SPEECH ENHANCEMENT USING A SWITCHING KALMAN FILTER WITH A PERCEPTUAL POST-FILTER

Jianping Deng, Martin Bouchard, and Tet Yeap

School of Information Technology and Engineering, University of Ottawa, 800 King Edward, Ottawa (Ontario), K1N 6N5, Canada {jdeng, bouchard, tet}@site.uottawa.ca

ABSTRACT

In this paper, a Switching Kalman Filter (SKF) with a Generalized Pseudo Bayesian (GPB) algorithm of order 1 is applied to the problem of speech enhancement. It is proposed to use the masking properties of human auditory systems as a perceptual post-filter concatenated with the GPB algorithm. Experiments show that the proposed algorithm can achieve an improvement both in terms of speech quality (PESQ score, ITU-T P.862) and of word recognition rate at low SNR.

1. INTRODUCTION

Hidden Markov Models (HMM), which are used widely in speech recognition, have also been proposed for speech enhancement tasks [1]. In these methods, the speech time signal is split up into a number of overlapping blocks; assuming that signals in each block are stationary. An autoregressive (AR) model is then applied. A problem with this approach, however, is that the framing often results in a relatively poor temporal resolution for fast varying speech sounds such as plosives [2]. Alternatively, a hidden filter model or the autoregressive Hidden Markov Model (AR-HMM) have been suggested in [2],[3]. By embedding AR models into dynamic Markov chains, AR-HMMs were shown to be suitable for non-stationary signal analysis. The Kalman filtering problem for such models includes a weighted sum of filters operating interactively in parallel, also known as the Switching Kalman Filters (SKF)[4],[5],[6],[7].

An important special problem with SKF is that the optimal minimum mean squared error estimator involves a bank of filters tuned to all possible parameter histories, which makes the cost in computation grow exponentially with data length. To solve this problem, several sub-optimal estimation algorithms have been proposed, including the Generalized Pseudo-Bayesian (GPB) algorithm [6] and the Interacting Multiple Model (IMM) algorithm [7]. They have been used mostly in the target-tracking problems. The IMM algorithm has recently been applied to the AR-HMM model for speech enhancement in [5] and compared with a Separate Multiple Model (SMM) algorithm suggested in [8].

In this paper, we investigate the use of a Switching Kalman Filter with a GPB of order 1 (GPB1) algorithm incorporating the masking properties of human auditory systems in a perceptual post-filter as in [9], to enhance noisy speech. Simulations show that an improved performance in speech quality or word recognition rate in low SNR can be obtained with the proposed algorithm compared with [5]. The rest of this paper is organized as follows: Section 2 describes how to model a clean speech signal by an AR-HMM model. The GPB1 algorithm with a perceptual post-filter is introduced in Section 3. Section 4 describes the experiments and presents the results. Conclusions are given in Section 5.

2. SIGNAL AND NOISE MODEL

Consider a Markov chain with N discrete hidden states, $s_t \in (1,\ldots,N)$ and a state transition matrix $P(s_t = j | s_{t-1} = i) = Z(i,j)$. For time t, the speech signal $x(t)$ conditioned on state $s_t = i$ is described as:

$$x(t) = A_i X_{t-1} + e_i(t) \quad (1)$$

where $A_i = [a_i(1),...,a_i(p)]$ is the vector of p AR coefficients on state i, $X_{t-1} = [x(t-1),...,x(t-p)]^T$, and $e_i(t)$ are Gaussian noise processes with variances Q_i.

