SPEECH ENHANCEMENT USING A MODIFIED KALMAN FILTER BASED ON COMPLEX LINEAR PREDICTION AND SUPERGAUSSIAN PRIORS

Thomas Esch and Peter Vary

Institute of Communication Systems and Data Processing (ind)
RWTH Aachen University, Germany
{esch|vary}@ind.rwth-aachen.de

ABSTRACT

This paper presents a modified Kalman Filter operating in the frequency domain for single channel speech enhancement. The proposed scheme uses a two step approach. In the first step, information from previous, enhanced speech DFT coefficients is exploited to perform an estimation of the current speech coefficients. Investigations show that the highest prediction gain is achieved by modeling the temporal trajectory of the speech DFT coefficients as a complex autoregressive (AR) process. In the second step, the first prediction is updated using three alternative spectral estimators, including the conventional Kalman Filter gain. Instrumental measurements show the improvement of the proposed scheme compared to purely statistical weighting rules.

Index Terms— Speech enhancement, noise reduction, adaptive Kalman filtering, linear prediction

1. INTRODUCTION

When a speech communication device is used in environments with high levels of ambient noise, the noise picked up by the microphone significantly impairs the quality and the intelligibility of the transmitted speech signal. In order to get a reliable separation from the noise signal (e.g., engine noise, street noise), noise reduction algorithms have become part of digital speech coding systems recently. They are used for example in mobile communications, in hearing aids and in hands-free devices.

State-of-the-art noise suppression systems are based on the so called 'spectral weighting' approach. The Discrete Fourier Transform (DFT) is used to perform noise suppression in the frequency domain by applying individual adaptive gains to each frequency bin. Most of the rules, proposed in literature, have been derived under certain assumptions about the statistics of the speech and noise DFT coefficients. The well-known Wiener Filter [1], for instance, is derived under the assumption that speech and noise samples are Gaussian distributed. Recently, the use of more sophisticated distributions to model the statistics of speech and noise were proposed, e.g., [2] and [3]. All of these assumptions may be classified as memory-less a priori knowledge, as only the probability distributions of either complex DFT coefficients, real-valued magnitudes or phase coefficients are considered. Correlation in time is not taken into account.

The authors in [4] were the first who proposed the use of a Kalman Filter for the purpose of speech enhancement. Compared to the Wiener filtering method, the performance of this model-based approach was shown to be considerably better. In order to reduce complexity, the authors in [5] introduced a Kalman filtering system in the sub-band domain that additionally achieved better results than the full-band time approach. In [6], the application of a Kalman Filter in sub-bands was further investigated and improved. However, most of these techniques only consider the temporal correlation within one frame and only a very limited number of proposals are known which also take into account the correlation of successive speech frames, e.g., [7].

In this paper, a Kalman Filter approach is presented that uses complex prediction to exploit the temporal correlation of successive speech DFT coefficients. The resulting prediction error is estimated in a second step applying different statistical estimators which are suitable for the statistics of the error signal. The remainder of this paper is organized as follows. In Sec. 2, a brief overview about the proposed system is given. Sec. 3 comprises the individual steps of the modified Kalman Filter in detail. Experimental results are reported in Sec. 4 and conclusions are drawn in Sec. 5.

2. SYSTEM OVERVIEW

The clean speech signal $s(k)$ is assumed to be degraded by an additive noise signal $n(k)$ to produce the noisy signal

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

where k is the discrete time index. Fig. 1 illustrates the simplified block diagram of the system that was considered within this work for noise reduction. The decomposition of speech and noise is performed in the frequency domain. Therefore, the noisy input signal $y(k)$ is segmented into overlapping frames of length L_{F}. After windowing (e.g., applying a Hann window), these frames are transformed via Fast Fourier Transform (FFT). The spectral coefficient of the noisy input signal at frequency bin μ and frame λ is given by

$$Y(\lambda,\mu) = S(\lambda,\mu) + N(\lambda,\mu) = R(\lambda,\mu)\mathrm{e}^{j\vartheta(\lambda,\mu)}, \tag{2}$$

where $S(\lambda,\mu)$ and $N(\lambda,\mu)$ represent the spectral speech and noise coefficients. $R(\lambda,\mu)$ and $\vartheta(\lambda,\mu)$ are the corresponding noisy magnitude and phase, respectively. Moreover, the magnitude of the

