

Improved Kalman Filter-Based Speech Enhancement

Jianqiang Wei, Limin Du, Zhaoli Yan and Hui Zeng

Institute of Acoustics

Chinese Academy of Sciences, Beijing, China

{weijq, dulm, yanzl and zengh}@iis.ac.cn

Abstract

In this paper, a Kalman filter-based speech enhancement algorithm with some improvements of previous work is presented. A new technique based on spectral subtraction is used for separation speech and noise characteristics from noisy speech and for the computation of speech and noise autoregressive (AR) parameters. In order to obtain a Kalman filter output with high audible quality, a perceptual post-filter is placed at the output of the Kalman filter to smooth the enhanced speech spectra. Experiments indicate that this newly proposed method works well.

1. Introduction

The enhancement of speech corrupted by noise is an important problem with numerous applications ranging from reduction of environmental noise for communication system to preprocessing for speech recognition system. Therefore, speech enhancement algorithms have attracted a great deal of interest in the past two decades.

There have been numerous studies [1], [2] dealing with enhancement of speech contaminated by noise. However, most approaches use the stationary Gaussian white noise assumption. But colored noise assumption proved to be more useful for speech enhancement [2].

Because of its high flexibility, the Kalman filter is widely used for signal enhancement. It can handle colored noise and has a reasonable numerical complexity. Moreover, Kalman filtering is a model based adaptive method, where speech as well as noise is modeled as AR processes. Thus, a key issue in Kalman filtering is the estimation of the AR parameters in the presence of noise. The traditional algorithm employs the EM method [3], [4] to iteratively estimate the AR parameters of speech and noise. Unfortunately, its computational complexity is high. The method used in our work is based on spectral subtraction for estimation of speech and noise AR parameters. It is computationally efficient and can be easily implemented. But since spectral subtraction is applied, musical noise [1] appears in the enhanced speech.

In order to obtain a Kalman filter output with high audible quality, a perceptual post-filter is placed at the output of the Kalman filter to decrease the musical noise level.

The paper is organized as follows. In Section 2, the signal model and Kalman filter are presented. In Section 3, the parameter estimation based on spectral subtraction is presented. The operation of the perceptual post-filter is described in Section 4. Experimental results are provided in Section 5. Finally, in Section 6, conclusions are given.

2. The signal model and Kalman filter

Let the noisy speech is modeled as a sum of two AR processes, i.e.

$$z(t) = s(t) + v(t) \quad (1)$$

where $z(t)$ represents the measured signal, $s(t)$ represents the speech and $v(t)$ represents the additive colored background noise. Further,

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

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

where p and q denote the model order for $s(t)$ and $v(t)$, respectively. The noises $w(t)$ and $u(t)$ are assumed to be zero mean Gaussian white noises with variances σ_w^2 and σ_u^2 , respectively. Due to the short-time stationarity of $s(t)$, $\{a_i\}$ may be assumed to be time invariant for 10-40ms. The noise parameters $\{b_i\}$ may typically be assumed to be constant for 1-2s [8].

In order to be able to use the Kalman filter, the double AR model Eqs. (1)-(3) may be converted to the state-space form, i.e.

$$\mathbf{x}(t) = \mathbf{\Phi}\mathbf{x}(t-1) + \mathbf{G}\mathbf{r}(t) \quad (4)$$

$$z(t) = \mathbf{h}^T \mathbf{x}(t) \quad (5)$$

where the state vector $\mathbf{x}(t)$ is defined by

$$\mathbf{x}^T(t) = [s(t-p+1) \cdots s(t), v(t-q+1) \cdots v(t)] \quad (6)$$

and the state transition matrix $\mathbf{\Phi}$ is given by

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}_v \end{bmatrix} \quad (7)$$

where

$$\mathbf{\Phi}_s = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ a_p & a_{p-1} & \cdots & \cdots & a_2 & a_1 \end{bmatrix} \quad (8)$$

$$\mathbf{\Phi}_v = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & \ddots & \ddots & & \vdots \\ \vdots & & & \ddots & \ddots & \vdots \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \\ b_q & b_{q-1} & \cdots & \cdots & b_2 & b_1 \end{bmatrix} \quad (9)$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{g}_v \end{bmatrix} \quad (10)$$

where \mathbf{g}_s and \mathbf{g}_v are the following p and q dimensional vectors, i.e.

$$\mathbf{g}_s^T = [0 \cdots 0 \ 1] \quad (11)$$

$$\mathbf{g}_v^T = [0 \cdots 0 \ 1] \quad (12)$$

and

$$\mathbf{h}^T = [\mathbf{h}_s^T \quad \mathbf{h}_v^T] \quad (13)$$

where \mathbf{h}_s and \mathbf{h}_v are the following p and q dimensional vectors, i.e.

