AUTOREGRESSIVE PARAMETER ESTIMATION FOR KALMAN FILTERING SPEECH ENHANCEMENT

Chang Huai YOU[+]*, Susanto RAHARDJA*[+]*, Soo Ngee KOH**

[+]Institute for Infocomm Research, Singapore 119613
*, [+] Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798

ABSTRACT

In this paper, autoregressive parameter estimation for Kalman filtering speech enhancement is studied. In conventional Kalman filtering speech enhancement, spectral subtraction is usually used for speech autoregressive (AR) parameter estimation. We propose log spectral amplitude (LSA) minimum mean-square error (MMSE) instead of spectral subtraction for the estimation of speech AR parameters. Based on an observation that full-band Kalman filtering speech enhancement often causes an unbalanced noise reduction between speech and non-speech segments, a spectral solution is proposed to overcome the unbalanced reduction of noise. This is done by shaping the spectral envelopes of the noise through likelihood ratio. Our simulation results show the effectiveness of the proposed method.

Index Terms— Speech Enhancement, Kalman Filtering, Autoregressive Model

1. INTRODUCTION

Single channel speech enhancement involves the application of speech-related characteristics in some of the signal processing techniques, such as short-term spectral amplitude (STSA) MMSE [1], Kalman filtering [2], hidden Markov model and signal subspace.

In this paper, a full-band Kalman filtering-based algorithm is investigated for the purpose of single channel speech enhancement. In comparison with spectral suppression, Kalman filtering speech enhancement has been proven to be effective for overcoming the tonal noise problem and non-stationary noise problem. It also achieves quite good speech quality by reducing the processing distortion introduced to the speech signals. This is because that the enhancement system uses well-established speech production model, short-term stationary nature of speech signal, as well as the well-representation of the observed measure in Kalman filtering.

The key problem of Kalman filtering lies on the state-space model, where the conventional speech enhancement approaches usually adopt AR model for both speech and noise. How to accurately estimate the AR parameters determines greatly the performance of the whole system. In [2], AR parameters are obtained by using the previously iterative estimate of the speech signal in an iterative Kalman filtering system. In [3, 4], AR parameters are estimated based on power spectral subtraction method.

In full-band Kalman filtering speech enhancement, it is observed that the enhanced speech contains much residual noise in speech segment and less residual noise in the non-speech segment. This causes an unbalanced reduction of the noise between speech and non-speech segments. In contrary, spectral suppression [1] and subband Kalman filtering [4] can uniformly suppress the noise in both pure-noise segment and speech segment.

In this paper, the LSA-MMSE is exploited to estimate the AR parameters for Kalman filtering speech enhancement. In order to moderate the unbalanced reduction, a spectral solution is introduced by applying speech spectral likelihood ratio into a time domain Kalman filtering system. In particular, the spectral likelihood ratio is used to shape the noise spectral envelope so that the AR parameters, including linear predictive coefficient (LPC) and the excitation variance in noise model, are modified accordingly over the full-band spectrum. In Section 2, a conventional Kalman filtering model for speech enhancement is introduced. In Section 3, the proposed Kalman filtering speech enhancement method is described. In Section 4, performance evaluation results are shown. The conclusion is given in Section 5.

2. KALMAN FILTER FOR SPEECH ENHANCEMENT

Let $s(n)$ and $v(n)$ denote the clean speech and noise respectively. The observed noisy speech, $x(n)$, is given by

$$x(n) = s(n) + v(n), \quad n = 1, 2, ... \tag{1}$$

The clean speech signal and noise are modeled as AR processes

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + w(n) \tag{2}$$

$$v(n) = \sum_{i=1}^{q} b_i v(n-i) + u(n) \tag{3}$$

where $w(n)$ and $u(n)$ are zero-mean white Gaussian processes with respective variances σ_w^2 and σ_u^2. Based on (1)-(3), Kalman process and Kalman measurement equations in state-space domain for speech enhancement are given by

$$\bar{\mathbf{s}}(n) = \bar{\mathbf{F}}\bar{\mathbf{s}}(n-1) + \bar{\mathbf{g}}\bar{\mathbf{w}}(n) \tag{4}$$

$$x(n) = \bar{\mathbf{C}}^T\bar{\mathbf{s}}(n) \tag{5}$$

where

$$\bar{\mathbf{s}}(n) = \begin{bmatrix}\mathbf{s}(n)\\ \mathbf{v}(n)\end{bmatrix}, \quad \bar{\mathbf{w}}(n) = \begin{bmatrix}w(n)\\ u(n)\end{bmatrix} \tag{6}$$