The parameters of the AR-HMM can be estimated using an Expectation-Maximisation (EM) algorithm. The details of the process can be found in [4]. For l multiple training sequences, the result is given here:

$$A_i = \left(\sum_{\ell} \sum_{t} \gamma_t^i X_{t-1} X_{t-1}^T \right)^{-1} \left(\sum_{\ell} \sum_{t} \gamma_t^i x(t) X_{t-1} \right) \quad (2)$$

$$Q_i = \left(\frac{1}{\sum_{\ell} \sum_{t} \gamma_t^i} \right) \left(\sum_{\ell} \sum_{t} \gamma_t^i \left(x(t) - A_i X_{t-1} \right)^2 \right) \quad (3)$$

$$Z(i,j) = \frac{\sum_{\ell} \sum_{t} \Pr\left(s_{t-1} = i, s_t = j | y_{1:t} \right)}{\sum_{\ell} \sum_{t} \gamma_t^i} \quad (4)$$

where $\gamma_t^i = \Pr(s_t = i \mid x_{1:t})$ is the probability that the HMM is in state i at time t and is calculated by the forward-backward algorithm [10], and y_t represents the noisy speech. When the measurement noise w_t is assumed additive white Gaussian, the state space model conditioned on s_t can be written as:

$$X_t = F_{s_t} X_{t-1} + G\, e_{s_t} \tag{5}$$

$$y_t = H\ X_t + w_t \tag{6}$$

$$F_{s_t} = \begin{bmatrix} A_{s_t} \\ I_{p-1} \quad 0_{p-1} \end{bmatrix} \tag{7}$$

$$H = G^T = [1 \quad 0 \quad \cdots \quad 0 \quad 0] \tag{8}$$

where I_{p-1} is an identity matrix of size $p-1$ and 0_{p-1} is a column of $p-1$ zeros. The variance R of w_t can be estimated from speech-free segments. In our experiments, it is estimated from the first 600 speech samples which are assumed to be noise.

3. PSEUDO BAYESIAN ALGORITHM WITH A PERCEPTUAL POST-FILTER

Some notations are first defined:

$$X_{t|t}^j = \mathrm{E}[X_t \mid y_{1:t}, s_t = j] \tag{9}$$

$$V_{t|t}^j = \mathrm{cov}[X_t \mid y_{1:t}, s_t = j] \tag{10}$$

$$L_t^j = \Pr(y_t \mid y_{1:t-1}, s_t = j) \tag{11}.$$

Given a noisy speech, each estimate X_t is found from a modified Switching Kalman Filter with the following steps performed in sequence:

$$\left(X_{t|t}^j, V_{t|t}^j, L_t^j\right) = Filter\left(X_{t-1|t-1}, V_{t-1|t-1}, y_t; F_j, G, H, Q_j, R\right) \tag{12}$$

$$M_{t|t}^{\ j} = \Pr\left(s_t = j \mid y_{1:t}\right) = \frac{L_t^j * \sum_i Z(i,j) M_{t-1|t-1}^{\ i}}{\sum_j L_t^j * \sum_i Z(i,j) M_{t-1|t-1}^{\ i}} \tag{13}$$

$$\left(X_{t|t}, V_{t|t}\right) = Collapse\left(X_{t|t}^j, V_{t|t}^j, M_{t|t}^j\right) \tag{14}$$

The definition of *Filter* is as follows:

$$X_{t|t-1}^j = F_j X_{t-1|t-1} \tag{15}$$

$$V_{t|t-1}^j = F_j V_{t-1|t-1} F_j^{\ T} + G Q_j G^T \tag{16}$$

$$e_t^j = y_t - H X_{t|t-1}^j \tag{17}$$

$$q_t^j = H V_{t|t-1}^j H^T + R \tag{18}$$

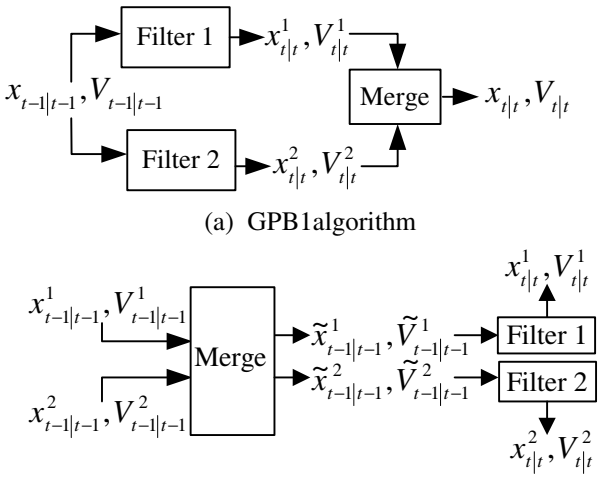

Fig 1. Comparison of GPB1 and IMM algorithms

$$K_t^j = V_{t|t-1}^j H^T (q_t^j)^{-1} \tag{19}$$

$$L_t^j = N\left(e_t^j; 0, q_t^j\right) \tag{20}$$

$$X_{t|t}^j = X_{t|t-1}^j + K_t^j e_t^j \tag{21}$$

$$V_{t|t}^j = \left(I - K_t^j H\right) V_{t|t-1}^j = V_{t|t-1}^j - K_t^j q_t^j K_t^{j\,T} \tag{22}$$

Collapse is a moment matching function, defined as:

$$X_{t|t} = \sum_j M_{t|t}^j X_{t|t}^j \tag{23}$$

$$V_{t|t} = \sum_j M_{t|t}^j V_{t|t}^j + \sum_j M_{t|t}^j \left(X_{t|t}^j - X_{t|t}\right)\left(X_{t|t}^j - X_{t|t}\right)^T \tag{24}$$

The enhanced speech signal $\hat{x}(t)$ (before post-filtering) is equal to the first component of the estimated $X_{t|t}$. It is the first-order Generalized Pseudo-Bayesian (GPB1) estimator [6], which limits the memory of the model history by combining the estimates from all models into a single estimate at the end of each processing cycle. If there are 2 hidden states in the AR-HMM model, Fig. 1 shows the comparison of the GPB1 and IMM filtering algorithms [4]. The collapsed output of the switching Kalman filter is post-filtered on a frame by frame basis with a frame length of 128 samples, using the masking properties of human auditory systems. The following procedure is taken [9]:

(1) Computing a 256-point FFT from a $\hat{x}(t)$ sequence to get the speech spectrum $\hat{S}(\omega_i)$.

(2) Adding up the energies of the FFT values in each critical bank [9], then convolving with the following spread function to get the spread critical band spectrum $C(k)$:

$$SF(k) = 15.81 + 7.5(k + 0.474) - 17.5\sqrt{1 + (k + 0.474)^2} \quad (25)$$

(3) Calculating the relative threshold offset $O(k)$ in decibels as:

$$O(k) = \alpha(14.5 + k) + (1 - \alpha)5.5 \quad (26)$$

where α is defined as:

$$\alpha = \min\left(\frac{SFM_{dB}}{SFM_{dB\max}}, 1\right) \quad (27)$$

$$SFM_{dB\max} = -60dB \text{ and } SFM_{dB} = 10\log_{10}\frac{G_m}{A_m}$$

(G_m is the geometric mean of the power spectrum, and A_m is the arithmetic mean of the power spectrum).

(4) Calculating the masking threshold $M(k)$ by subtracting the threshold offset $O(k)$ from the spread critical band spectrum $C(k)$:

$$M(k) = 10^{(\log_{10}[C(k)] - [(O(k)/10])} \quad (28)$$

(5) Mapping the total masking level $M_t(k)$ $(k = 1,2,\ldots,18)$ in each critical band to the frequency domain (FFT bins), to obtain $T(\omega_i)$ $(i = 1,2,\cdots 256)$.

(6) Performing the thresholding on half of the speech spectrum $\hat{S}(\omega_i)$: (29),(30)

if $\omega_i < 64$

$$\left|\tilde{\hat{S}}(\omega_i)\right| = \begin{cases} \left|\hat{S}(\omega_i)\right| \times \alpha^2 & \text{if } \left|\hat{S}(\omega_i)\right|^2 < T(\omega_i) \\ \left|\hat{S}(\omega_i)\right| & \text{otherwise} \end{cases}$$

if $\omega_i > 64$

$$\left|\tilde{\hat{S}}(\omega_i)\right| = \begin{cases} \left|\hat{S}(\omega_i)\right| \times \alpha^3 & \text{if } \left|\hat{S}(\omega_i)\right|^2 < T(\omega_i) \\ \left|\hat{S}(\omega_i)\right| \times \alpha^2 & \text{otherwise} \end{cases}$$

where $i = 1,2,\cdots 128$, and α is the tonality coefficient computed with the simultaneous masking threshold. The other half of the masked speech spectrum is obtained by symmetry.

