter

The proposed subband iterative KF algorithm is applied to $y_h(n)$ while keeping $y_l(n)$ unchanged. It is operated on a frame-by-frame basis, including two loops, namely, the inner and the outer loop. For each frame of N samples, in the inner loop, the KF parameters are updated sample-by-sample through an iterative procedure. The additive noise components are reduced significantly when the inner loop is completed for one entire frame. Then the linear prediction coefficients (LPCs) and other state-space model (SSM) parameters are re-estimated from the processed speech for the next inner loop iteration. The outer loop iteration stops when the KF converges or the preset maximum number of iterations is exhausted, giving the further enhanced speech frame. This procedure will repeat for subsequent frames until all the noisy speech frames have been processed.

The SSM of the proposed subband iterative KF is represented by the following two equations, where the bold faced letters represent vectors or matrices

State Equation:

$$\boldsymbol{x}(n) = \boldsymbol{\Phi}\boldsymbol{x}(n-1) + \boldsymbol{H}^T u(n) \tag{5}$$

Observation Equation:

$$z(n) = \boldsymbol{H}\boldsymbol{x}(n) + v(n) \tag{6}$$

Here $\boldsymbol{x}(n)$ is a P-dimensional state parameter vector at time n which can be expressed in terms of the HF signal samples

$$\boldsymbol{x}(n) = [y_h(n-p+1) \quad y_h(n-p+2) \quad \ldots \quad y_h(n)]^T \tag{7}$$

In (5), $u(n)$ is called the process noise and $\boldsymbol{\Phi}$ is a $P \times P$-dimensional state transition matrix, which is given as

$$\boldsymbol{\Phi} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ a_p & a_{p-1} & a_{p-2} & \ldots & a_1 \end{bmatrix},$$

where a_i is the i^{th} LPC coefficient, P is the LPC order, and $\boldsymbol{H}$ is the $1 \times P$ observation row vector as given by

$$\boldsymbol{H} = [0 \quad 0 \quad 0 \quad \ldots \quad 1].$$

In (6), $z(n)$ is the noisy observation of the SSM and $v(n)$ is the measurement noise at time n. For each frame of N samples, we set D as the maximum number of iterations. The proposed iterative KF based SE is summarized below.

```
    Estimate LPCs $a_k, k = 1, 2, 3, \ldots, P$, from the subband noisy speech $z(n)$. Let $\hat{s}_h^{(0)} = z(n), n = 1, 2, 3, \ldots, N$.
For $j = 1 \quad to \quad D$ do [outer loop]

    Initialization:
        $\hat{\boldsymbol{x}}^{(j)}(0|0) = 0$    (8)
        ${\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{(j)}(0|0) = [0]_{p \times p}$    (9)
        $\boldsymbol{\Phi}^{(j)} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ a_p & a_{p-1} & a_{p-2} & \ldots & a_1 \end{bmatrix}$    (10)

    For $n = 1 \quad to \quad N$ do [inner loop]

        Time update (predictor):
            $\hat{\boldsymbol{x}}^{(j)}(n|n-1) = \boldsymbol{\Phi}^{(j)}\hat{\boldsymbol{x}}^{(j)}(n-1|n-1)$    (11)
            ${\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{(j)}(n|n-1) = \boldsymbol{\Phi}^{(j)}{\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{(j)}(n-1|n-1){\boldsymbol{\Phi}^{(j)}}^T + \boldsymbol{H}^T\sigma_u^2\boldsymbol{H}$    (12)

        Measurement update (corrector):
            $e^{(j)}(n) = \hat{s}_h^{(j-1)} - \boldsymbol{H}\hat{\boldsymbol{x}}^{(j)}(n|n-1)$    (13)
            $\boldsymbol{K}^{(j)}(n) = {\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{(j)}(n|n)\boldsymbol{H}^T(\boldsymbol{H}{\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{(j)}(n|n)\boldsymbol{H}^T + \sigma_v^2)^{-1}$    (14)
            $\hat{\boldsymbol{x}}^{(j)}(n|n) = \hat{\boldsymbol{x}}^{(j)}(n|n-1) + \boldsymbol{K}^{(j)}(n)e^{(j)}(n)$    (15)
            ${\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{(j)}(n|n) = (\boldsymbol{I} - \boldsymbol{K}^{(j)}(n)\boldsymbol{H}){\boldsymbol{\Sigma}_{\boldsymbol{x}}}^{(j)}(n|n-1)$    (16)

            Estimate enhanced speech (at time $n$):
            $\hat{s}_h^{(j)}(n) = \boldsymbol{H}\hat{\boldsymbol{x}}^{(j)}(n|n)$    (17)

    End for [inner loop]
    If $|1 - k_1^{(j)}||\hat{a}_P| < 1$ (where $k_1^{(j)}$ is the $1^{st}$ element of
        $\boldsymbol{K}^{(j)}(n)$) [KF Converges]
        Output the enhanced speech $\hat{s}_h(n)$ and stop.
End for [outer loop]
    Else
        Re-estimate LPCs from the $j^{th}$ processed frame $\hat{s}_h^{(j)}(n)$, giving a new set of $a_k$'s, $k = 1, 2, 3, \ldots, P$.
Repeat for [outer loop]
```