$$\mathbf{h}_s^T = [0 \quad \dots \quad 0 \quad 1] \quad (14)$$

$$\mathbf{h}_v^T = [0 \quad \dots \quad 0 \quad 1] \quad (15)$$

Further,

$$\mathbf{r}(t) = [w(t) \quad u(t)]^T \quad (16)$$

$$\mathbf{Q} = \mathbf{G}E[\mathbf{r}(t)\mathbf{r}(t)^T]\mathbf{G}^T = \mathbf{G}\mathbf{R}\mathbf{G}^T \quad (17)$$

where

$$\mathbf{R} = \begin{bmatrix} \sigma_w^2 & 0 \\ 0 & \sigma_u^2 \end{bmatrix} \quad (18)$$

From the Kalman filtering theory [5], Kalman filtering is performed in two steps:

- the data step:

$$\mathbf{L}(t) = \mathbf{P}(t|t-1)\mathbf{h}^T(\mathbf{h}\mathbf{P}(t|t-1)\mathbf{h}^T)^{-1} \quad (19)$$

$$\hat{\mathbf{x}}(t|t) = \hat{\mathbf{x}}(t|t-1) + \mathbf{L}(t)(z(t) - \mathbf{h}^T\hat{\mathbf{x}}(t|t-1)) \quad (20)$$

$$\mathbf{P}(t|t) = \mathbf{P}(t|t-1) - \mathbf{L}(t)\mathbf{h}\mathbf{P}(t|t-1) \quad (21)$$

- the time step:

$$\hat{\mathbf{x}}(t+1|t) = \Phi\hat{\mathbf{x}}(t|t) \quad (22)$$

$$\mathbf{P}(t+1|t) = \Phi\mathbf{P}(t|t)\Phi^T + \mathbf{Q} \quad (23)$$

where $\mathbf{P}(t+1|t)$ is the covariance matrix for the prediction error, $\mathbf{P}(t|t)$ is the covariance matrix for the estimation error, $\mathbf{L}(t)$ is the Kalman gain which controls the step-size, $\hat{\mathbf{x}}(t+1|t)$ is the priori estimate for the state vector and $\hat{\mathbf{x}}(t|t)$ is the posteriori estimate.

Since only an estimate of the noise-free speech signal is needed, the output equation of the Kalman filter will be

$$s(t) = \mathbf{h}_2^T \mathbf{x}(t) \quad (24)$$

where \mathbf{h}_2 contains zeros except $\mathbf{h}_2(p) = 1$.

It must be noted that Kalman filter offers optimal estimate when the system parameters Φ , \mathbf{h} , \mathbf{G} and \mathbf{Q} are known [2], so that it is important that system parameters be estimated as accurate as possible.

3. A new parameter estimation method

Since the AR model is built into the structure of the Kalman filter, the choice of parameter estimation method is very important. The traditional algorithm employs the EM method to iteratively estimate the AR parameters of speech and noise. Unfortunately, its computational complexity is high.

In this section, technique based on spectral subtraction will be used for separation speech and noise characteristics from noisy speech and for the computation of speech and noise AR parameters. It is computationally efficient and can be easily implemented.

From the double AR model for noisy speech mentioned above, the power spectral density $P_z(\omega)$ of noisy speech may be divided into a sum of the power spectral density $P_s(\omega)$ of speech and the power spectral density $P_v(\omega)$ of background noise, i.e.

$$P_z(\omega) = P_s(\omega) + P_v(\omega) \quad (25)$$

from Eq. (2) it follows that

$$P_s(\omega) = \frac{\sigma_s^2}{\left|1 + \sum_{i=1}^p a_i e^{-j\omega i}\right|^2} \quad (26)$$

similarly from Eq. (3) it follows that

$$P_v(\omega) = \frac{\sigma_v^2}{\left|1 + \sum_{i=1}^q b_i e^{-j\omega i}\right|^2} \quad (27)$$

From Eqs. (1)-(3) it follows that $z(t)$ may be denoted by an autoregressive moving average (ARMA) model with power spectral density $P_z(\omega)$. An estimate of $P_z(\omega)$ can be achieved by an autoregressive (AR) model, i.e.

$$\hat{P}_z(\omega) = \frac{\hat{\sigma}_z^2}{\left|1 + \sum_{i=1}^r \hat{c}_i e^{-j\omega i}\right|^2} \quad (28)$$

where $\{\hat{c}_i\}$ and $\hat{\sigma}_z^2$ are the estimated parameters of the AR model

$$z(t) = \sum_{i=1}^r c_i z(t-i) + \eta(t) \quad (29)$$

where the variance of $\eta(t)$ is given by σ^2 , $p \leq r \leq N$ and N is the frame length.