$$\bar{\mathbf{F}} = \begin{bmatrix}\mathbf{F} & \mathbf{0}\\ \mathbf{0} & \mathbf{F}_v\end{bmatrix}, \quad \bar{\mathbf{g}} = \begin{bmatrix}\mathbf{g} & \mathbf{0}\\ \mathbf{0} & \mathbf{g}_v\end{bmatrix}, \quad \bar{\mathbf{C}} = \begin{bmatrix}\mathbf{C}\\ \mathbf{C}_v\end{bmatrix} \tag{7}$$

$$\mathbf{F} = \begin{bmatrix}0 & 1 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & 1\\ a_p & a_{p-1} & \dots & a_1\end{bmatrix}, \mathbf{F}_v = \begin{bmatrix}0 & 1 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & 1\\ b_q & b_{q-1} & \dots & b_1\end{bmatrix} \tag{8}$$

$$\mathbf{s} = \begin{bmatrix}s(n-p+1) & \dots & s(n-1) & s(n)\end{bmatrix}^{\mathrm{T}}_{p\times 1} \tag{9}$$

$$\mathbf{v} = \begin{bmatrix}v(n-q+1) & \dots & v(n-1) & v(n)\end{bmatrix}^{\mathrm{T}}_{q\times 1} \tag{10}$$

$$\mathbf{g} = \mathbf{C} = \begin{bmatrix}0 & \dots & 0 & 1\end{bmatrix}^{\mathrm{T}}_{p\times 1}, \quad \mathbf{g}_v = \mathbf{C}_v = \begin{bmatrix}0 & \dots & 0 & 1\end{bmatrix}^{\mathrm{T}}_{q\times 1} \tag{11}$$

The estimate of the speech signal, $\hat{s}(n)$, can be obtained from the estimated state space of Kalman filtering by [1]:

$$\hat{s}(n) = \mathbf{C}_1^T\hat{\bar{\mathbf{s}}}(n|\mathbf{x}_n), \quad \mathbf{C}_1 = [\mathbf{C}^T \underbrace{0 \quad \dots \quad 0}_{q}]^T \tag{12}$$

3. PROPOSED KALMAN FILTERING SPEECH ENHANCEMENT

3.1. Speech AR Parameter Estimation

In Kalman filtering, there are many ways to estimate the speech AR parameters $\mathbf{a} = [a_p \quad ... \quad a_2 \quad a_1]^T$ and the variance of $w(n)$.

A conventional method to estimate the AR parameters is to use half-wave rectification applied to the power spectral density (PSD) (i.e., the so-called power spectral subtraction).

$$\tilde{\Psi}_s(k) = \max\Big\{\Psi_x(k) - \Psi_n(k), \quad \epsilon\Psi_x(k)\Big\} \tag{13}$$

where Ψ_x, Ψ_s and Ψ_n denote the respective PSDs of the observed noisy signal, speech signal and noise, and k denotes the frequency bin. The autocorrelation, $R_s(i)$ $(i = 0, ..., p)$, can be obtained through inverse discrete Fourier transform (IDFT), i.e.,

$$R_s = \mathrm{IDFT}\{\tilde{\Psi}_s(k) \mid k = 1, ..., K\} \tag{14}$$

In this paper, it is proposed to estimate the speech spectral amplitude using the LSA-MMSE estimation [1] and the PSD of speech signal is given by

$$\tilde{\Psi}_s(k) = \max\Big\{|G_{LSA}(k)X_k|^2, \quad \epsilon\Psi_x(k)\Big\} \tag{15}$$

where X_k is the discrete Fourier transform (DFT) of $x(n)$, and G_{LSA} is the gain function of LSA-MMSE [1] [2]. The AR parameters of speech $\mathbf{a}$ can be optimally estimated through Yule-Walker equation and obtained by using the Levinson-Durbin algorithm. Finally, the variance of $w(n)$, $\sigma_w^2(n)$, is obtained by

$$\begin{aligned}\sigma_w^2(n) &= \mathrm{E}\big\{|s(n) - \sum_{i=1}^{p} a_i^*(n)s(n-i)|^2\big\}\\ &= R_s(0) - 2\Re\big\{\sum_{i=1}^{p} a_i^* R_s(i)\big\} + \sum_{i=1}^{p}\sum_{j=1}^{p} a_i^* a_j R_s(i-j)\end{aligned} \tag{16}$$

where $\Re$ is the real part operator.

