14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

SUBBAND PARTICLE FILTERING FOR SPEECH ENHANCEMENT

Ying Deng and V. John Mathews

Dept. of Electrical and Computer Eng., University of Utah
50 S. Central Campus Dr., Rm. 3280 MEB, Salt Lake City, UT 84112, USA
phone: + (1)(801) 581-6941, fax: + (1)(801) 581-5281, email: yd4@utah.edu, mathews@ece.utah.edu

ABSTRACT

Particle filters have recently been applied to speech enhancement when the input speech signal is modeled as a time-varying autoregressive process with stochastically evolving parameters. This type of modeling results in a nonlinear and conditionally Gaussian state-space system that is not amenable to analytical solutions. Prior work in this area involved signal processing in the fullband domain and assumed white Gaussian noise with known variance. This paper extends such ideas to subband domain particle filters and colored noise. Experimental results indicate that the subband particle filter achieves higher segmental SNR than the fullband algorithm and is effective in dealing with colored noise without increasing the computational complexity.

1. INTRODUCTION

Speech enhancement has been an active area of research during the past forty years. Speech enhancement algorithms available in the literature can be broadly divided into two categories - non-model based and model based algorithms. Representative approaches in the non-model based algorithms include spectral subtractive-type algorithms [1, 2] and signal subspace-based algorithms [3, 4]. Model based algorithms employ models of speech in the enhancement process. Autoregressive (AR) models are widely used to represent the vocal tract transfer function. Example of model based speech enhancement algorithms include the iterative Wiener filtering approaches [6, 7]. Kalman filtering based algorithms [8, 9, 10] form another class of model-based speech enhancement algorithms. Almost all such methods are based on autoregressive modeling of speech signals and linear Gaussian state-space representation of the system. Speech enhancement algorithms that assume specific probability distributions of speech signals and then derive minimum mean-square error estimates of the clean speech signals [11, 12] and those that assume composite source models (a composite source model is composed of a finite set of statistically independent subsources with each subsource representing a particular class of statistically similar speech sounds) such as Hidden Markov Models (HMMs) and use different estimators for different classes of speech signals [13, 14] also belong to the class of model based methods.

AR models of speech used in the iterative Wiener filtering and Kalman filtering approaches [6, 7, 8, 9, 10] assume that the articulatory shape of the vocal tract remains fixed throughout the analysis interval. However, in reality the vocal tract is changing continuously. To better model the non-stationarity of speech signals, time-varying autoregressive (TVAR) models of speech have been proposed [15]. In [16], a TVAR model with stochastically evolving parameters was adopted and shown to outperform standard AR models. By transforming between the AR coefficients and the reflection coefficients using standard Levinson recursion, Fong [17] used a time-varying partial correlation (TV-PARCOR) model and showed that the TV-PARCOR is a better physical representation of audio signals than the TVAR model. We adopt the TV-PARCOR model in our method. With the TVAR or TV-PARCOR modeling of speech signals, the system can be represented in a nonlinear conditionally Gaussian state-space form. Analytic solutions for recursive Bayesian state estimation exist only for a small number of specific cases [18]. For nonlinear conditionally Gaussian state-space models as those discussed in [16, 17] and also in this paper, the integrations used to compute the filtering distribution and the integrations employed to estimate the clean speech signal and model parameters do not have closed-form analytical solutions. Approximation methods have to be employed for these computations. The approximation methods developed so far can be grouped into three classes: (1) analytic approximations such as the Gaussian sum filter [19] and the extended Kalman filter [20], (2) numerical approximations which make the continuous integration variable discrete and then replace each integral by a summation [21], and (3) sampling approaches such as the unscented Kalman filter [22] which uses a small number of deterministically chosen samples and the particle filter [23] which uses a larger number of random (Monte Carlo simulation) samples for the computations. The analytic approximations are computationally simple but usually fail in complicated situations. The numerical approximations are only suited for low-dimensional state-spaces. In [16, 17], particle filters have been successfully employed for speech enhancement. The methods developed in [16, 17] were in the fullband domain and only white Gaussian noise with known variance was considered.

In this paper, we present a speech enhancement algorithm that employs particle filters in the subband domain. Typically, subband speech signals have flatter power spectrum as compared to the corresponding fullband signals. We can therefore use lower order TV-PARCOR models for the subband signals. This usually reduces the computational complexity of the algorithm. We show in this paper that while maintaining similar computational complexity, the subband modeling can model the speech power spectrum more accurately and result in better enhancement results. The enhanced fullband speech signals are obtained by synthesizing the enhanced subband speech signals. The particle filter based speech enhancement algorithms in [16, 17] assume white Gaussian noise with known variance, which is unrealistic in practical applications. This work extends the algorithm to solve the enhancement problem in colored noise environments. In order to accomplish this, we model the colored noise by an AR model and augment the state-space model for the white noise case. Experimental results show that the subband particle filter achieves higher segmental SNR improvement than the fullband scheme in white Gaussian noise without increasing the computational complexity. The subband particle filter is also effective in dealing with colored noise.