From Eqs. (25)-(29) it follows that if the PSD estimates $\hat{P}_s(\omega)$ and $\hat{P}_v(\omega)$ can be achieved based on the noisy speech signal, the PSD estimate $\hat{P}_z(\omega)$ of speech will be

$$\hat{P}_s(\omega) = \hat{P}_z(\omega) - \delta \hat{P}_v(\omega) \quad (30)$$

where δ is a scalar design variable, typically lying in the interval $0 < \delta < 4$. In normal cases δ has a value around 1 ($\delta = 1$ corresponds to Eq. (25)).

Furthermore, the needed system parameters can be calculated based on the estimated PSD $\hat{P}_s(\omega)$.

Hence, the primary problem is to estimate the PSD $\hat{P}_s(\omega)$ and $\hat{P}_v(\omega)$ from the noisy speech which can be measured.

The structure of the new parameter estimation technique is illustrated in Fig. 1.

3.1. Frame spectrum estimation

Here any kind of PSD estimator [6] may be used, for example parametric or non-parametric estimation. Two of the main methods are periodogram and minimum entropy (MEM) estimate.

In the periodogram method, the frame is first multiplied by Hanning window (or other suitable window type) and then the FFT is used. The estimate is further averaged (or recursively smoothed)

$$\hat{P}_z(\omega) = \alpha \hat{P}_z(\omega) + (1 - \alpha) P_z(\omega) \quad (31)$$

where α is selectable smoothing factor.

In the MEM method, the AR model parameters of the signal are first estimated, for example by using Burg's method, and then estimated spectrum is obtained according to Eq. (28).

3.2. Frame noise estimation

The spectral minima tracking algorithm by Doblinger [8] can be applied for frame noise spectrum estimation. It is computationally more efficient and can be easily adjusted to different kinds of noise disturbances by means of only two parameters, i.e.

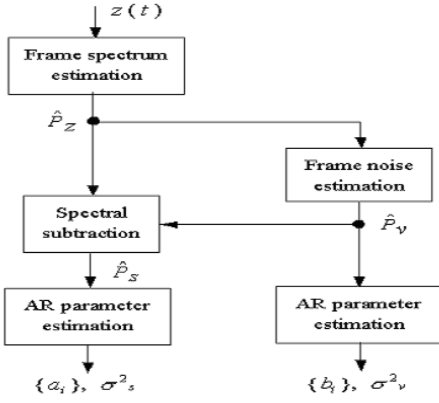


Figure 1: The structure of the new parameter estimation.

If $\hat{P}_v(\omega, m-1) < \hat{P}_z(\omega, m)$ then

$$\hat{P}_z(\omega, m) = \gamma \hat{P}_v(\omega, m-1) + \frac{1-\gamma}{1-\beta} (\hat{P}_z(\omega, m) - \beta \hat{P}_z(\omega, m-1)) \quad (32)$$

else

$$\hat{P}_v(\omega, m) = \hat{P}_z(\omega, m) \quad (33)$$

where $\hat{P}_v(\omega, m)$ denotes the noise spectral component of frequency ω in the m^{th} frame, $\hat{P}_z(\omega, m)$ denotes the smoothed spectral component of frequency ω in the m^{th} frame for noisy speech signal (corresponding to Eq. (31)), γ and β are adjustable variables.

An alternative noise spectrum estimator has been presented by Martin in [7].

3.3. Spectral subtraction

Based on $\hat{P}_z(\omega)$ and $\hat{P}_v(\omega)$, calculate the estimate of the speech PSD $\hat{P}_s(\omega)$ using Eq. (30). Here several subtraction rules can be used [1], simple power spectral subtraction, power spectral subtraction with half wave rectification or full wave rectification, Wiener filter and modified power spectral subtraction with half wave rectification.

3.4. AR parameter estimation

Finally, AR parameters of speech and noise can be achieved from the estimates of PSD $\hat{P}_s(\omega)$ and $\hat{P}_v(\omega)$, respectively. Here the estimated speech and noise PSD $\hat{P}_s(\omega)$ and $\hat{P}_v(\omega)$ are first converted to autocorrelation sequences and then some system identification methods such as Burg's and exact AR modeling algorithm may be applied to get the needed AR parameters.

