Fodor *et al. EURASIP Journal on Advances in Signal Processing* (2015) 2015:13
DOI 10.1186/s13634-015-0201-6

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH **Open Access**

Linking speech enhancement and error concealment based on recursive MMSE estimation

Balázs Fodor*, Florian Pflug and Tim Fingscheidt

Abstract

Speech enhancement and error concealment have seen a considerable progress over the past decades. Although both fields deal with distorted speech signals, there has rarely been an attempt to relate respective approaches to each other. In this paper, for the first time, a clear synopsis of recursive minimum mean square error (MMSE) estimation in both fields is provided. Our work intentionally does not propose a certain algorithm furthering the state of the art, nor does it provide simulation results of algorithms. Instead, our aim is threefold: First we revisit the basics of Bayes estimation in a recursive manner, covering both kinds of distortion acoustic noise as well as transmission channel noise. Second, we present recursive MMSE estimation applied to speech enhancement (in the frequency domain, as typical) and applied to error concealment (in the time domain, as typical) in strictly coherent notations and provide respective overview diagrams. Finally, we discuss commonalities and differences between both approaches, identify a particular strength of error concealment in general, and provide possible research directions for speech enhancement. A particularly interesting observation is that noise introduced by error concealment is far from being Gaussian and that additive acoustic noise can be expressed in terms of bit errors in DFT coefficients providing a potential interface to error concealment approaches.

Keywords: MMSE estimation; Speech enhancement; Error concealment

1 Introduction

Minimum mean square error (MMSE) estimation is omnipresent in a wide range of research fields and applications. It belongs to the family of Bayesian estimators which are based on the following model: The unobservable quantity to be estimated is considered to be the outcome of a random process such as a speech sample [1,2]. These outcomes can be measured through a channel introducing distortions, resulting in so-called observations. Besides the observations, Bayes estimators use *a priori* knowledge about the aforementioned random process and the channel, resulting in improved estimation results [2].

Based on this model, MMSE estimators minimize the estimation error variance conditioned on the observations. As an example, the widely used Wiener filter is optimal with respect to the MMSE error criterion [2]. In speech enhancement, it is widely assumed that the observations are statistically independent of each other, therefore, MMSE estimation of speech is carried out sequentially by means of the current observation only [3]. Signal history, such as the last speech estimate, is merely used for smoothing purposes in a practical system (cf. [3], Section V). Assuming, however, a dynamic signal model in the form of an autoregressive (AR) speech process, e. g., in conjunction with the source-filter model [4,5], the optimal MMSE estimator is able to exploit signal redundancy by employing both the current and previous observations [6]. In this case, under certain assumptions, the estimation can be carried out *recursively* and can be split into two steps typically decreasing computational complexity and relaxing memory requirements [7]. The first step exploits signal history in the form of previous observations providing an *a priori* estimate which is subsequently corrected in a second step taking into account

*Correspondence: b.fodor@tu-bs.de
Technische Universität Braunschweig, Institute for Communications Technology, Schleinitzstr. 22, 38106, Braunschweig, Germany

also the current observation, resulting in an *a posteriori* estimate.

In speech enhancement, the following model is widely employed for MMSE estimation: The unobservable speech is distorted by the unobservable acoustic noise, resulting in the observations in the form of noisy speech. The aim of speech enhancement is to estimate the speech by means of some *a priori* knowledge and the observations. State-of-the-art speech enhancement is often carried out in the short-time Fourier transform (STFT) domain and it is assumed that the discrete Fourier transform (DFT) coefficients of the speech and the acoustic noise are statistically independent. Thus, the *a priori* knowledge employed for MMSE estimation is a specific distribution of the speech (Gaussian [3,8,9], super-Gaussian [10-12]) and the acoustic noise (typically Gaussian). Classical MMSE estimators for speech enhancement under a Gaussian speech assumption are, e. g., the Wiener filter (e. g., [13]), the MMSE short-time speech amplitude (STSA) estimator [3], or the MMSE log-spectral amplitude (LSA) estimator [9]. These estimators mainly differ with respect to the estimation domain, i. e., instead of the complex-valued DFT coefficients, an arbitrary function of it is estimated (typically its amplitude or the logarithm of its amplitude).

