# MODEL-BASED SPEECH ENHANCEMENT USING SNR DEPENDENT MMSE ESTIMATION

*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 contribution presents a modified Kalman filter approach for single channel speech enhancement which is operating in the frequency domain. In the first step, temporal correlation of successive frames is exploited yielding estimates of the current speech and noise DFT coefficients. This first prediction is updated in the second step applying an SNR dependent MMSE estimator which is adapted to the (measured) statistics of the speech prediction error signal. Objective measurements show consistent improvements compared to estimators which do not take into account the temporal correlation or the influence of the input SNR on the statistics of the prediction error signal.

***Index Terms—*** Speech enhancement, noise reduction, Kalman filter, generalized Gamma distribution

## 1. INTRODUCTION

The problem of improving the speech quality of modern communication devices in noisy environments remains challenging and is still an active field of research, even though many techniques have been introduced in the past decades.

One of the popular methods for enhancing degraded speech is based on applying individual adaptive gains to the noisy input coefficients in the short-time Fourier transform (STFT) domain. Most of the rules proposed in literature have been derived under certain assumptions about the statistics of the speech and noise signal. While a Gaussian model is often used for the noise signal, the distribution of the speech signal is typically modeled either as Gaussian [1, 2] or as super-Gaussian [3, 4, 5]. Except for smoothing purposes as in [2], these statistical approaches only rely on memory-less a priori knowledge.

In contrast to the statistical estimators, the Kalman filter performs optimal estimation in linear dynamic systems in which a non-stationary target signal is disturbed by additive noise. The authors of [6] were the first who proposed the use of a Kalman filter for the purpose of speech enhancement. Compared to the common Wiener filtering method, the performance of this model-based approach was shown to be considerably better. In order to reduce complexity, the authors of [7] introduced a Kalman filtering system in the sub-band domain that additionally achieved better results than the full-band time domain approach. In [8], the application of a Kalman filter in sub-bands was further investigated and improved. In addition to the exploitation of intra-frame correlation, model-based approaches that consider the correlation of successive speech frames can be found, e.g., in [9, 10].

In this paper, a modified *Kalman filter* approach is considered which is applied in the frequency domain to the complex-valued discrete Fourier transform (DFT) coefficients. The proposed system consists of two steps, namely *propagation* and *update step*. It is shown that the input signal-to-noise ratio (SNR) influences the statistics of the speech prediction error signal in the propagation step. In contrast to [10], this characteristic is taken into account in the update step by using an SNR dependent minimum mean square error (MMSE) estimator that relies on generalized Gamma priors. The remainder of this paper is organized as follows: In Sec. 2, a brief overview of the considered system is given. Section 3 investigates the influence of the input SNR on the statistics of the speech prediction error signal and proposes the application of an adaptive weighting rule exploiting the SNR dependency. Experimental results are shown in Sec. 4 and conclusions are drawn in Sec. 5.

## 2. SYSTEM OVERVIEW

Figure 1 illustrates a simplified block diagram of the proposed system. It is assumed that the noisy input signal $y(k)$ consists of the clean speech signal $s(k)$ which is degraded by an additive noise signal $n(k)$ according to:

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

where $k$ is the discrete time index. For the decomposition of the speech and the noise signal, the noisy signal $y(k)$ is segmented into overlapping frames and is transformed into the frequency domain. Therefore, the fast Fourier transform (FFT) is applied to these frames after windowing and zero-padding. Hence, the spectral coefficients of the noisy input signal at frequency bin $\mu$ and frame $\lambda$ are given by:

$$Y(\lambda, \mu) = S(\lambda, \mu) + N(\lambda, \mu), \tag{2}$$

where $S(\lambda, \mu)$ and $N(\lambda, \mu)$ represent the spectral coefficients of the speech and the noise signal.

The proposed system is based on a Kalman filter structure which is applied to the complex-valued DFT coefficients $Y(\lambda, \mu)$, cf. [10]. Therefore, a two step approach is used consisting of a propagation and an update step. In contrast to [10], where both steps are carried out only for the speech signal, the propagation step in this system is extended to the noise signal in order to additionally take into account correlated noise signals.