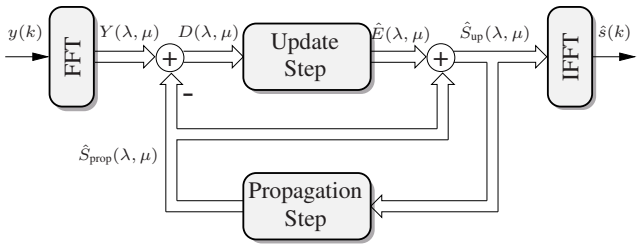

Fig. 1. System block diagram

This work was supported by Nokia, Tampere, Finland.

speech coefficient $S(\lambda,\mu)$ is denoted by $A(\lambda,\mu)$ and the corresponding phase by $\alpha(\lambda,\mu)$.

The system used for speech enhancement is based on a *Kalman Filter* structure that can be divided into two steps. In the first step, called *propagation step*, the temporal correlation of successive speech frames is exploited. The estimate $\hat{S}_{\text{prop}}(\lambda,\mu)$ of the current speech coefficient $S(\lambda,\mu)$ is propagated in time based on information taken from the previous N_{K} enhanced coefficients, i.e.,

$$\hat{S}_{\text{prop}}(\lambda,\mu) = f\left(\hat{S}_{\text{up}}(\lambda-1,\mu),\,\ldots\,,\hat{S}_{\text{up}}(\lambda-N_{\text{K}},\mu)\right). \quad (3)$$

In the second step, called *update step*, this first estimation is updated by utilizing the noisy coefficient $Y(\lambda,\mu)$ of the current frame. Considering the differential signal

$$D(\lambda,\mu) = Y(\lambda,\mu) - \hat{S}_{\text{prop}}(\lambda,\mu), \quad (4)$$

the update step estimates the prediction error

$$E(\lambda,\mu) = S(\lambda,\mu) - \hat{S}_{\text{prop}}(\lambda,\mu) \quad (5)$$

of the propagation step. As will be seen later in Sec. 3.2, an adequate statistical estimator can be used for this purpose. Thus, the estimate $\hat{E}(\lambda,\mu)$ of the prediction error can be stated as a spectral weighting of the differential signal by multiplying the Kalman gain $K(\lambda,\mu)$:

$$\hat{E}(\lambda,\mu) = K(\lambda,\mu)D(\lambda,\mu). \quad (6)$$

The results of both steps are combined to get the enhanced speech coefficient

$$\hat{S}'_{\text{up}}(\lambda,\mu) = \hat{S}_{\text{prop}}(\lambda,\mu) + \hat{E}(\lambda,\mu). \quad (7)$$

It turned out that speech distortions are reduced by omitting the phase of $\hat{S}'_{\text{up}}(\lambda,\mu)$ and using instead the short-time phase of the noisy input signal for reconstruction, i.e.,

$$\hat{S}_{\text{up}}(\lambda,\mu) = \left|\hat{S}'_{\text{up}}(\lambda,\mu)\right| \mathrm{e}^{j\vartheta(\lambda,\mu)}. \quad (8)$$

In order to obtain the enhanced signal in the time domain, an Inverse Fast Fourier Transform (IFFT) and the overlap-add method are applied.

3. MODIFIED KALMAN FILTER

This section addresses the basic principles of the afore mentioned propagation and update steps in Fig. 1. The main differences compared to a conventional Kalman Filter (e.g., [4], [5], [6]) used for speech enhancement are presented.

3.1. Propagation Step

For the estimation of the current speech coefficient $S(\lambda,\mu)$ within the propagation step, the autoregressive (AR) speech model is used which has been proven to be very effective for modeling the human speech production system. In contrast to most other speech processing algorithms, the AR process is used here in the frequency domain to model the temporal trajectory of each frequency bin. Thus, the speech coefficient $S(\lambda,\mu)$ can be stated as:

$$S(\lambda,\mu) = \sum_{i=1}^{N_{\text{K}}} \hat{a}_i(\lambda,\mu)\hat{S}_{\text{up}}(\lambda-i,\mu) + E(\lambda,\mu), \quad (9)$$

where N_{K} is the model order and $\hat{a}_i(\lambda,\mu)$ is the i-th AR coefficient that has to be estimated in advance.