In speech enhancement, the widely used Kalman filter [6] can be employed as a *recursive* MMSE estimator. In Kalman filtering, a Gaussian assumption is made for the acoustic noise and the error of the *a priori* speech estimate. In [14], a Kalman filter was proposed for time-domain speech enhancement. The proposed approach models the speech as an AR process based on a predictor, therefore, the estimation employs the current and the previous observations and, according to Kalman theory, the estimation of the speech is carried out recursively in two steps. A DFT-domain Kalman filter for speech enhancement was introduced, e. g., in [15-17]. In [15], the Kalman filter operates on complex-valued DFT coefficients and the speech was modeled as an AR process, while the noise process was assumed to be memoryless. In [16], different approaches to calculate the prediction coefficients and different estimators for calculating the *a posteriori* speech were investigated. Additionally, the memoryless assumption for the noise was replaced by an AR noise model in [17].

In error concealment, the following model is widely utilized: On the transmitter side, speech samples or source-coded parameters are quantized and mapped to corresponding bit combinations. These are transmitted over digital error-prone transmission channels and are received as bit-wise log-likelihood ratios (LLRs) comprising channel reliability information by the decoder. The aim of error concealment is the disguise of transmission errors which would otherwise lead to an unacceptable degradation of speech quality on the receiver side. This is achieved by MMSE estimation of unobservable speech samples or source-coded parameters, requiring *a posteriori* probabilities. These are proportional to both a likelihood term resulting from the received LLRs and a prior term comprising the available inherent signal redundancy as *a priori* knowledge at decoding time. This approach has been employed for robust source decoding of speech signals [18-23], source-coded audio signals [24,25], and uncompressed audio [26] that exploit signal redundancy in sample values or various source codec parameters (e. g., scaling factors, line spectral frequencies (LSFs) vectors, vector-quantized gains, adaptive codebook indices). Thereby, the signal redundancy is exploited by a time-variant modeling of the prior either using Markov chains [18-24] or employing approaches based on linear prediction in [25,26]. Typical applications are speech and audio transmission systems such as mobile phones or digital wireless microphones.

The aim of this paper is to reveal links between the fields of speech enhancement and error concealment focusing on recursive MMSE estimation approaches. We will show that the main structure of recursive MMSE estimation is the same in both disciplines which allows for drawing interesting links between tackling acoustic noise and transmission channel noise. In speech enhancement, the well-known Kalman filter can be employed as optimal recursive MMSE estimator which assumes that the acoustic noise is Gaussian distributed. Transmission channel noise as found in distorted speech signals, however, is far from a Gaussian distribution and — motivated by the nature of digital transmission — is rather modeled on *bit level* in error concealment which allows for exploiting powerful bit reliability information. This extra *a priori* knowledge can be identified as a definite strength of error concealment compared to speech enhancement. Based on this finding, this paper sketches as outlook new research directions for speech enhancement exploiting bit likelihoods.

This paper is structured as follows: Section 2 gives an introduction to recursive MMSE estimation. Section 3 shows how recursive MMSE estimation is commonly used for *speech enhancement* in the STFT domain. Section 4 gives an example for employing recursive MMSE estimation in *error concealment* in the time domain, presented in strong analogy to Section 3. Section 5 discusses links between speech enhancement and error concealment based on the recursive MMSE approaches from the previous Sections 3 and 4. In addition, some new research recommendations for speech enhancement motivated by error concealment are sketched. Finally, Section 6 closes the paper with conclusions.