(7) Doing an IFFT using $\left|\tilde{\hat{S}}(\omega_i)\right|$ and the phase of $\hat{S}(\omega_i)$, keeping the last 128 values of the size-256 IFFT outputs to obtain the enhanced speech signal.

A new threshold setting different from [9] is used in step (6). The tradeoff between noise reduction and speech distortion has been modified to obtain a better performance in speech recognition (i.e. in this work both speech quality and recognition rate are to be improved).

5. EXPERIMENTS

This section presents the performance evaluation of the proposed speech enhancement method. First, the performance is measured in terms of speech quality with the ITU-T P.862 PESQ score [11], which has a close match with subjective speech quality scores. Then the enhancement method is tested as a front-end to compensate the noise for robust speech recognition.

The speech data is from the "Numbers v1.3" corpus provided by Oregon Health & Science University (OGI). The corpus is a collection of 8kHz telephone speech data, including both isolated digit and continuous digit strings [12]. In our experiment, speech files with fixed-length 5 connected digits were used. In the testing stage, background noises (from the ITU-T Supplement P.23 database) were artificially added to the speech signals by a computer, with SNR varying from 0 dB to 10dB.

All the experiments use the same AR-HMM speech model which is built by 20 training strings. The order of the AR model is 10, and the number of states for the AR-HMM is five. HMMs trained using the EM algorithm are guaranteed to reach only a local maximum likelihood solution and are sensitive to the initialization. In our experiment, the AR-HMM model is initialized using a procedure of Kalman filter AR model described in [3]. Unlike in [5], in our experiments the training speech and speakers will not appear in the enhancement tests. Tables 1 and 2 first show the average PESQ score of the proposed method for the 20 speech files in the test set, and compare the proposed algorithm with the GPB1 and the IMM [5]. There is not much difference between the PESQ scores for the IMM algorithm and the GPB1 algorithm. But with the perceptual post-filter, the GPB1 algorithm can produce PESQ scores of 0.1 to 0.2 better than the IMM algorithm.

Secondly, the proposed enhancement algorithm is applied as a pre-processor to an automatic speech recognition (ASR) system. For the connected digits recognition, each phoneme is represented by the left-to-right monophone HMM containing 5 states (3 observation states, an entry and an exit state) and 4 mixtures for each

Input SNR	10dB	5dB	0dB
SKF (IMM)	2.59	2.25	1.89
SKF (GPB1)	2.60	2.27	1.89
Proposed	2.66	2.41	2.11

Table 1 Results for speech with white noise (PESQ scores)

Input SNR	10dB	5dB	0dB
SKF (IMM)	2.63	2.33	2.05
SKF (GPB1)	2.62	2.31	2.03
Proposed	2.67	2.40	2.10

Table 2 Results for speech with street noise (PESQ scores)

state. They were trained using 550 digit strings spoken by different people and then evaluated on an independent test set. Therefore, it is the kind of system which is trained in a clean environment and tested in noisy ones. The acoustic features are 39 Perceptual Linear Prediction (PLP) coefficients, with 12 PLP cepstrum, log energy and their first and second order derivatives. Since the speech data is collected from both analog and digital phone lines, cepstrum mean normalization is applied to remove the input channel difference. Tables 3 and 4 show the average word recognition rate in 8 training iterations for white noise and street noise. The proposed algorithm is compared to the GPB1, the IMM, the conventional spectral subtraction (SS) and the Kalman filter algorithm [13]. 'No NR' denotes a result without noise reduction.

The results show that all the Switching Kalman Filtering algorithms produced a performance advantage over the Kalman filter and the spectral substraction methods. For 0dB SNR, the proposed algorithm performs better than all the other methods. For 5 dB SNR, it also produced results as good or better than the GPB1 and the IMM algorithms. For 10 dB SNR, and especially for the case of white noise, the proposed algorithm degrades slightly the performance compared to the GPB1 and the IMM. The explanation could be that speech recognizers are more sensitive to speech distortion compared to the way the signal is perceived by human listeners.