In the propagation step, temporal correlation (a priori information of higher order) of successive frames is exploited. The current DFT coefficients of speech $S(\lambda, \mu)$ and noise $N(\lambda, \mu)$ are propagated in time based on information taken from previous, enhanced samples using linear prediction techniques. The resulting estimates

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

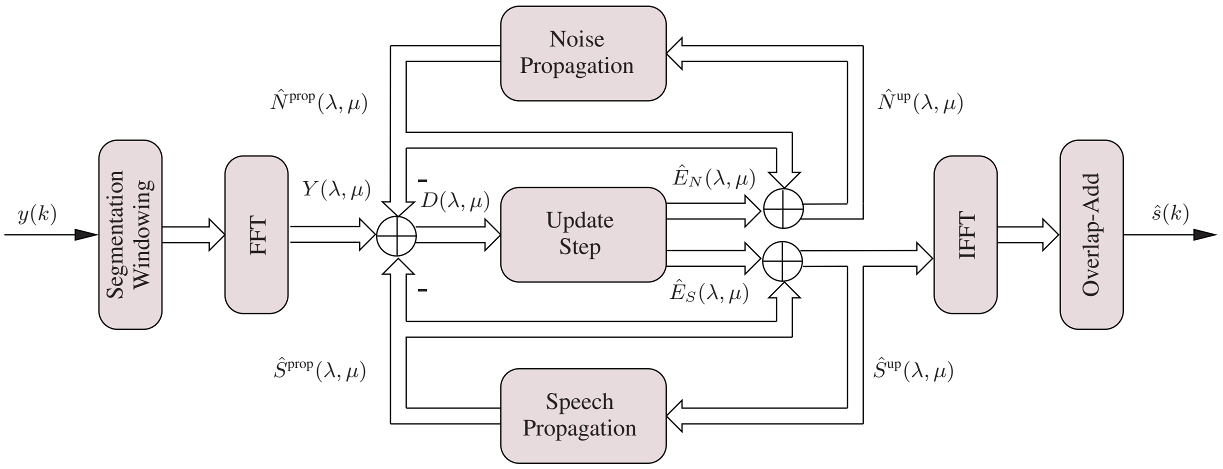

**Fig. 1**. Block diagram of the proposed Kalman filter structure.

$\hat{S}^{\text{prop}}(\lambda,\mu)$ and $\hat{N}^{\text{prop}}(\lambda,\mu)$ can be stated as:

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

$$\hat{N}^{\text{prop}}(\lambda,\mu) = \sum_{i=1}^{M_{\text{K}}} \hat{b}_i(\lambda,\mu)\hat{N}^{\text{up}}(\lambda-i,\mu), \tag{4}$$

where $N_{\text{K}}$ and $M_{\text{K}}$ represent the orders of the speech and the noise models, respectively. The problem of estimating the required autoregressive (AR) coefficients $a_i(\lambda,\mu)$ and $b_i(\lambda,\mu)$ in noisy environments has been extensively studied in literature, e.g., [11]. In this work, a simpler approach is used which minimizes the energies of the prediction errors. Therefore, the well-known Levinson-Durbin algorithm is used, e.g., [12]. The required autocorrelation vectors and matrices are calculated from the previous $L_{\text{AC}}$ enhanced DFT coefficients of either speech or noise.

The estimates $\hat{S}^{\text{prop}}(\lambda,\mu)$ and $\hat{N}^{\text{prop}}(\lambda,\mu)$ are summed up to get an estimation of the current noisy DFT coefficients:

$$\hat{Y}^{\text{prop}}(\lambda,\mu) = \hat{S}^{\text{prop}}(\lambda,\mu) + \hat{N}^{\text{prop}}(\lambda,\mu). \tag{5}$$

The prediction in the propagation step generally is erroneous, resulting in the following prediction errors:

$$E_{\text{S}}(\lambda,\mu) = S(\lambda,\mu) - \hat{S}^{\text{prop}}(\lambda,\mu) \quad \text{and} \tag{6}$$

$$E_{\text{N}}(\lambda,\mu) = N(\lambda,\mu) - \hat{N}^{\text{prop}}(\lambda,\mu) \tag{7}$$

for the speech and the noise, respectively. The objective in the following update step is to estimate the prediction errors $E_{\text{S}}(\lambda,\mu)$ and $E_{\text{N}}(\lambda,\mu)$ based on the differential signal $D(\lambda,\mu)$:

$$D(\lambda,\mu) = Y(\lambda,\mu) - \hat{Y}^{\text{prop}}(\lambda,\mu). \tag{8}$$

As shown in [10], the differential signal $D(\lambda,\mu)$ consists of the sum of the two prediction errors:

$$D(\lambda,\mu) = E_{\text{S}}(\lambda,\mu) + E_{\text{N}}(\lambda,\mu). \tag{9}$$