2 On recursive MMSE estimation

As pointed out in Section 1, the aim of recursive MMSE estimation is to estimate the unobservable speech samples $s(n)$, with n being the discrete time index, which are transmitted through either an acoustic or a transmission channel. This channel distorts the speech signal by superimposing unobservable noise samples $d(n)$ being modeled as statistically independent, resulting in the observed noisy speech signal (cf. Figure 1)[a]

$$y(n) = s(n) + d(n). \tag{1}$$

The estimation of the clean speech $s(n)$ is carried out by means of all previous and the current observations $\mathbf{y}_0^n = \left[y(0), y(1), \ldots, y(n-1), y(n)\right]^T$, with $(\cdot)^T$ denoting the transpose operation, as well as by some *a priori* knowledge about the speech signal and the channel, resulting in the clean speech estimate $\hat{s}(n)$ (cf. Figure 1).

The *a priori* knowledge about the speech includes the following autoregressive model [14]: The current speech sample $s(n)$ is assumed to be a sum of the predicted speech $s^+(n)$ and the prediction error $e(n)$ (cf. Figure 1), the latter being a zero-mean (random) signal which is statistically independent of $s^+(n)$. The predicted speech is generated by a predictor of the order N_p as $s^+(n) = \mathbf{a}^T \cdot \mathbf{s}_{n-N_p}^{n-1}$ with $\mathbf{s}_{n-N_p}^{n-1} = [s(n-N_p), s(n-N_p+1), \ldots, s(n-1)]^T$ and $\mathbf{a} = \left[a_{N_p}, a_{N_p-1}, \ldots, a_1\right]^T$ being the so-called prediction coefficients. Please note that these coefficients are time-variant in practice, therefore, they need to be estimated. However, assuming a slow time variability, $\mathbf{a}$ is treated as a constant for the moment.

According to the speech signal model, the current speech sample $s(n)$ is statistically dependent not only on the current observation $y(n)$ but also on the previous ones $\mathbf{y}_0^{n-1}$, thus, these are also included in the estimation process [6]. This paper deals with *recursive* estimation, therefore, the estimation process is typically split into two parts, namely the *propagation* step and the *update* step. The propagation step exploits the previous observations to provide an *a priori* estimate of the current speech sample $s(n)$. Since this estimate can yield a relatively high error variance, the update step improves it incorporating the current observation $y(n)$, resulting in the *a posteriori* speech estimate $\hat{s}(n)$. Hence, the information carried by the previous observations becomes successively part of the *a priori* knowledge during the estimation process.

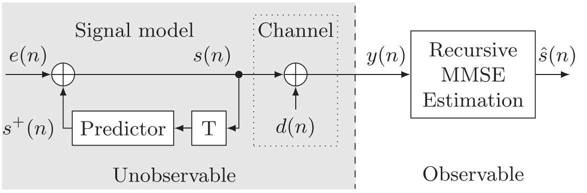

Figure 1 Signal model, channel, and recursive MMSE estimation.

2.1 The estimator

The recursive MMSE estimation with the underlying signal model as in Figure 1 yields

$$\hat{s}(n) = E\left\{s \,\middle|\, y(n), \mathbf{y}_0^{n-1}\right\} = \int_{\mathbb{R}} s \cdot p\left(s \,\middle|\, y(n), \mathbf{y}_0^{n-1}\right) ds \tag{2}$$

with $E\{\cdot\}$ being the expectation operator, $p(\cdot)$ being a probability density function (pdf), and $p(s(n)|y(n), \mathbf{y}_0^{n-1}) = p\left(s(n)|\mathbf{y}_0^n\right)$ being the so-called posterior. Usually, the posterior is computed by means of Bayes' rule, therefore, (2) can be rewritten as

$$\hat{s}(n) = \frac{\int_{\mathbb{R}} s \cdot p\left(y(n) \,\middle|\, s, \mathbf{y}_0^{n-1}\right) \cdot p\left(s \,\middle|\, \mathbf{y}_0^{n-1}\right) ds}{p\left(y(n) \,\middle|\, \mathbf{y}_0^{n-1}\right)} \tag{3}$$

with $p\left(y(n) \,\middle|\, s(n), \mathbf{y}_0^{n-1}\right)$, $p\left(s(n) \,\middle|\, \mathbf{y}_0^{n-1}\right)$, and $p\left(y(n) \,\middle|\, \mathbf{y}_0^{n-1}\right) = \int_{\mathbb{R}} p\left(y(n) \,\middle|\, s, \mathbf{y}_0^{n-1}\right) \cdot p\left(s \,\middle|\, \mathbf{y}_0^{n-1}\right) ds$ being the so-called likelihood, the prior, and the evidence, respectively. Please note that the evidence is typically calculated by marginalizing the pdf product of the numerator in (3).