The rest of this paper is organized as follows. Section 2 describes the subband system model and the estimation objectives. In Section 3, we present the subband particle filter and the noise estimation algorithm. Section 4 provides experimental results. Finally, we make our concluding remarks in Section 5.

2. THE SUBBAND SYSTEM MODEL

In the fullband domain, the noisy speech $y(t)$ can be expressed as

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

where $s(t)$ and $n(t)$ are the clean speech signal and the additive background noise, respectively. The fullband signal is decomposed into a set of subband signals using an analysis filter bank and the subband signal can be written as

$$y_i(t) = s_i(t) + n_i(t), \tag{2}$$

where i is the subband index.

The component of the clean speech signal $s_i(t)$ in the ith subband is modeled as a p-th order TVAR process, *i.e.*,

$$s_i(t) = \sum_{k=1}^{p} a_{i,t}(k)s_i(t-k) + \sigma_{s_{i,t}} e_{s_i}(t). \quad (3)$$

Here, $\mathbf{a}_{i,t} = [a_{i,t}(1), \cdots, a_{i,t}(p)]^T$ is time-varying AR coefficients vector associated with the ith subband and $e_{s_i}(t)$ is a white Gaussian excitation with unit variance. The variance of the excitation is $\sigma^2_{s_{i,t}}$.

We assume that the colored noise statistics change sufficiently slowly so that they can be approximated as not changing during short time intervals. In such short time intervals, we model the ith subband component $n_i(t)$ of the colored noise as a q-th order AR process, *i.e.*,

$$n_i(t) = \sum_{k=1}^{q} b_i(k)s_i(t-k) + \sigma_{n_i} e_{n_i}(t). \quad (4)$$

Here, $\mathbf{b}_i = [b_i(1), \cdots, b_i(q)]^T$ is the i-th subband AR coefficients vector and $e_{n_i}(t)$ is a white Gaussian excitation with unit variance. The variance of the excitation is $\sigma^2_{n_i}$ and is not known *a priori*.

Given the clean speech and noise models in the subbands, we can develop a state-space system model in the following manner. Let us define $\mathbf{x}_i(t) = [s_i(t), \cdots, s_i(t-p+1), n_i(t), \cdots, n_i(t-q+1)]^T$, $\mathbf{e}_i(t) = [e_{s_i}(t) e_{n_i}(t)]^T$ and $\mathbf{y}_i(t) = [y_i(t)]^T$ and the system matrices

$$\mathbf{A}_{i,t} = \begin{bmatrix} \mathbf{A}^s_{i,t} & \mathbf{0}_{p\times q} \\ \mathbf{0}_{q\times p} & \mathbf{A}^n_{i,t} \end{bmatrix}_{(p+q)\times(p+q)} \quad (5)$$

where

$$\mathbf{A}^s_{i,t} = \begin{bmatrix} a_{i,t}(1) & a_{i,t}(2) & \cdots & a_{i,t}(p-1) & a_{i,t}(p) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{p\times p} \quad (6)$$

and

$$\mathbf{A}^n_{i,t} = \begin{bmatrix} b_i(1) & b_i(2) & \cdots & b_i(q-1) & b_i(q) \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{q\times q}. \quad (7)$$

Let

$$\mathbf{B}_{i,t} = \begin{bmatrix} \sigma_{s_{i,t}} & 0 \\ \mathbf{0}_{(p-1)\times 1} & \mathbf{0}_{(p-1)\times 1} \\ 0 & \sigma_{n_i} \\ \mathbf{0}_{(q-1)\times 1} & \mathbf{0}_{(q-1)\times 1} \end{bmatrix}_{(p+q)\times 2} \quad (8)$$

and

$$\mathbf{C}_{i,t} = [1\overbrace{0\cdots 0}^{p-1} 1\overbrace{0\cdots 0}^{q-1}]_{1\times(p+q)}. \quad (9)$$