While in [7] the authors propose a system that depends on two separate Kalman Filters for real and imaginary part, a *complex* predictor is required here in order to compute the spectral coefficient $D(\lambda,\mu)$, see Eq. (4). For this purpose, three alternative methods are investigated in the following to perform the complex estimation:

1) *Predict magnitude and phase separately:*

$$\hat{A}_{\text{prop}}(\lambda,\mu) = \sum_{i=1}^{N_{\text{K}}} \hat{a}_{i,\text{abs}}(\lambda,\mu)\left|\hat{S}_{\text{up}}(\lambda-i,\mu)\right| \quad (10)$$

$$\hat{\alpha}_{\text{prop}}(\lambda,\mu) = \sum_{i=1}^{N_{\text{K}}} \hat{a}_{i,\angle}(\lambda,\mu)\angle\left\{\hat{S}_{\text{up}}(\lambda-i,\mu)\right\} \quad (11)$$

$$\Rightarrow \hat{S}^{(1)}_{\text{prop}}(\lambda,\mu) = \hat{A}_{\text{prop}}(\lambda,\mu)\mathrm{e}^{j\hat{\alpha}_{\text{prop}}(\lambda,\mu)} \quad (12)$$

2) *Predict real and imaginary part separately:*

$$\hat{S}^{(2)}_{\text{prop}}(\lambda,\mu) = \sum_{i=1}^{N_{\text{K}}} \hat{a}_{i,\text{Re}}(\lambda,\mu)\text{Re}\left\{\hat{S}_{\text{up}}(\lambda-i,\mu)\right\} + j\sum_{i=1}^{N_{\text{K}}} \hat{a}_{i,\text{Im}}(\lambda,\mu)\text{Im}\left\{\hat{S}_{\text{up}}(\lambda-i,\mu)\right\} \quad (13)$$

3) *Apply complex AR coefficients:*

$$\hat{S}^{(3)}_{\text{prop}}(\lambda,\mu) = \sum_{i=1}^{N_{\text{K}}} \hat{a}_i(\lambda,\mu)\hat{S}_{\text{up}}(\lambda-i,\mu) \quad (14)$$

$\text{Re}\{\cdot\}$ and $\text{Im}\{\cdot\}$ denote real and imaginary part and $\angle\{\cdot\}$ represents the phase operator. For the computation of the AR coefficients, the minimization of the prediction error energy is used as optimization criterion. Hence, the real AR coefficients in Eqs. (10), (11) and (13) and also the complex coefficients[1] in Eq. (14) can be obtained by using the Yule-Walker equations [8], where the required autocorrelation vector and matrix are calculated from the past L_{AC} enhanced speech coefficients. In order to find out which one of these methods performs best, the prediction gain

$$G_P^{(j)} = \frac{\mathcal{E}\left\{|S(\lambda,\mu)|^2\right\}}{\mathcal{E}\left\{\left|S(\lambda,\mu) - \hat{S}^{(j)}_{\text{prop}}(\lambda,\mu)\right|^2\right\}}, \quad j\in\{1,2,3\} \quad (15)$$

was measured with the expectation operator $\mathcal{E}\{\cdot\}$. Here, *idealistic* conditions were assumed, i.e., the prediction was based on clean speech coefficients and ideal AR coefficients determined from the previous L_{AC} clean samples. Fig. 2 depicts the results over the model order N_{K}. The data is obtained from about 30 minutes of samples selected randomly from the NTT speech database (sampling frequency $f_s = 8$ kHz). Moreover, the frame size was set to 20 ms ($L_{\text{F}} = 160$), the shift size to 5 ms and $L_{\text{AC}} = 8$ was used.

The results show that the highest prediction gain is obtained by using complex AR coefficients. Note that even negative values are

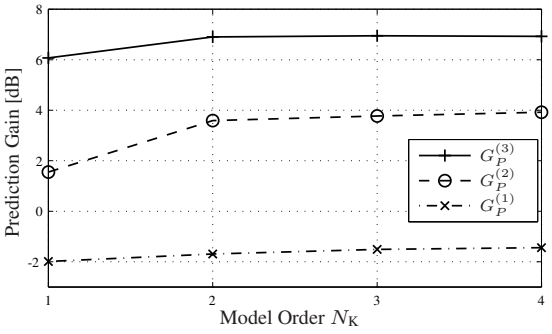

Fig. 2. Prediction gain for clean speech

[1] The derivation of the complex AR coefficients can be carried out analogue to the real case.

achieved when magnitude and phase are predicted separately. This is due to the fact that there is almost no correlation in successive phase samples. On the basis of these results, complex AR coefficients determined from previous enhanced speech coefficients were used to estimate the current speech coefficient in the propagation step, following Eq. (14).