Other parameters required for KF are as follows

1) $e(n)$ is the measurement innovation or prediction.
2) $\boldsymbol{K}(n)$ is the Kalman gain function.
3) $\boldsymbol{\Sigma}_{\boldsymbol{x}}(n|n)$ is the error covariance matrix of the *a posteriori* estimate, $\hat{\boldsymbol{x}}(n|n)$.

The above procedure is repeated for the following frames until the last one being processed, resulting in the enhanced HF subband speech $\hat{s}_h(n)$. Finally, the enhanced fullband speech $\hat{s}(n)$ is obtained as

$$\hat{s}(n) = \hat{s}_h(n) + y_l(n) \tag{18}$$

D. Parameter Estimation

The LPC coefficients used in the subband iterative KF are updated based on the partially enhanced speech in each frame for a better accuracy. Fig. 4 shows the improved LPC power spectra (dashed), which can preserve the shapes of all the four formants as compared to the clean speech LPC power spectrum (solid). The noise variance σ_v^2 is estimated from $y_h(n)$ using a method proposed based on the finite difference approximation of Taylor series. The clean speech samples given in equation (1) can be well approximated locally at any point by a lower order polynomial, which can be thought of

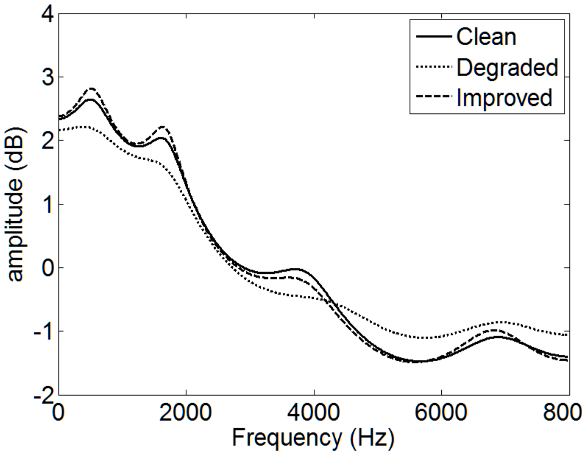

Fig. 4. Power spectra comparison between the clean speech (solid), degraded speech (dotted), and improved (dashed) for babble noise (SNR = 0dB).

TABLE I. DERIVATIVE TEMPLATES [8].

Template (w)	Differentiation Order
[-1 1]	First Derivative
[1 -2 1]	Second Derivative
[1 -3 3 -1]	Third Derivative
[1 -4 6 -4 1]	Forth Derivative

as a truncated local Taylor series approximation. Our idea is to apply a finite difference operation to the truncated series so that the lower order terms are eliminated, while leaving behind mainly the additive noise components, from which the noise variance is estimated. The differentiation can be represented mathematically as a convolution of the noisy observation with the derivative templates (see Table I) [8]. We apply the difference operation to $y_h(n)$, namely,

$$\hat{y}_h(n) = \frac{1}{M} \sum_{i=0}^{M-1} w[i] y_h[n-i] \tag{19}$$

where w is the derivative template in TABLE I and M is the length of w. Finally, σ_v^2 is estimated from $\hat{y}_h(n)$ using the sample variance formula,

$$\sigma_v^2 = \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_h(n) - \bar{\mu})^2 \tag{20}$$

where $\bar{\mu}$ is the sample mean of $\hat{y}_h(n)$ and N is the number of sample points in the analysis speech.

III. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the performance of the proposed algorithm, 30 speech utterances belonging to six speakers are taken from NOIZEUS speech corpus database [9]. The speech is sampled at 16 kHz and corrupted by white Gaussian, babble and car noises taken from the Noisex-92 database [10]. The LPC order is set to 8 and the wavelet *'sym13'* is used. The criteria used for performance evaluation are the perceptual evaluation of speech quality (PESQ) and segmental SNR (dB) [9]. The PESQ takes values between 1 (worst) and 4.5 (best). The performance of the proposed method (Proposed-SBIT-KF) is compared with existing methods, namely, LPCs enhancement in iterative KF (LPC-IT-KF) [6], and fast converging iterative KF (FC-IT-KF) [5]. From Fig. 5, it is observed that the proposed method performs much better than the two existing methods for all three noises in terms of PESQ. The segmental SNR results presented in Fig. 6 also show that the proposed method performs better than the existing methods.