Then, we can rewrite (2), (3) and (4) in state-space form as

$$\mathbf{x}_i(t) = \mathbf{A}_{i,t}\mathbf{x}_i(t-1) + \mathbf{B}_{i,t}\mathbf{e}_i(t) \quad (10)$$

$$\mathbf{y}_i(t) = \mathbf{C}_{i,t}\mathbf{x}_i(t). \quad (11)$$

In order for a sequential minimum mean square error (MMSE) estimation of the state vector, we have to know the distribution function $p(\mathbf{x}_i(t)|\mathbf{y}_i(1:t))$. This distribution can be obtained using the recursions

$$p(\mathbf{x}_i(t+1)|\mathbf{y}_i(1:t)) = \int p(\mathbf{x}_i(t)|\mathbf{y}_i(1:t))p(\mathbf{x}_i(t+1)|\mathbf{x}_i(t))d\mathbf{x}_i(t) \quad (12)$$

$$p(\mathbf{x}_i(t+1)|\mathbf{y}_i(1:t+1)) = \frac{p(\mathbf{y}_i(t+1)|\mathbf{x}_i(t+1))p(\mathbf{x}_i(t+1)|\mathbf{y}_i(1:t))}{p(\mathbf{y}_i(t+1)|\mathbf{y}_i(1:t))} \quad (13)$$

Once the distribution function is known, the MMSE estimate of the state vector is given by

$$\hat{\mathbf{x}}_i(t) = E\{\mathbf{x}_i(t)|\mathbf{y}_i(1:t)\} = \int \mathbf{x}_i(t)p(\mathbf{x}_i(t)|\mathbf{y}_i(1:t))d\mathbf{x}_i(t). \quad (14)$$

From the state-space representation (10) and (11), we can see that if the model parameters $\mathbf{a}_{i,t}$, $\sigma_{s_{i,t}}$, $\mathbf{b}_i$ and σ_{n_i} are known, the estimation problem can be solved using a Kalman filter. However, the parameters are unknown and have to be jointly estimated with the state vector $\mathbf{x}_i(t)$. This results in a conditionally Gaussian state-space system and has no closed form solution for the computation of the filtering distribution and the state estimation. Particle filter as an approximation method is then adopted to solve the estimation problem in this paper.

Let us define a speech parameter vector $\theta_i(t) = [a_{i,t}(1), \cdots, a_{i,t}(p), \log\sigma^2_{s_{i,t}}]^T$ that is to be estimated with the state vector $\mathbf{x}_i(t)$. The noise parameters will be estimated separately during intervals where speech is absent from the signal. To facilitate a particle filter solution, we further assume a TV-PARCOR model [17] for the time-varying AR coefficients of the speech signal. That is, the time-varying AR coefficients are first transformed to a set of time-varying reflection coefficients using Levinson recursion. The corresponding reflection coefficients is applied a constrained Gaussian random walk model. The constraints imposed are such that stability of the model is ensured. The constrained random walk model for the reflection coefficients is

$$p(\rho_{i,t}|\rho_{i,t-1}) = \begin{cases} N(\rho_{i,t-1}, \delta^2_\rho I); & \text{if } \max_k |\rho_{i,t}(k)| < 1 \\ 0; & \text{otherwise} \end{cases} \quad (15)$$

Here, $\rho_{i,t} = [\rho_{i,t}(1), \cdots, \rho_{i,t}(p)]^T$ is the set of reflection coefficients associated with the speech signal at time t. The logarithm of speech excitation variance also follows a Gaussian random walk model, *i.e.*, we assume that

$$p(\log\sigma^2_{s_{i,t}}|\log\sigma^2_{s_{i,t-1}}) = N(\log\sigma^2_{s_{i,t-1}}, \delta^2_{e_s}). \quad (16)$$

The estimation objectives then become the computation of the joint distribution $p(\mathbf{x}_i(t), \theta_i(t)|\mathbf{y}_i(1:t))$ and the MMSE estimates $E\{\mathbf{x}_i(t), \theta_i(t)|\mathbf{y}_i(1:t))\}$.

3. SUBBAND PARTICLE FILTERING AND NOISE ESTIMATION

The subband particle filter based speech enhancement algorithm is illustrated in Figure 1. The algorithm first decomposes the input signal into subband components, performs enhancement in the subband domain, and then reconstructs the enhanced fullband signal using a synthesis filter bank. In subsection 3.1, a sequential Monte Carlo method for estimating the state and speech parameter vector from the observed noisy signal is presented. In subsection 3.2, we will discuss the method used to estimate the noise parameters.

3.1 Subband particle filter

The subband particle filter adopted in this paper is the Rao-Blackwellized particle filter similar to those developed in [16, 17]. For a tutorial discussion of particle filtering, refer to [23]. We present the algorithm according for our state-space model in what follows.