3.2. Noise Shaping Based on Likelihood Ratio

The conditional probability density function (PDF) of a noisy spectral component X_k at frequency bin k, given speech absence H_k^0 and speech presence H_k^1, is assumed to be a statistically independent Gaussian distribution, i.e.,

$$p(X_k|H_k^0) = \frac{1}{\pi\eta_{n,k}}\exp\{-\frac{|X_k|^2}{\eta_{n,k}}\} \tag{17}$$

$$p(X_k|H_k^1) = \frac{1}{\pi(\eta_{s,k}+\eta_{n,k})}\exp\{-\frac{|X_k|^2}{\eta_{n,k}+\eta_{s,k}}\} \tag{18}$$

where $\eta_{s,k}$ and $\eta_{n,k}$ are the respective variances of the speech signal and noise. The speech spectral likelihood ratio (LR) at the kth spectral bin, Λ_k, is obtained by

$$\Lambda_k = \frac{p(X_k|H_k^1)}{p(X_k|H_k^0)} = \frac{1}{1+\xi_k}\exp\{\frac{\xi_k}{1+\xi_k}\gamma_k\} \tag{19}$$

ξ_k and γ_k are the *a priori* and *a posteriori* SNRs respectively, which are defined as $\xi_k = \eta_{s,k}/\eta_{n,k}$ and $\gamma_k = |X_k|^2/\eta_{n,k}$. Usually the *a priori* SNR is approximated by using a 'decision-directed' method as follows

$$\hat{\xi}_k^{(l)} = \alpha\frac{|\hat{S}_k^{(l-1)}|^2}{\eta_{n,k}^{(l-1)}} + (1-\alpha)\max\{\gamma_k^{(l)}, 0\} \tag{20}$$

[1] The recursive equations for the state-space estimation for Kalman filtering speech enhancement can be found in our previous work [5].

[2] In practical implementation of the Kalman filtering system, in order to avoid the singular transition matrix due to speech AR parameter estimation, the term $\epsilon\Psi_x(k)$ of (15) is introduced in place of a very small constant, e.g., $\epsilon = 2.2 \times 10^{-16}$.

where the weighting factor α is set to 0.98 empirically; the enhanced spectral amplitude of the previous frame $(l-1)$, $|\hat{S}_k^{(l-1)}|$, is obtained by the LSA-MMSE method.

The AR parameters of a signal represents the spectral envelope of the signal. By shaping the spectral density envelope of the signal, the AR parameters will be changed accordingly. On the other hand, spectral likelihood ratio represents the ratio of the probability of the observed signal under speech presence situation to the one under speech absence situation at particular frequency bins.

In full-band Kalman filtering speech enhancement, it is noticed that the enhanced speech contains much residual noise in between the peaks of the enhanced speech spectrum during speech segment. Although the speech distortion is mitigated, the noise cannot be clearly removed. On the other hand, it is also observed that the noise is reduced very much in the non-speech segment. In order to overcome the unbalanced noise reduction, we propose to shape the spectral envelopes of the noise through the spectral likelihood ratio. Applying logarithm to (19), a log spectral likelihood is given by

$$\lambda_k = \log(\Lambda_k) = \frac{\xi_k}{1+\xi_k}\gamma_k - \log(1+\xi_k) \qquad (21)$$

From the log spectral likelihood equation (21), it can be seen that when the instantaneous SNR (*a posteriori* SNR-1, i.e., $\gamma_k - 1$) is high, according to Cappé's analysis [6], ξ_k is just the one sample delay of instantaneous SNR, i.e., it leads to a high ξ_k, so that $\xi_k/(1+\xi_k)$ approaches 1. Subsequently λ_k is mainly determined by γ_k, the first term of (21). When instantaneous SNR is low, ξ_k is a smoothed version of $(\gamma_k - 1)$ [6], $\lambda_k \approx \xi_k(\gamma_k - 1)$, i.e., it is a small value.

In order to solve the unbalanced problem of the residual noise, we introduce a smoothed log spectral likelihood ratio

$$\bar{\lambda}_k^{(l)} = \rho\bar{\lambda}_k^{(l-1)} + (1-\rho)\lambda_k^{(l)} \qquad (22)$$

where ρ is a smoothing factor. By the definition of the smoothed log spectral likelihood ratio, the ratio contains its past information, so that it can be mitigated in the case of a sudden change of speech spectral energy. Moreover, a frame log likelihood ratio is defined as follows:

$$\Xi = \frac{1}{K}\sum_{k=0}^{K-1}\lambda_k \qquad (23)$$

Obviously, the frame log likelihood ratio represents the relative strength of the speech segment in noise. In order to apply Ξ into Kalman filtering speech enhancement system, Ξ needs to be constrained and normalized by