Hence, the estimation problem in the update step reduces to a 'classical' noise reduction problem: The target signal $E_{\text{S}}(\lambda,\mu)$ is degraded by the additive 'noise' signal $E_{\text{N}}(\lambda,\mu)$. For the decomposition of the 'noisy' signal $D(\lambda,\mu)$, a conventional statistical estimator, e.g., the well-known Wiener filter [1] can be used. In [10], the use of estimators relying on super-Gaussian speech models is evaluated within the update step. Therefore, a spectral weighting gain $G(\lambda,\mu)$ is determined for each frequency bin of each frame that is multiplied with the differential signal in order to estimate the two prediction errors:

$$\hat{E}_{\text{S}}(\lambda,\mu) = G(\lambda,\mu) \cdot D(\lambda,\mu) \tag{10}$$

$$\hat{E}_{\text{N}}(\lambda,\mu) = (1 - G(\lambda,\mu)) \cdot D(\lambda,\mu). \tag{11}$$

To obtain the final enhanced DFT coefficients $\hat{S}^{\text{up}}(\lambda,\mu)$ and $\hat{N}^{\text{up}}(\lambda,\mu)$, the initial predictions of the propagation step are updated:

$$\hat{S}^{\text{up}}(\lambda,\mu) = \hat{S}^{\text{prop}}(\lambda,\mu) + \hat{E}_{\text{S}}(\lambda,\mu) \tag{12}$$

$$\hat{N}^{\text{up}}(\lambda,\mu) = \hat{N}^{\text{prop}}(\lambda,\mu) + \hat{E}_{\text{N}}(\lambda,\mu). \tag{13}$$

The application of an inverse fast Fourier transform (IFFT) and the overlap-add method yield an estimate of the enhanced output signal $\hat{s}(k)$ in the time domain.

## 3. SNR INFLUENCE ON STATISTICS OF PREDICTION ERROR SIGNAL

It was shown in [10] that the prediction gain within the propagation step is depending on the input SNR: the higher the SNR, the higher the prediction gain. Therefore, it seems obvious that the input SNR also influences the statistics of the prediction error signals, which can be exploited within the update step by using an appropriate statistical, SNR dependent estimator. In order to keep the algorithm general and to not become dependent on a particular noise signal, it is still assumed that the noise prediction error coefficients $E_{\text{N}}$ follow a complex Gaussian distribution and only the statistics of the speech prediction error signal $E_{\text{S}}$ are investigated in the following.

For the evaluation, the Kalman filter system of Sec. 2 based on the Wiener filter applied in the update step was used. Depending on the input SNR, which was varied in the range from -20 dB to 35 dB (step size: 5 dB), the histogram of the speech prediction error $E_{\text{S}}$ was measured. For this, about 1 hour of speech (taken randomly from the NTT database) was disturbed by white Gaussian noise. Fig. 2 shows the measured histograms of the absolute value of $E_{\text{S}}$ for different SNR values. The magnitudes have been normalized to a power of $\sigma^2_{E_S} = 1$ to illustrate the dependencies of the shape

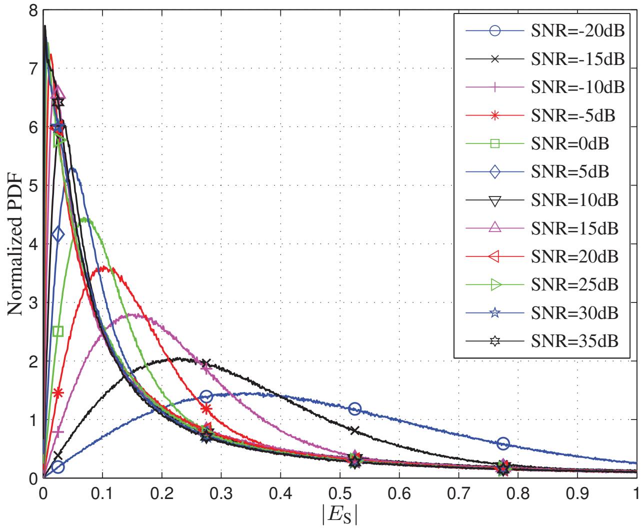

**Fig. 2**. Normalized histograms of $|E_\mathrm{S}|$ dependent on input SNR.

of the PDF on the input SNR. The SNR dependency can clearly be seen. The steepness of the respective probability functions (PDFs) around zero is getting larger for higher input SNR values showing that smaller prediction error values occur proportionally more often at higher SNR values. This behavior goes along with the fact that the prediction in the propagation step performs better the higher the input SNR is.