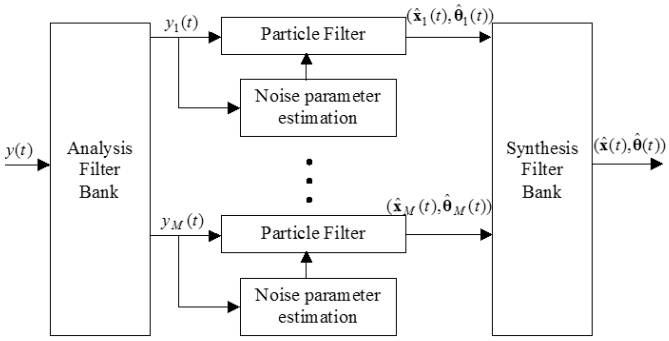

Figure 1: Subband speech enhancement system.

3.1.1 Sequential Bayesian importance sampling

Suppose that it is possible to sample N particles $\{\mathbf{x}_i^m(1:t),\theta_i^m(1:t);m=1,\ldots,N\}$ according to $p(\mathbf{x}_i(1:t),\theta_i(1:t)|\mathbf{y}_i(1:t))$. An empirical estimate of this distribution is

$$\overline{p}_N(\mathbf{x}_i(1:t),\theta_i(1:t)|\mathbf{y}_i(1:t))=\frac{1}{N}\sum_{m=1}^{N}\delta_{(\mathbf{x}_i^m(1:t),\theta_i^m(1:t))}, \quad (17)$$

where $\delta_{(\cdot)}$ is the Dirac delta function. Using this empirical distribution, the MMSE state and speech parameters estimates can be obtained as

$$\begin{aligned}&(\hat{\mathbf{x}}_i(1:t),\hat{\theta}_i(1:t))\\&=\int(\mathbf{x}_i(1:t),\theta_i(1:t))\overline{p}_N(d\mathbf{x}_i(1:t),d\theta_i(1:t)|\mathbf{y}_i(1:t))\\&=\frac{1}{N}\sum_{m=1}^{N}(\mathbf{x}_i^m(1:t),\theta_i^m(1:t))\end{aligned} \quad (18)$$

According to the strong law of large numbers, this estimate converges to the true estimate as N goes to infinity [23].

Unfortunately, $p(\mathbf{x}_i(1:t),\theta_i(1:t)|\mathbf{y}_i(1:t))$ is usually too complicated to sample directly. Instead, a simpler distribution $\pi(\mathbf{x}_i(1:t),\theta_i(1:t)|\mathbf{y}_i(1:t))$ which can be easily sampled from and whose support includes that of $p(\mathbf{x}_i(1:t),\theta_i(1:t)|\mathbf{y}_i(1:t))$ is employed. This method is called Bayesian importance sampling (BIS) [24]. An empirical estimate of $p(\mathbf{x}_i(1:t),\theta_i(1:t)|\mathbf{y}_i(1:t))$ using BIS is given by

$$\hat{p}_N(\mathbf{x}_i(1:t),\theta_i(1:t)|\mathbf{y}_i(1:t))=\sum_{m=1}^{N}\bar{\omega}_{1:t}^m\delta_{(\mathbf{x}_i^m(1:t),\theta_i^m(1:t))}, \quad (19)$$

where, the normalized importance weights $\bar{\omega}_{1:t}^m=\frac{\omega_{1:t}^m}{\sum_{m=1}^N\omega_{1:t}^m}$ and the importance weights $\omega_{1:t}^m\propto\frac{p(\mathbf{x}_i^m(1:t),\theta_i^m(1:t)|\mathbf{y}_i(1:t))}{\pi(\mathbf{x}_i^m(1:t),\theta_i^m(1:t)|\mathbf{y}_i(1:t))}$. With the BIS, the MMSE state and speech parameters estimates can be obtained as

$$(\hat{\mathbf{x}}_i(1:t),\hat{\theta}_i(1:t))=\sum_{m=1}^{N}\bar{\omega}_{1:t}^m(\mathbf{x}_i^m(1:t),\theta_i^m(1:t)). \quad (20)$$