$$\tilde{\Xi} = \frac{\max\{\min(\Xi,\ \Xi_{top}),\ \Xi_{low}\} - \Xi_{low}}{\Xi_{top} - \Xi_{low}} \qquad (24)$$

where Ξ_{top} and Ξ_{low} are two empirically constant values. In Kalman filtering speech enhancement application, it is observed that the reduction degree of noise can be controlled by regulating the level of input noise energy, and the shape of residual noise spectrum can be modified by changing the spectral shape of the input noise to Kalman filter. We propose to shape the PSD of noise by the normalized log spectral likelihood ratio and the normalized frame likelihood ratio, i.e.,

$$\tilde{\Psi}_n(k) = \tau_1\big[\tau_2 - (\{\tilde{\Xi}\}^{\tau_4} + \tau_3)\{\tilde{\lambda}_k\}^{\tau_5}\big]\Psi_n(k) \qquad (25)$$

where τ_i $(i = 1, ..., 5)$ are constant values determined empirically. $\tilde{\lambda}_k$ is the normalized log spectral likelihood ratio

$$\tilde{\lambda}_k = \frac{\bar{\lambda}_k - \min_k(\bar{\lambda}_k)}{\max_k(\bar{\lambda}_k) - \min_k(\bar{\lambda}_k)} \qquad (26)$$

The shaped noise autocorrelation can be obtained by

$$R_n = \mathrm{IDFT}\{\tilde{\Psi}_n(k) \mid k = 1, ..., K\} \qquad (27)$$

With R_n, the shaped noise AR parameters $\mathbf{b} = [b_q \quad ... \quad b_2 \quad b_1]^T$ can be obtained. Subsequently, the modified variance of $u(n)$, $\sigma_u^2(n)$, is obtained by

$$\sigma_u^2(n) = R_n(0) - 2\Re\Big\{\sum_{i=1}^{q} b_i^* R_n(i)\Big\} + \sum_{i=1}^{q}\sum_{j=1}^{q} b_i^* b_j R_n(i-j) \qquad (28)$$

3.3. Implementation of the Proposed System

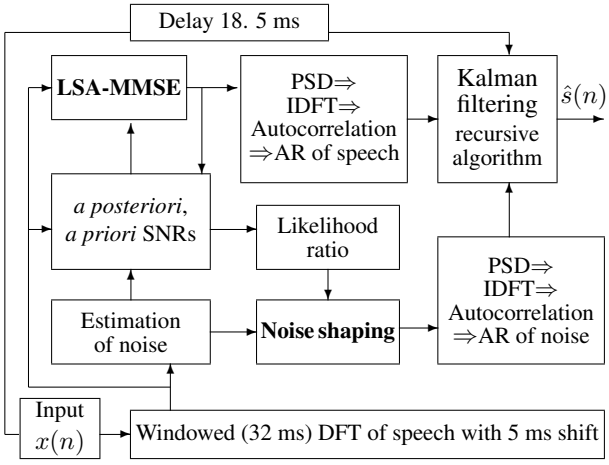

Fig. 1. The proposed Kalman filtering speech enhancement.

The implementation of the proposed speech enhancement system is described in Fig. 1. The observed noisy speech signal inputs to a buffer (e.g., 256 samples, corresponding to 32 ms at 8 kHz sampling rate) sample by sample. The noisy $x(n)$ is extracted from a central small frame (e.g., 40 samples (5 ms)) sample by sample and input to the Kalman filter. At the same time, the (32 ms) buffer is used to estimate the AR parameters of speech and noise. In each 5 ms, the AR parameters are updated and input to the Kalman filtering system [3].

[3]Theoretically, the AR parameters need to be updated sample by sample

Table 1. The performances of the speech enhancement methods including conKal-Ny: conventional Kalman filtering (denoted by conKal) with AR parameter estimation by using noisy speech; conKal-ITn: the nth iteration by using the iterative Kalman filtering scheme [2]; conKal-SS: conKal with AR estimation by using power spectral subtraction method (13); conKal-Ideal: conKal with AR estimation by using clean speech; Spect-subtr: power spectral subtraction (13); LSA-MMSE; Kal-LSA; Kal-LSA-n1: Kal-LSA and noise shaping where q = 5; and Kal-LSA-n2: Kal-LSA and noise shaping where q = 15.