4. Operation of the perceptual post-filter

Since spectral subtraction is used for estimation speech and noise parameters, musical noise appears in the enhanced speech. In order to obtain a Kalman filter output with high audible quality, a perceptual post-filter is placed at the output of the Kalman filter to further suppress musical noise. By utilizing a perceptual filter, averaging is performed in a manner similar to that of the human auditory system [9].

The perceptual filter minimizes signal distortion while constraining the noise spectrum residual to be beneath the masking threshold [10]. The constrained optimization

problem will be

$$\mathbf{T}^* = \arg \min_{\mathbf{T}} [E \{ \| (\mathbf{T}\mathbf{F}^H - \mathbf{F}^H) \hat{\mathbf{s}} \|_2^2 \}]$$

$$\text{subject to: } E \{ \|\mathbf{f}_i^H \mathbf{T} \hat{\mathbf{v}}\|^2 \} < m_i^2 \quad (34)$$

where $\mathbf{T}^* = \text{diag}\{t_1^*, \dots, t_N^*\}$, N denotes the frame size, \mathbf{F} is the Fourier transform matrix, and m_i is the masking threshold calculated according to the model described in the Perceived Audio Quality ITU Recommendation (ITU-R BS. 1387).

The optimal filter can be formulated as

$$t_i^* = \begin{cases} 1 & m_i \geq \hat{P}_v(\omega_i)^{\frac{1}{2}} \\ \frac{m_i}{\hat{P}_v(\omega_i)^{\frac{1}{2}}} & m_i < \hat{P}_v(\omega_i)^{\frac{1}{2}} \end{cases} \quad (35)$$

where $\hat{P}_v(\omega_i)$ is the power spectral density of the noise signal in frequency ω_i .

5. Experimental results

In order to evaluate the performance of the newly proposed improved Kalman filter-based speech enhancement (IKF) algorithm, both objective and subjective tests are conducted in comparison with the following algorithms:

- The spectral subtraction based on minimum statistics by Martin [7].
- The computationally efficient speech enhancement by spectral minima tracking in subbands by Doblinger [8].

In the experiments described below, the AR orders used to model the speech signal and noise signal are both eight, the frame size is 32ms and 75% overlap. The speech signal is artificially degraded (mixed) by additive noise at various SNR ranging from -10 to 10dB. Various recorded noise sources, including computer fan, highway and helicopter, are considered. The use of artificially mixed signals is motivated by the need of making the original clean speech signal available in order to evaluate the performances.

5.1. SNR improvements

In this experiment, the segmental SNR of the clean and enhanced signals is applied to conduct the objective test. This distortion measure is known to be correlated with the subjective perception of speech quality.

Let $s(t)$ and $\hat{s}(t)$ denote the clean and enhanced speech signal, respectively. Thus, the segmental SNR is calculated according to the following formula

$$\text{SNR}_{\text{seg}} = \frac{1}{M} \sum_{t=0}^{M-1} 10 \log_{10} \frac{\sum_{t=T_i}^{T_i+T-1} s^2(t)}{\sum_{t=T_i}^{T_i+T-1} (\hat{s}(t) - s(t))^2} \quad (36)$$

where the frame length $T = 128$ and averaging is performed across all frames.

Fig.2 shows the sample mean of the output SNR_{seg} for various input SNR_{seg} at various noise sources for the total three algorithms mentioned above.

According to Fig. 2, it can be seen that in terms of output SNR_{seg} IKF (improved Kalman filter-based speech enhancement with perceptual post-filter) outperforms the other two algorithms for input $\text{SNR}_{\text{seg}} > 0\text{dB}$, while for the lower input SNR_{seg} , Martin's algorithm is preferable.

5.2. Spectrograms

An alternative objective test is performed by comparing the spectrograms of the enhanced signal (some speech segment) processed by all the algorithms mentioned above with that of the clean speech. Fig.3 shows the spectrograms of some clean speech segment (a), the corresponding noisy segment (b), the enhanced segment processed by Martin's algorithm (c), the enhanced segment processed by Doblinger's algorithm (d) and the enhanced segment processed by IKF (e).

As can be seen, the IKF algorithm shows better noise reduction without obvious distortion of the speech signal. Furthermore, the distribution of residual noise is very uniform so as to fewer musical noise.

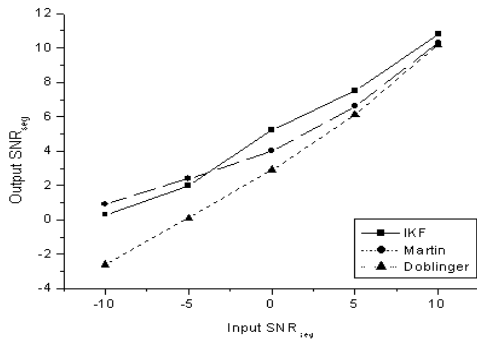


Figure 2: The dependence of output SNR_{seg} on input SNR_{seg} for all the algorithms tested.