In order to estimate $p(\mathbf{x}_i(1:t),\theta_i(1:t)|\mathbf{y}_i(1:t))$ at any time t without changing the past simulated trajectories $(\mathbf{x}_i^m(1:t-1),\theta_i^m(1:t-1)),m=1,\ldots,N$, we employ a sequential BIS scheme. The basic idea of sequential BIS is that $\pi(\mathbf{x}_i(1:t-1),\theta_i(1:t-1)|\mathbf{y}_i(1:t-1))$ is a factor of $\pi(\mathbf{x}_i(1:t),\theta_i(1:t)|\mathbf{y}_i(1:t))$, *i.e.*,

$$\begin{aligned}&\pi(\mathbf{x}_i(1:t),\theta_i(1:t)|\mathbf{y}_i(1:t))=\\&\quad\pi(\mathbf{x}_i(1:t-1),\theta_i(1:t-1)|\mathbf{y}_i(1:t-1))\times\\&\quad\pi(\mathbf{x}_i(t),\theta_i(t)|\mathbf{x}_i(1:t-1),\theta_i(1:t-1),\mathbf{y}_i(1:t)).\end{aligned} \quad (21)$$

The importance weights can also be recursively evaluated as

$$\begin{aligned}\omega_{1:t}^m&\propto\frac{p(\mathbf{x}_i^m(1:t-1),\theta_i^m(1:t-1)|\mathbf{y}_i(1:t-1))}{\pi(\mathbf{x}_i^m(1:t-1),\theta_i^m(1:t-1)|\mathbf{y}_i(1:t-1))}\times\\&\quad\frac{p(\mathbf{x}_i^m(t),\theta_i^m(t)|\mathbf{x}_i^m(1:t-1),\theta_i^m(1:t-1),\mathbf{y}_i(1:t))}{\pi(\mathbf{x}_i^m(t),\theta_i^m(t)|\mathbf{x}_i^m(1:t-1),\theta_i^m(1:t-1),\mathbf{y}_i(1:t))}\\&=\omega_{1:t-1}^m\frac{p(\mathbf{x}_i^m(t),\theta_i^m(t)|\mathbf{x}_i^m(1:t-1),\theta_i^m(1:t-1),\mathbf{y}_i(1:t))}{\pi(\mathbf{x}_i^m(t),\theta_i^m(t)|\mathbf{x}_i^m(1:t-1),\theta_i^m(1:t-1),\mathbf{y}_i(1:t))}\end{aligned} \quad (22)$$

and the normalized importance weights are $\bar{\omega}_{1:t}^m=\frac{\omega_{1:t}^m}{\sum_{m=1}^N\omega_{1:t}^m}$.

Thus, given an estimate of $p(\mathbf{x}_i(1:t-1),\theta_i(1:t-1)|\mathbf{y}_i(1:t-1))$, the estimate of $p(\mathbf{x}_i(1:t),\theta_i(1:t)|\mathbf{y}_i(1:t))$ is obtained by augmenting $(\mathbf{x}_i^m(1:t-1),\theta_i^m(1:t-1))$ with $(\mathbf{x}_i^m(t),\theta_i^m(t)),m=1,\ldots,N$ and recursively updating the importance weights according to (22). $(\mathbf{x}_i^m(t),\theta_i^m(t)),m=1,\ldots,N$ are sampled from $\pi(\mathbf{x}_i(t),\theta_i(t)|\mathbf{x}_i^m(1:t-1),\theta_i^m(1:t-1),\mathbf{y}_i(1:t))$. The marginal distribution $p(\mathbf{x}_i(t),\theta_i(t)|\mathbf{y}_i(1:t))$ is estimated as

$$\hat{p}_N(\mathbf{x}_i(t),\theta_i(t)|\mathbf{y}_i(1:t))=\sum_{m=1}^{N}\bar{\omega}_{1:t}^m\delta(\mathbf{x}_i^m(t),\theta_i^m(t)). \quad (23)$$

3.1.2 Rao-Blackwellization

Recall that $p(\mathbf{x}_i(t),\theta_i(1:t)|\mathbf{y}_i(1:t))=p(\mathbf{x}_i(t)|\theta_i(1:t),\mathbf{y}_i(1:t))p(\theta_i(1:t)|\mathbf{y}_i(1:t))$ and that $p(\mathbf{x}_i(t)|\theta_i(1:t),\mathbf{y}_i(1:t))$ is a Gaussian distribution that can be analytically evaluated using a Kalman filter. We assume that the noise parameters are already estimated and known. Then from the Rao-Blackwell theorem [25], we can reduce the estimation variance by only sampling $p(\theta_i(1:t)|\mathbf{y}_i(1:t))$ and analytically evaluating $p(\mathbf{x}_i(t)|\theta_i^m(1:t),\mathbf{y}_i(1:t))$ to obtain an estimate of $p(\mathbf{x}_i(t),\theta_i(1:t)|\mathbf{y}_i(1:t))$. We can summarize the Rao-Blackwellized particle filter as follows.