3.2. Update Step

While in the propagation step the temporal correlation of the speech signal is exploited, the update step utilizes the statistical characteristics of speech and noise. The objective in this step is to estimate the prediction error E of the propagation step. Reorganizing Eq. (4) and using Eq. (2), it can be shown that the differential signal D consists of the prediction error degraded by the initial noise signal N:

$$D(\lambda,\mu) = S(\lambda,\mu) + N(\lambda,\mu) - \hat{S}_{\text{prop}}(\lambda,\mu) \tag{16}$$
$$= E(\lambda,\mu) + N(\lambda,\mu). \tag{17}$$

Thus, the task of the update step eases to the classical noise reduction problem: Decomposition of the noisy input sample D into the new wanted coefficient E and the noise coefficient N. Therefore, a conventional statistical estimator can be applied which is adapted to the statistics of E and N.

In the following, E and N are assumed to be statistically independent. Whereas a Gaussian signal model is considered for the noise coefficients, the statistical distribution of the prediction error is investigated by means of histogram measurements. Fig. 3 depicts the histogram of the real part of the DFT coefficients, averaged over approximately 2 hours of speech taken from the NTT speech database after normalization to $\mathcal{E}\{|\text{Re}\{E(\lambda,\mu)\}|^2\} = 1$ along with the analytic Gaussian, Laplacian and two-sided Gamma probability density functions (PDFs). It can be seen that the PDF of $\text{Re}\{E\}$ lies somewhere between a Gaussian and Laplacian PDF. The same distribution holds for the imaginary part. Based on these results, three suitable spectral estimators are proposed for the calculation of the weighting gains $K(\lambda,\mu)$ which are briefly described in the following:

a) *Gaussian MMSE Estimator/Wiener Filter*
This Gaussian *minimum mean square error* (MMSE) estimator corresponds to the Wiener Filter solution and is derived from the optimal filter theory [1]. It is a linear estimator that minimizes the mean square error between clean and enhanced coefficient. Applied to the update step, the enhanced coefficient $\hat{E}(\lambda,\mu)$ can be stated as:

$$\hat{E}(\lambda,\mu) = \mathcal{E}\big\{E(\lambda,\mu)|D(\lambda,\mu)\big\}. \tag{18}$$

Note that this MMSE estimator in the update step equals the conventional Kalman filter gain as it arises from the same assumption that prediction error and noise are Gaussian distributed [5].

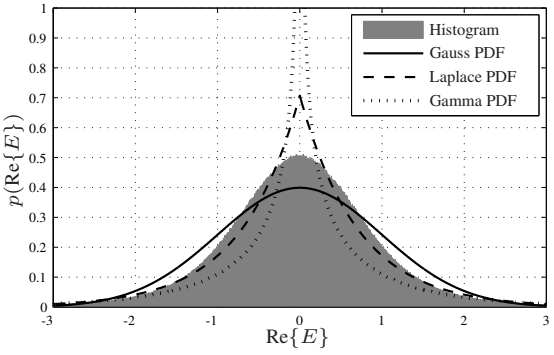

Fig. 3. Histogram of prediction error E

b) *Supergaussian MMSE Estimator*
In [2], the authors propose an MMSE estimator that uses a Laplacian PDF for the statistics of the wanted signal and a Gaussian model for the noise signal. This approach can be utilized here and a separate estimator for real and imaginary part of $E(\lambda,\mu)$ is obtained under the assumption that real and imaginary parts of $D(\lambda,\mu)$, $E(\lambda,\mu)$ and $N(\lambda,\mu)$ are statistically independent:

$$\hat{E}(\lambda,\mu) = \mathcal{E}\big\{\text{Re}\{E(\lambda,\mu)\}|\text{Re}\{D(\lambda,\mu)\}\big\} + j\mathcal{E}\big\{\text{Im}\{E(\lambda,\mu)\}|\text{Im}\{D(\lambda,\mu)\}\big\}. \tag{19}$$

c) *Supergaussian Joint MAP Estimator*
Applied to the update step, this more generalized supergaussian estimator [3] uses a parametric function to approximate the PDF of the spectral magnitude $|E|$. The Kullback-Leibler distance between measured and modeled PDF is used to obtain the optimal approximation [3]. In combination with a Gaussian noise model, this *maximum a posteriori* (MAP) estimator jointly maximizes the a posteriori PDF of amplitude *and* phase of the prediction error E, given the noisy sample D:

$$|\hat{E}| = \arg\max_{|E|} p\left(|E|, \angle\{E\}|D\right) \tag{20}$$

$$\angle\{\hat{E}\} = \arg\max_{\angle\{E\}} p\left(|E|, \angle\{E\}|D\right). \tag{21}$$