In order to exploit the SNR dependency, the complex DFT estimator of [5] is used within the update step. This MMSE estimator relies on a (complex) Gaussian distribution for the noise signal (here: $E_N$). For the magnitude of the target signal $E_S$, the following single-sided generalized Gamma density is assumed [5]:

$$p_{|E_S|}(x) = \frac{\gamma \delta^\nu}{\Gamma(\nu)} x^{\gamma\nu-1} \exp(-\delta x^\gamma) \tag{14}$$

with $\delta > 0$, $\gamma > 0$, $\nu > 0$ and $0 \leq x < \infty$. $\Gamma(\cdot)$ represents the Gamma function and $\gamma, \delta$ and $\nu$ are model parameters which can be adjusted according to the measured histograms. Thereby the parameter $\delta$ depends on $\gamma$, $\nu$ and $\sigma^2_{E_S}$. Several special cases are included in Eq. 14, e.g., a Rayleigh or a Gamma PDF. In [5], the complex DFT estimator is derived for the cases $\gamma = 1$ and $\gamma = 2$ and computes the conditional expectation $\mathcal{E}\{E_\mathrm{S}|D\}$ which directly can be used here in the update step to obtain the weighting gains required in Eqs. 10 and 11.

In order to get a good approximation to the measured histograms of Fig. 2, the Kullback Leibler distance [4] between modeled and measured PDFs is minimized. The resulting parameter settings are given in Tab. 1 and contribute to different MMSE estimators dependent on the input SNR.

In the simulations, the averaged and quantized a priori SNR estimates of the previous $N_\mathrm{K}$ frames decide which parameter settings are used in the current frame. The decision is made individually for each frequency bin.

## 4. RESULTS

The investigation is based on four different noise suppression techniques. The purely statistical weighting rules Wiener filter [1] and super-Gaussian MMSE estimator [3] on the one hand are compared with two Kalman filter techniques based on the system presented in Sec. 2 on the other hand. In the first Kalman filter approach, the SNR dependent MMSE estimator (Kalman SNR dependent) as illustrated in Sec. 3 was applied in the update step. For comparison, the complex DFT estimator of [5] was also adapted to the statistics of the prediction error signal $E_\mathrm{s}$ independent of the input SNR. Therefore, the normalized data which was recorded for the separate evaluation at different SNR values in Sec. 3 was merged in order to obtain an overall histogram of $|E_\mathrm{s}|$, which does not reflect its dependency on the input SNR. This measured histogram was also approximated by the model PDF of Eq. 14 (resulting in $\gamma = 1$ and $\nu = 0.64$) and used within the MMSE estimator [5] in the update step independent of the input SNR (Kalman SNR independent).

| SNR [dB] | $\leq -20$ | -15 | -10 | -5 | 0 | 5 |
|---|---|---|---|---|---|---|
| $\gamma$ | 1 | 1 | 1 | 1 | 1 | 1 |
| $\nu$ | 1.41 | 1.05 | 0.87 | 0.76 | 0.72 | 0.67 |

| SNR [dB] | 10 | 15 | 20 | 25 | 30 | $\geq 35$ |
|---|---|---|---|---|---|---|
| $\gamma$ | 1 | 1 | 1 | 1 | 1 | 1 |
| $\nu$ | 0.63 | 0.60 | 0.57 | 0.54 | 0.52 | 0.50 |

**Table 1**. Parameter settings for complex DFT estimator.

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 and noise attenuation (SA and NA) and the segmental speech SNR (SegSSNR) were calculated according to [13]. The main parameter settings that were used in the simulations are listed in Tab. 2. Five speech signals (3 male, 2 female, each 8 s length) from the NTT speech database were each degraded by six different noise sequences (f16, babble, car, factory1, factory2, white) taken from the NOISEX-92 database.

The averaged results are depicted in Figs. 3 and 4. Figure 3 shows the difference between noise and speech attenuation over the input SNR and Fig. 4 the segmental speech SNR plotted over the noise attenuation with the input SNR as control variable. In Fig. 3, a higher score indicates a better performance in which a value greater than 0 dB justifies the application of noise suppression. In Fig. 4, a high SegSSNR and a high NA is desirable.