- Sample $\pi(\theta_i(t)|\theta_i(1:t-1),\mathbf{y}_i(1:t))$ for $\theta_i^m(t),m=1,\ldots,N$ and $\theta_i^m(1:t)=(\theta_i^m(1:t-1),\theta_i^m(t))$.
- For $m=1,\ldots,N$, evaluate the importance weights up to a normalizing constant $\omega_{1:t}^m\propto\omega_{1:t-1}^m\frac{p(\theta_i^m(t)|\theta_i^m(1:t-1),\mathbf{y}_i(1:t))}{\pi(\theta_i^m(t)|\theta_i^m(1:t-1),\mathbf{y}_i(1:t))}$
- $p(\theta_i(1:t)|\mathbf{y}_i(1:t))$ can then be approximated by $\hat{p}_N(\theta_i(1:t)|\mathbf{y}_i(1:t))=\sum_{m=1}^N\bar{\omega}_{1:t}^m\delta_{\theta_i^m(1:t)}$.
- The estimates of the speech parameters and the state vector can be expressed as

$$\hat{\theta}_i(t)=\sum_{m=1}^{N}\bar{\omega}_{1:t}^m\theta_i^m(t) \quad (24)$$

$$\hat{\mathbf{x}}_i(t)=\sum_{m=1}^{N}\bar{\omega}_{1:t}^mE\{\mathbf{x}_i(t)|\theta_i^m(1:t),\mathbf{y}_i(1:t)\}, \quad (25)$$

where, $E\{\mathbf{x}_i(t)|\theta_i^m(1:t),\mathbf{y}_i(1:t)\}$ can be computed using a Kalman filter. For details of Kalman filtering, please refer to [26].

3.1.3 Resampling

One problem with the sequential BIS is that after several time steps, many importance weights will have insignificant values. This will cause large estimation variances. In order to alleviate this problem, many resampling schemes have been proposed such as sampling importance resampling [27], residual resampling [28] and stratified resampling [29]. The generic stratified resampling scheme is adopted in this paper. For details of the algorithm, please refer to [29].

3.2 Noise parameter estimation

For noise parameter estimation, we first design a voice activity detector in each subband. Then we collect all the noise only segments

and construct a sequence of noise samples. We can then estimate the noise parameters using the Yule-Walker method from the noise only sequence.

The voice activity detector we adopt here is based on the minimum controlled recursive averaging noise spectrum estimation method [12]. We summarize the algorithm as follows. For each subband noisy signal $y_i(t)$, we first estimate the energy recursively as,

$$S_i(t) = \alpha_s S_i(t-1) + (1-\alpha_s) y_i^2(t), \tag{26}$$

where $0 < \alpha_s < 1$ is a forgetting factor. Then we track the minimum value of $S_i(t)$ denoted by $S_{i,min}(t)$. A samplewise comparison of the smoothed energy and the corresponding variable in the previous frame allows the following update for the minimum value

$$S_{i,min}(t) = \min\{S_{i,min}(t-1), S_i(t)\}. \tag{27}$$

Finally, we compute the ratio of the smoothed energy to its minimum value and compare the ratio with a threshold T. If the ratio is larger than the threshold, we consider it speech active. Otherwise, we consider it noise only. Once the noise only sequence is obtained, a standard Yule-Walker autoregressive parameter estimation algorithm [30] is applied to get the noise parameters.

4. EXPERIMENTAL RESULTS

The filter bank we used to obtain our experimental results was a nonuniform pseudo-QMF bank [31, 32] which achieves critical band division. The length of the prototype filter is 896 samples. A 1s long clean speech signal with sampling rate 16kHz was used. For noise parameter estimation, the forgetting factor α_s was chosen to be 0.8 and the threshold T was set to 5. For colored noise, we chose a first order AR model with $q = 1$ to represent the noise signal in each subband. We also employed a first order time-varying autoregressive model $p = 1$ for the speech signal in each subband. The importance distribution was chosen to be the prior distribution, *i.e.*, $\pi(\theta_i(t)|\theta_i(1:t-1), \mathbf{y}_i(1:t)) = p(\theta_i(t)|\theta_i(t-1))$. With the Gaussian random walk models defined in (15) and (16), it is easy to sample this prior distribution. The number of particles N was selected as 100 in all the experiments. The variances of the Gaussian random walk models in (15) and (16) were set to $\delta_\rho^2 = 0.001$ and $\delta_{e_s}^2 = 0.01$.