All the afore mentioned estimators require the *a posteriori* SNR γ and the *a priori SNR* ξ as input parameters. For the application within the update step they are defined as follows:

$$\gamma(\lambda,\mu) = \frac{|D(\lambda,\mu)|^2}{\hat{\sigma}_N^2(\lambda,\mu)} \quad \text{and} \quad \xi(\lambda,\mu) = \frac{\mathcal{E}\{|E(\lambda,\mu)|^2\}}{\hat{\sigma}_N^2(\lambda,\mu)}. \tag{22}$$

While the noise power spectral density (PSD) $\hat{\sigma}_N^2(\lambda,\mu)$ can be estimated, e.g., using [9], the a priori SNR is usually estimated using the recursive *decision-directed* approach [10].

4. EVALUATION

For the evaluation, five different noise types (babble, car, f16, factory, white) from the NOISEX-92 database were added to three male and two female speech sequences (each with a length of 8 s taken from the NTT speech database) at an input SNR varying between -10 dB and 35 dB with an increment of 5 dB. Investigated estimators were the weighting rules introduced above: Gaussian MMSE estimator/Wiener Filter, supergaussian MMSE estimator (MMSE-LapGauss) and supergaussian joint MAP estimator (JMAP). On the one hand, these suppression rules were used as purely statistical estimators, directly applied to the noisy input signal and on the other hand, they were embedded in the proposed Kalman Filter structure within the update step. According to Fig. 2, complex prediction with $N_{\text{K}} = 3$ and $L_{\text{AC}} = 8$ was used in the prior propagation step. Moreover, the transformation in the frequency domain was obtained by using 75% overlapping Hann analysis windows of 20 ms length for all investigated techniques. The required a priori SNR and noise PSD estimation was performed by the decision-directed approach [10] and minimum statistics [9], respectively.

In the simulation, the speech and noise signal can be filtered separately with weighting gains adapted for the noisy signal. Hence, the output signal can additionally be stated as $\hat{s}(k) = \tilde{s}(k) + \tilde{n}(k)$, where $\tilde{s}(k)$ is merely the filtered speech signal and $\tilde{n}(k)$ the filtered noise signal. Based on these quantities, the segmental speech SNR (SegSSNR), the segmental noise attenuation (NA) and the segmental speech attenuation (SA) were calculated (e.g., Chap. 4 in [11]).

Figs. 4 and 5 depict the averaged results for SA and SegSSNR, respectively, both plotted over NA with the input SNR as control