Methods	Objective Measurements		
	seg.SNR (dB)	IS distortion	MBSD
Noisy	0.00	0.263	1.496
conKal-Ny	2.65	0.205	0.821
conKal-IT1	4.05	0.187	0.540
conKal-IT3	9.02	0.183	0.459
conKal-SS	5.20	0.179	0.463
conKal-Ideal	9.67	0.139	0.308
Spect. Subtr	2.47	0.328	0.845
LSA-MMSE	6.02	0.273	0.564
Kal-LSA	8.16	0.178	0.349
Kal-LSA-n1	9.36	0.172	0.323
Kal-LSA-n2	9.50	0.170	0.320

4. PERFORMANCE EVALUATION

To evaluate the performance of the proposed method, 10 utterances from the TIMIT database are selected and downsampled to 8 kHz. In this simulation, the proposed enhancement system has $\tau_1 = 1.5$, $\tau_2 = 2$, $\tau_3 = 0.83$, $\tau_4 = 0.2$, $\tau_5 = 0.2$, $p = 10$, $\rho = 0.5$, $\Xi_{top} = 1600$, and $\Xi_{low} = 0.05$. For the noise model, since we focus on the noise spectral shaping, the order q is set to 15 (conventional q = 5).

The proposed Kalman filtering speech enhancement with LSA-MMSE for AR parameter estimation (Kal-LSA) is evaluated by comparing with other speech enhancement methods. Table 1 shows the simulation results where the speech signal is contaminated by white noise. It can be seen that our proposed method (Kal-LSA-n) is consistently closer to the upperbound case (conKal-Ideal) than the other methods in terms of the segmental SNR, Itakura-Saito (IS) distortion and modified Bark spectral distortion. Figure 2 shows the spectrograms of the speech enhanced by using the proposed Kal-LSA without and with noise shaping. It can be seen that the noise level is further reduced in the speech segment when applying the proposed noise shaping.

for each output of Kalman filter. However, it is observed that the 5 ms AR update is enough for the accuracy of the Kalman filtering speech enhancement. Therefore the computational complexity is greatly reduced.

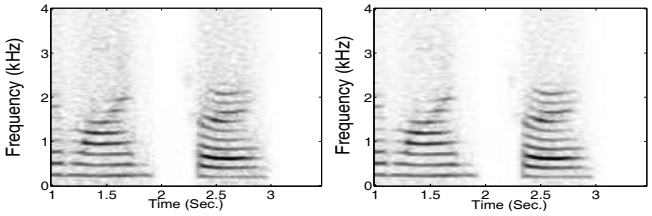

Fig. 2. Spectrograms of the speech enhanced by using (a) Kal-LSA without noise shaping; (b) Kal-LSA with noise shaping.

5. CONCLUSION

In this paper, we consider the advantages of Kalman filtering and MMSE spectral suppression into a single channel speech enhancement system. It is found that the use of the LSA-MMSE method to estimate the AR parameter of speech signal can greatly improve the performance of Kalman filtering speech enhancement system as compared to the use of the power spectral substraction method. Furthermore, the unbalanced noise reduction problem of full-band Kalman filtering is highlighted. For the purpose of solving the problem, a noise spectral shaping method based on speech likelihood ratio is introduced to moderate the unbalanced noise reduction. Simulation results confirm that the proposed noise shaping method is effective.

6. REFERENCES

[1] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-33, pp. 443-445, Apr. 1985.

[2] J.D. Gibson, B. Koo, and S.D. Gray, "Filtering of Colored Noise for Speech Enhancement and Coding," *IEEE Trans. Signal Processing*, Vol. 39, No. 8, pp. 1732-1742, Aug.1991.

[3] P. Sörqvist, P. Händel and B. Ottersten, "Kalman Filtering for Low Distortion Speech Enhancement in Mobile Communication," *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, ICASSP-1997, Vol 2, pp. 1219-1222, 1997.

[4] C.H. You, S.N. Koh and S. Rahardja, "Kalman Filtering Speech Enhancement Incorporating Masking Properties for Mobile Communication in a Car Environment," *Proc. IEEE International Conference on Multimedia and Expo*, ICME'2004, pp. 1343-1346, Jun. 2004.

[5] C.H. You, S. Rahardja and S.N. Koh , "Perceptual Kalman Filtering Speech Enhancement," *Proc. IEEE International Conference on Acoustic, Speech and Signal Processing*, ICASSP-2006, Toulouse, France, May 2006.

[6] O. Cappé, "Elimination of Musical Noise Phenonmenon with the Ephraim and Malah Noise Suppression," *IEEE Trans. Speech and Audio Processing*, Vol. 2, pp. 345-349, Apr. 1994.