Normally, speech recognition produces better results if noise is added to the templates rather than if it is subtracted from the data [14]. But at low SNR, it is more advantageous to enhance speech in order to improve the recognition rate. In this case, our enhancement method would work well as a pre-processor.

6. CONCLUSIONS

Given the AR-HMM model trained from clean speech, the recursive Switching Kalman Filters provide a better performance than conventional Kalman filters. To further improve the performance of the SKF using a GPB algorithm, this paper proposed a GPB algorithm with a perceptual post-filter. Experiments have shown that the proposed algorithm can achieve an improvement both in terms of speech quality (PESQ score, ITU-T P.862) and of word recognition rate at low SNR.

7. REFERENCES

[1] Y. Ephraim, "Statistical-Model-Based Speech Enhancement Systems," *Proc. IEEE*, Vol.80, pp 1526-1555, Oct.1992.

[2] H. Sheikhzadeh and L. Deng, "Waveform-Based Speech Recognition Using Hidden Filter Models: Parameter Selection And Sensitivity To Power Normalization," *IEEE Trans. Speech Audio Processing*, Vol. 2, pp. 80-89, Jan. 1994

[3] William D. Penny and Stephen J. Roberts, "Dynamic Models for nonstationary signal segmentation," *Technical Report TR-98-14*, Computers and Biomedical Research, pp. 483-502,1999

[4] Kevin P. Murphy, "Switching Kalman Filters, *Technical Report* TR-8-10, Compaq Cambridge Research Laboratory, 1998

[5] J. Bum Kim, K.Y. Lee, and C.W. Lee, "On the application of the Interacting Multiple Model Algorithm for Enhancing Noisy Speech," *IEEE Trans. Speech and Audio Processing*, Vol. 8, No 3, May 2000

[6] Y. Bar-Shalom and X. Li, *Estimation and Tracking: Princples, Techniques and Software*, Artech House, 1993

[7] C-J. Kim, "Dynamic linear Models with Markov Switching," *J. of Econometrics*, Vol. 60, pp1-22, 1994

[8] K.Y. Lee and K. Shirai, "Efficient Recursive Estimation For Speech Enhancement In Colored Noise," *IEEE Signal processing Lett.*, Vol. 3, pp196-199, July 1996

[9] N. Ma, M. Bouchard and R. Goubran, "Perceptual Kalman filtering for speech enhancement in colored noise," *Proceedings of IEEE ICASSP 2004*, Vol. 1, pp.717-720, Montreal, May 2004.

[10] L.R.Rabiner, "A Tutorial on Hidden Markov Models and Selected Application in Speech recognition", *Proc. IEEE*, Vol.77, pp.257-286, Feb. 1989

[11] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs," *ITU-T Recommendation* P.862, Feb. 2001

[12] http://www.cslu.ogi.edu/corpora/

[13] M. Gabrea, "Adaptive Kalman Filtering-based Speech enhancement algorithm," *Proc. of Canadian Conference on Electrical and Computer Engineering 2001*, Vol. 1, pp. 521-526, Fredericton, New-Brunswick, 2001

[14] 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

Input SNR	>25dB	10dB	5dB	0dB
No NR	97.77%	77.8%	66.7%	16.6%
SS	/	80.5%	66.7%	49.5%
KF	/	84.4%	75.0%	59.8%
SKF (IMM)	/	90.5%	80.0%	64.4%
SKF (GPB1)	/	89.0%	78.1%	67.1%
Proposed	/	84.4%	79.8%	69.6%

Table 3 Word recognition rate for speech mixed with white noise

Input SNR	>25dB	10dB	5dB	0dB
No NR	97.77%	81.1%	75.6%	28.9%
SS	/	72.2%	63.3%	45.6%
KF	/	84.4%	73.3%	53.3%
SKF (IMM)	/	86.7%	80.0%	66.7%
SKF (GPB1)	/	87.8%	80.0%	67.8%
Proposed	/	85.6%	80.0%	68.9%

Table 4 Word recognition rate for speech mixed with street noise