For the first set of experiments, we compare the algorithm of this paper with the fullband Rao-Blackwellized particle filter [17] in white Gaussian noise to show that our subband particle filter algorithm achieves higer segmental SNR improvement and at the same time takes similar CPU time per iteration as compared to the fullband particle filter algorithm. The segmental SNR is a widely used objective measure for speech enhancement systems and is defined as

$$\text{SegSNR} = \frac{1}{M}\sum_{m=0}^{M-1} 10\log_{10}\left(\frac{\sum_{n=0}^{L-1}|s(n+mL)|^2}{\sum_{n=0}^{L-1}|s(n+mL)-\hat{s}(n+mL)|^2}\right), \tag{28}$$

where $s(n)$ and $\hat{s}(n)$ denote the clean speech and the enhanced speech, respectively. Here, M is the number of frames in the speech segment and L is the number of samples per frame. The segmental SNR improvement was estimated by subtracting the SegSNR of the enhanced speech from the SegSNR associated with the noisy speech. White Gaussian noise was added to a clean speech signal at different segmental SNRs. Without optimization, the CPU times were $0.9080s$ per iteration for the subband algorithm and $0.8982s$ per iteration for the fullband algorithm on a standard 1.4 GHz PC. Table 1 shows the comparison of the SegSNR improvements in this example. Figure 2 shows the comparison of the estimated clean speech signal for these two approaches at an input SegSNR of 5 dB. From Table 1 and Figure 2, we can see that the subband domain speech enhancement algorithm exhibits much smaller estimation variance and thus much higher segSNR improvement. This is because while maintaining similar computation complexity, the subband modeling can model the speech power spectrum more accurately and result in better enhancement results.

Input SegSNR(dB)	SegSNR improvement(dB)	
	Fullband	Subband
-5	6.90	12.37
0	5.15	10.25
5	3.42	9.61
10	2.44	7.52

Table 1: Comparison of SegSNR improvement in white Gaussian Noise.

Input SegSNR(dB)	SegSNR improvement(dB)			
	Cohen [33]		Subband PF	
	Traffic	F-16	Traffic	F-16
-5	7.64	5.98	7.82	8.12
0	5.67	4.07	5.30	6.77
5	3.30	2.59	3.14	4.75
10	1.23	0.83	1.11	2.89

Table 2: Comparison of SegSNR improvement in colored noise.

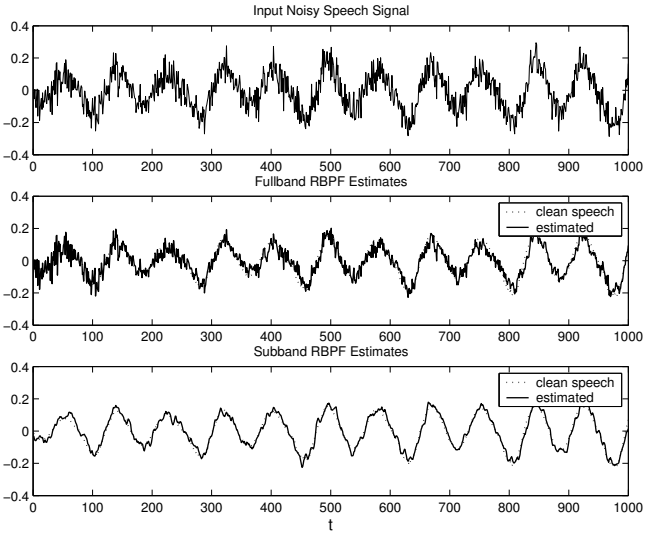

Figure 2: Comparison of the estimated clean speech at input SegSNR of 5 dB.

For the second set of experiments, we compare the SegSNR improvement of our algorithm to the two-state modeling algorithm [33] when the clean speech signal is corrupted by colored noise. Two colored noise signals - traffic and F-16 noise - were added to a clean speech signal at different Segmental SNRs. Table 2 shows the experimental results. From Table 2, we can see that the subband particle filter performs $2-3$dB better for the F-16 noise case than the two-state modeling algorithm. The two-state modeling algorithm performs slightly better or similar to the particle filter for the traffic noise case. However, the performance difference is less than 0.4dB in all cases tabulated in Table 2. Informal listening tests have also shown that the subband particle filter method exhibits lower residual noise than the two-state modeling approach [33].

5. CONCLUSIONS

This paper presented a subband domain particle filter based speech enhancement system. We have shown through experiments that the subband domain particle filter performs better in terms of segmental SNR as compared to the corresponding fullband domain algorithm. The algorithm is able to deal with colored noise, whereas only white Gaussian noise with known variance was considered in previous application of particle filters to speech enhancement [16, 17]. We compared our speech enhancement results in colored noise with a two-state modeling algorithm [33] and demonstrated that our method is effective in dealing with colored noise. We have assumed that the colored noise is stationary in our paper. A noise parameter estimation that can track the non-stationarity of the background noise is under development.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113-120, Apr. 1979.