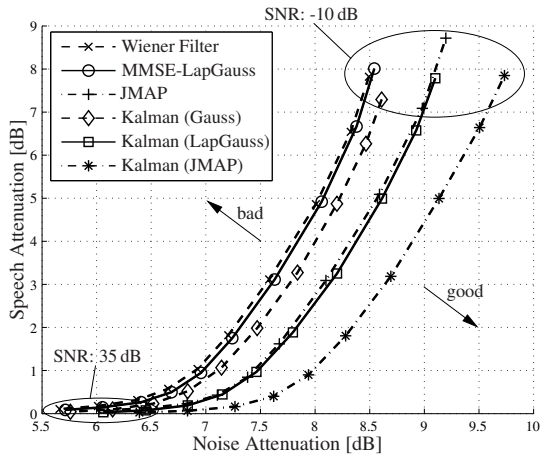

Fig. 4. Speech attenuation vs. noise attenuation

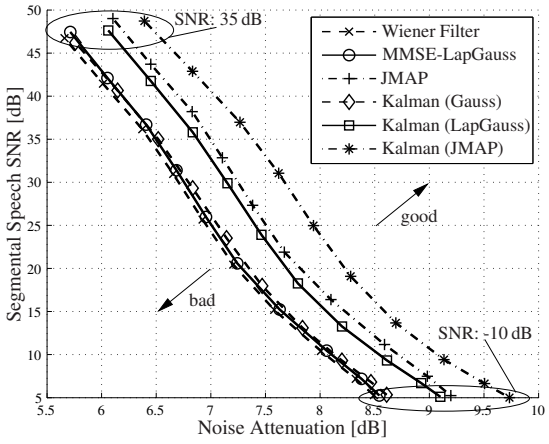

Fig. 5. Segmental speech SNR vs. noise attenuation

variable. Thus, a fair comparison with respect to the tradeoff between noise attenuation and speech distortion is possible. In Fig. 4, the points of best performance would be placed in the lower right corner, in Fig. 5, in the upper right corner. A lower a priori SNR threshold was applied to all estimators in a way that they yield nearly the same SA at 5 dB input SNR (cf. Fig. 4).

While keeping the SA and SegSSNR constant, it can be seen that the Kalman Filter approaches achieve a higher NA than the corresponding, purely statistical estimators. Comparing the three Kalman Filters, the use of supergaussian PDFs to model the statistics of the prediction error yields better results than the conventional Kalman (Gauss) Filter gain. Especially the utilization of the JMAP weighting rule outperforms all other approaches. The achieved results correspond to the subjectively perceived speech quality (informal listening tests). Furthermore, the gain that was achieved due to the exploitation of the temporal correlation in the propagation step was investigated. Fig. 6 depicts the effective prediction gain of all three proposed estimators over the input SNR, compared to the clean speech case (cf. Fig. 2). It can be seen that the proposed system already starts to benefit from the propagation step at -10 dB input SNR and reaches the level of ideal prediction nearly at 25 dB.

Even though the computational complexity is moderately increased by the proposed estimators, the evaluation clearly shows the advantages of these modified Kalman Filters and motivates further investigations in combined model-based and statistical approaches.

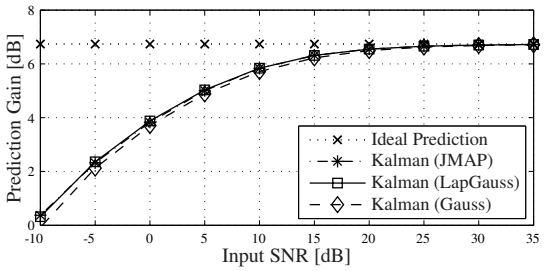

Fig. 6. Effective prediction gain

5. CONCLUSIONS

In this contribution, a modified Kalman Filter for single-channel speech enhancement is presented. The approach is based on a two step operation. In the first step, temporal correlation of successive speech frames is exploited by using complex linear prediction. In the second step, the first prediction is updated utilizing the statistical characteristics of the error signal. Here, not only the conventional Kalman Filter gain (relying on a Gaussian model for speech and noise) is taken into account, but also statistical estimators adapted explicitly to the PDF of the prediction error signal. Instrumental measurements have shown that the proposed modified Kalman Filters outperform the purely statistical estimators in terms of speech/ noise attenuation and segmental speech SNR. Moreover, the use of a supergaussian PDF to model the statistics of the prediction error has yielded better results than the conventional Kalman Filter gain. The results have been confirmed by informal listening tests.

6. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[2] R. Martin and C. Breithaupt, "Speech Enhancement in the DFT Domain Using Laplacian Speech Priors," in *Proc. of IWAENC*, Kyoto, Japan, 2003.

[3] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, pp. 1110–1126, 2005.

[4] K. Paliwal and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering," in *Proc. of ICASSP*, Dallas, USA, 1987.

[5] W.-R. Wu and P.-C. Chen, "Subband Kalman Filtering for Speech Enhancement," *IEEE Transactions on Circuits and Systems II*, vol. 45, no. 8, pp. 1072–1083, Aug. 1998.

[6] H. Puder, "Kalman-Filters in Subbands for Noise Reduction with Enhanced Pitch-Adaptive Speech Model Estimation," *European Transactions on Telecommunications*, vol. 13, no. 2, pp. 139–148, 2002.

[7] E. Zavarehei and S. Vaseghi, "Speech Enhancement in Temporal DFT Trajectories Using Kalman Filters," in *Proc. of INTERSPEECH*, Lisbon, Portugal, 2005.

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

[9] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 501–512, 2001.

[10] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transaction on Acoustic, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[11] J. Benesty, S. Makino, and J. Chen, Eds., *Speech Enhancement*. Springer, Berlin, 2005.