The results show that both Kalman filter approaches yield a better performance than the corresponding statistical estimators espe-

| *Parameter* | *Settings* |
|---|---|
| Sampling frequency | 8 kHz |
| Frame length | 160 (20 ms) |
| FFT length | 256 (including zero-padding) |
| Frame overlap | 75% (Hann window) |
| Input SNR | -10 dB ... 35 dB (step size: 5 dB) |

| *Propagation Step* | |
|---|---|
| AC length $L_\mathrm{AC}$ | 6 |
| Model order $N_\mathrm{K}$ | 3 |
| Model order $M_\mathrm{K}$ | 2 |

| *Update Step* | |
|---|---|
| Noise estimation | Minimum Statistics [14] |
| SNR estimation | Decision-directed approach [2] |

**Table 2**. System settings.

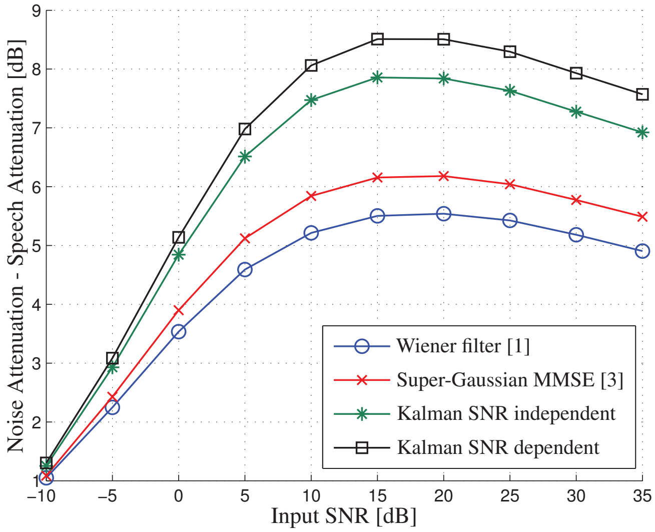

**Fig. 3**. Difference between noise attenuation and speech attenuation plotted over input SNR.

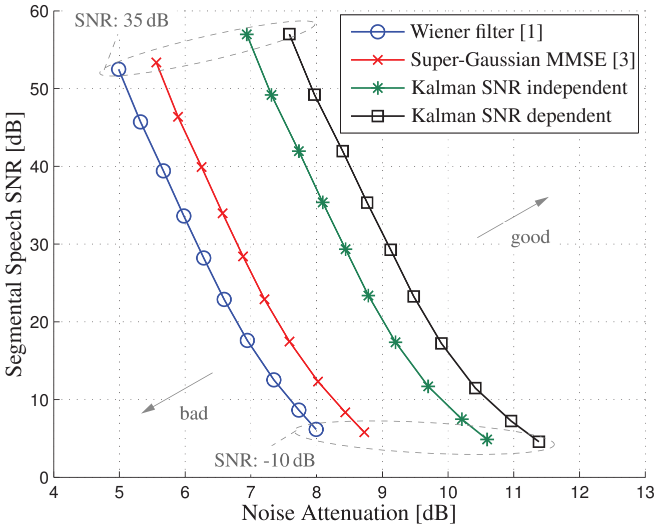

**Fig. 4**. Segmental speech SNR plotted over noise attenuation.

cially for input SNR values greater than 0 dB. Compared to the SNR independent MMSE estimator applied in the update step, the new approach achieves better results in terms of noise attenuation and speech distortion for the entire SNR range. Thus, the additional exploitation of the prediction error SNR dependency leads to a further improvement. The proposed approach yields the best compromise between speech and noise attenuation in Fig. 3 and the highest noise attenuation if the SegSSNR is kept constant in Fig. 4. The instrumental measurements were confirmed by informal listening tests.

## 5. CONCLUSIONS

The noise reduction system proposed in this paper uses a modified Kalman filter approach in the frequency domain. The method considers the SNR dependency of the resulting speech prediction error signal by modeling the measured histograms separately for several quantized SNR values with generalized gamma priors. A complex DFT estimator which is applied in the update step exploits these SNR dependent statistics and shows a better performance compared to the Wiener filter [1], the super-Gaussian MMSE estimator [3] and the corresponding Kalman filter system which is not adapted to the SNR dependency.

## 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] 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.

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

[4] 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.

[5] J. S. Erkelens, R. C. Hendriks, and R. Heusdens, “On the Estimation of Complex Speech DFT Coefficients Without Assuming Independent Real and Imaginary Parts,” *IEEE Signal Processing Letters*, vol. 15, pp. 213–216, 2008.

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

[7] 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, 1998.

[8] 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.

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

[10] T. Esch and P. Vary, “Speech Enhancement Using a Modified Kalman Filter Based on Complex Linear Prediction and Supergaussian Priors,” in *Proc. of ICASSP*, Las Vegas, USA, 2008.

[11] M. Kuropatwinski and W.B. Kleijn, “Estimation of the Short-Term Predictor Parameters of Speech Under Noisy Conditions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1645 –1655, 2006.

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

[13] S. Gustafsson, R. Martin, P. Jax, and P. Vary, “A Psychoacoustic Approach to Combined Acoustic Echo Cancellation and Noise Reduction,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, 2002.

[14] 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.