[2] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 126-137, Mar. 1999.

[3] Y. Ephraim and H.L. Van Trees, "Signal subspace approach for speech enhancement," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 251-266, Jul. 1995.

[4] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 334-341, Jul. 2003.

[5] J. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, pp. 1586-1604, Dec. 1979.

[6] J. Lim and A.V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 197-210, Jun. 1978.

[7] J.H.L. Hansen and M. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795-805, Apr. 1991.

[8] K.K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," *Proc. ICASSP 1987*, Dallas, USA, April 6-9. 1987, pp. 177-180.

[9] S. Gannot, D. Burshtein and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 373-385, Jul. 1998.

[10] K.Y. Lee and S. Jung, "Time-domain approach using multiple Kalman filters and EM algorithm to speech enhancement with nonstationary noise," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 282-291, May 2000.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109-1121, Dec. 1984.

[12] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, pp. 2403-2418, Nov. 2001.

[13] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, pp. 1526-1555, Oct. 1992.

[14] H. Sameti, H. Sheikhzadeh, L. Deng and R.L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 445-455, Sept. 1998.

[15] M.G. Hall, A.V. Oppenheim and A.S. Willsky, "Time-varying parametric modeling of speech," *Signal Processing*, vol. 5, pp. 267-285, May 1983.

[16] J. Vermaak, C. Andrieu, A. Doucet and S.J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech singlas," *IEEE Trans. Speech and Audio Processing*, vol. 10, pp. 173-185, Mar. 2002.

[17] W. Fong, S.J. Godsill, A. Doucet and M. West, "Monte Carlo smoothing with application to audio signal enhancement," *IEEE Trans. Signal Processing*, vol. 50, pp. 438-449, Feb. 2002.

[18] B. Ristic, S. Arulampalam and N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Norwood, MA: Artech House, 2004.

[19] H.W. Sorenson and D.L. Alspach, "Recursive Bayesian estimation using Gaussian sums," *Automatica*, vol. 7, pp. 465-479, 1971.

[20] B.D.O. Anderson and J.B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

[21] S.C. Kramer and H.W. Sorenson, "Recursive Bayesian estimation using piece-wise constant approximations," *Automatica*, vol. 24, pp. 789-801, 1988.

[22] S. Julier, J. Uhlmann and H.F. Durrant-White, "A new method for nonlinear transformation of means and covariance in filters and estimators," *IEEE Trans. Automatic Control*, vol. 45, pp. 477-482, Mar. 2000.

[23] A. Doucet, N. de Freitas and N.J. Gordon, *Sequential Monte Carlo Methods in Practice*. New York: Springer, 2001.

[24] J.M. Bernardo and A.F.M. Smith, *Bayesian Theory*. New York: John Wiley & Sons, 1994.

[25] E.L. Lehmann and G. Casella, *Theory of Point Estimation*. New York: Springer-Verlag, 1998.

[26] C.K. Chui and G. Chen, *Kalman Filtering with Real-Time Applications*. New York: Springer-Verlag, 1999.

[27] N.J. Gordon, D.J. Salmond and A.F.M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," *IEE Proc. F, Radar Signal Process.*, vol. 140, pp. 107-113, Apr. 1993.

[28] J.S. Liu and R. Chen, "Sequential Monte Carlo methods for dynamic systems," *J. Amer. Statist. Assoc.*, vol. 93, no. 443, pp. 1032-1044, 1998.

[29] G. Kitagawa, "Monte Carlo filter and smoother for non-Gaussian nonlinear state space models," *J. Comput. Graph. Statist.*, vol. 5, no. 1, pp. 1-25, 1996.

[30] J.G. Proakis and D.G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications*. Upper Saddle River, New Jersy: Prentice Hall, 1996.

[31] J. Li, T.Q. Nguyen and S. Tantaratana, "A simple design method for near-perfect-reconstruction nonuniform filter banks," *IEEE Trans. Signal Process.*, vol. 45, pp. 2105-2109, Aug. 1997.

[32] T.Q. Nguyen, "Near-Perfect-Reconstruction Pseudo-QMF Banks," *IEEE Trans. Signal Process.*, vol. 42, pp. 65-76, Jan. 1994.

[33] I. Cohen, "Enhancement of speech using bark-scaled wavelet packet decomposition," *Proc. 7th European Conf. Speech, Communication and Technology*,Sept. 2001, pp. 1933-1936.