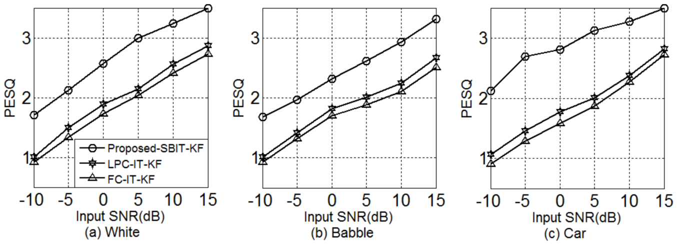

Fig. 5. Performance comparison between the proposed and existing methods in terms of PESQ in the presence of (a): White, (b): Babble and (c): Car noises for a wide range of input SNRs (-10dB to 15dB).

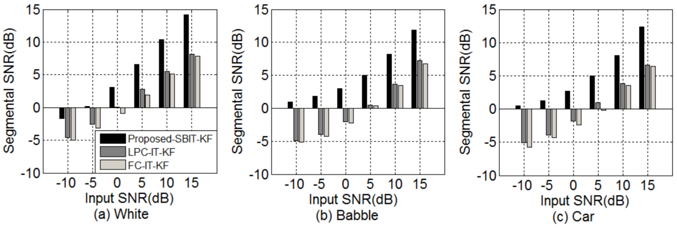

Fig. 6. Performance comparison between the proposed and existing methods in terms of segmental SNR (dB) in the presence of (a): White, (b): Babble and (c): Car noises for a wide range of input SNRs (-10dB to 15dB).

IV. CONCLUSIONS

In this paper, we have proposed an efficient single channel SE algorithm using subband iterative KF. A wavelet filterbank has been used to generate 16 reconstructed subbands of the noisy speech. An iterative KF has then been applied only to the HF subband speech that is obtained by using the CMSE synthesis method for noise reduction. The LPC coefficients used in KF were updated based on the partially enhanced speech in each frame for a better accuracy. By using a truncated Taylor series expansion of the partially reconstructed HF subband speech along with a difference operation serving as a high-pass filter, a method for the noise variance estimation was also proposed. The proposed method not only provides a superior SE performance but also reduces the computational complexity of conventional subband KF based methods. Through extensive simulation studies, we have found that the proposed method works effectively in adverse noise environments for a wide range of input SNRs, and outperforms several existing SE methods in literature.

REFERENCES

[1] K. Paliwal, K. Wojcicki, and B. Schwerin, "Single-channel Speech Enhancement using Spectral Subtraction in the Short-time Modulation Domain," *Speech Communication*, vol. 52, no. 5, pp. 450-475, May 2010.

[2] C. V. R. Rao, M. B. R. Murthy, and K. S. Rao, "Speech Enhancement using Sub-band Cross-correlation Compensated Wiener Filter Combined with Harmonic Regeneration," *Int. J. of Electronics and Communications*, 66(6):459-464, 2012.

[3] R. Ishaq, B. G. Zapirain, M. Shahid, and B. Lovstrom, "Subband Modulator Kalman filtering for Single Channel Speech Enhancement," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7442-7446, May 2013.

[4] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. on Signal Processing*, pp. 1732-1742, Aug 1991.

[5] S. So and K. K. Paliwal, "Fast Converging Iterative Kalman Filtering for Speech Enhancement using Long and Overlapped Tapered Windows with Large Side Lobe Attenuation," *INTERSPEECH*, pp. 1081-1084, Sep 2010.

[6] T. Mellahi and R. Hamdi, "LPCs Enhancement in Iterative Kalman Filtering for Speech Enhancement using Overlapped Frames," in *Proc. of Int. Conf. on Electrical Sciences and Technologies*, Nov 2014.

[7] J. Wan-lu, "Orthogonal Wavelet Packet Analysis Based Chaos Recognition Method," *Frontiers of Electrical and Electronic Engineering in China*, Volume 1, Issue 1, pp. 13-19, January 2006.

[8] T. O'Haver, "A Pragmatic Introduction to Signal Processing," available at https://terpconnect.umd.edu/ toh/spectrum/.

[9] P. C. Loizou, *Speech Enhancement: Theory and Practice*, Signal Processing and Communications, CRC Press, 2007.

[10] Noisex-92 database, "Speech at CMU," available at http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.