5.3. Listening tests

Although the segmental SNR and spectrogram measurement provide a valuable information about the performance of speech enhancement algorithms, they do not characterize all its aspects. For example, they do not take into account the unvoiced signal sections at all. Hence, the listening tests are also incorporated in our experiments.

Our informal test involves a few listeners. The speech signal is degraded by some recorded noise signals at various SNR conditions. Each listener is demanded to compare the quality of the enhanced speech signal with that of corrupted one without knowing which signal corresponds to which algorithm. All listeners indicate that the quality of the speech processed by IKF algorithm is superior to that of the other algorithms because of better noise reduction (especially fewer musical noise) without obvious distortion of the speech. It is in accordance with the objective tests.

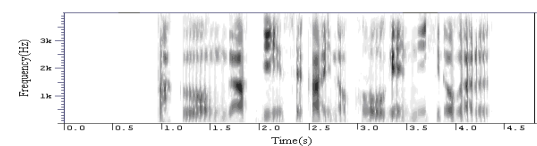
6. Conclusions

In this work, an improved Kalman filter-based speech enhancement algorithm with perceptual post-filter is presented. As opposed to the traditional methods, a new parameter estimation technique based on spectral subtraction has been applied for separation speech and noise characteristics from noisy speech and for the computation of speech and noise autoregressive (AR) parameters. Finally, in order to further decrease musical noise level, a perceptual post-filter is applied. Experiments show that the presented algorithm really works well.

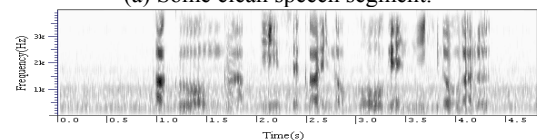
7. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using

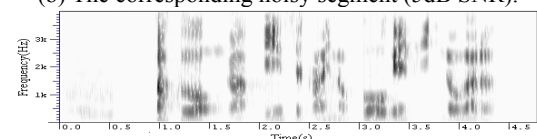
- spectral subtraction", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113-120, 1979.
- [2] Gibson J. D., Koo B. and Gray S. D., "Filtering of colored noise for speech enhancement and coding", *IEEE Trans. Signal Processing*, vol. 39, no. 8, pp. 1732-1742, Aug. 1991.
- [3] B. G. Lee, K. Y. Lee and S. Ann, "An EM-based approach for parameter enhancement with an application to speech signals", *Signal Processing*, vol. 46, no. 1, pp. 1-14, Sep. 1995.
- [4] S. Gannot, D. Burshtein and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithm", *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 4, July 1998.
- [5] S. Haykin, *Adaptive filter theory*, Prentice-Hall, 1993.
- [6] A. V. Oppenheim and R. W. Schaffer, *Digital signal processing*, Prentice-Hall, 1975.
- [7] R. Martin, "Spectral subtraction based on minimum statistics", *Proc. Seventh European Signal Processing Conference*, pp. 1182-1185, Sept. 1994.
- [8] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands", *Proceedings of EUROSPEECH'95*, vol. 2, pp. 1513-1516, Sept. 1995.
- [9] N. Virag, "Signal channel speech enhancement based on masking properties of the human auditory system", *IEEE Trans. on Speech and Audio Processing*, vol. 7, pp. 126-137, Mar. 1999.
- [10] *Method for Objective measurements of perceived audio quality*, Recommendation ITU-R BS. 1387, International Telecommunication Union, July 1999.



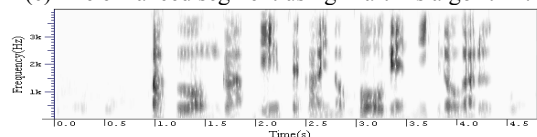
(a) Some clean speech segment.



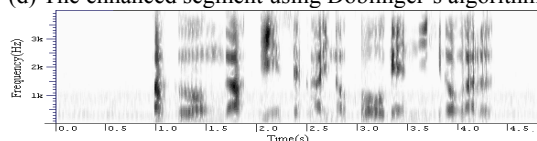
(b) The corresponding noisy segment (5dB SNR).



(c) The enhanced segment using Martin's algorithm.



(d) The enhanced segment using Doblinger's algorithm.



(e) The enhanced segment using IKF algorithm.

Figure 3: Comparison of speech spectrograms.