2.2 The prior

As can be seen above, the prior is a function of the previous observations $\mathbf{y}_0^{n-1}$. However, the prior can also be determined recursively by marginalization using the signal model in Figure 1 as

$$p\left(s(n) \,\middle|\, \mathbf{y}_0^{n-1}\right) = \int \cdots \int_{\mathbb{R}^{N_p}} p\left(s(n) \,\middle|\, \mathbf{s}_{n-N_p}^{n-1}\right) \cdot p\left(\mathbf{s}_{n-N_p}^{n-1} \,\middle|\, \mathbf{y}_0^{n-1}\right) d\mathbf{s}_{n-N_p}^{n-1}. \tag{4}$$

The first pdf in the integral is a predictor pdf, the second one is the joint pdf of the last N_p posteriors. Therefore, the current prior is obviously dependent on the (distribution of the) last N_p estimates. Moreover, the mean of the prior using the signal model in Figure 1 turns out to be

$$\begin{aligned} E\left\{s(n) \,\middle|\, \mathbf{y}_0^{n-1}\right\} &= E\left\{e(n) \,\middle|\, \mathbf{y}_0^{n-1}\right\} + E\left\{s^+(n) \,\middle|\, \mathbf{y}_0^{n-1}\right\} \\ &= 0 + \mathbf{a}^T \cdot \begin{pmatrix} E\left\{s(n-N_p) \,\middle|\, \mathbf{y}_0^{n-N_p}\right\} \\ E\left\{s(n-N_p+1) \,\middle|\, \mathbf{y}_0^{n-N_p+1}\right\} \\ \vdots \\ E\left\{s(n-2) \,\middle|\, \mathbf{y}_0^{n-2}\right\} \\ E\left\{s(n-1) \,\middle|\, \mathbf{y}_0^{n-1}\right\} \end{pmatrix} \\ &= \mathbf{a}^T \cdot \hat{\mathbf{s}}_{n-N_p}^{n-1} = \hat{s}^+(n) \end{aligned} \tag{5}$$

which is the result of the propagation step. Please note that the prediction error $e(n) = s(n) - s^+(n)$ (also called innovation, e. g., [7]) is orthogonal to the previous speech samples and, therefore, also to the previous observations (e. g., [7,27]). Thus, we obtain $E\{e(n)|\mathbf{y}_0^{n-1}\} = E\{e(n)\} = 0$.

Assuming that this *a priori* speech estimate $\hat{s}^+(n) = f\left(\mathbf{y}_0^{n-1}\right)$ is a sufficient statistic for $s(n)$, the prior turns out to be

$$p(s(n)|\mathbf{y}_0^{n-1}) = p(s(n)|\hat{s}^+(n)). \tag{6}$$

This pdf describes the clean speech given the *a priori* speech estimate, in other words, it is the pdf of the propagation error $\bar{e}(n) = s(n) - \hat{s}^+(n)$ which can be written as

$$p\left(s(n)\left|\hat{s}^+(n)\right.\right) = p_{\bar{e}}\left(\bar{e}(n) = s(n) - \hat{s}^+(n)\right). \tag{7}$$

Using the signal model in Figure 1, the variance of the propagation error turns out to be [28]

$$\begin{aligned} E\{(s(n) - \hat{s}^+(n))^2\} &= E\{e^2(n)\} + E\{(s^+(n) - \hat{s}^+(n))^2\} \\ &= \sigma_e^2(n) + \sigma_{e^+}^2(n) = \sigma_{\bar{e}}^2(n) \end{aligned} \tag{8}$$

with $\sigma_e^2(n) = E\left\{e^2(n)\right\}$ being the prediction error variance and $\sigma_{e^+}^2(n) = E\left\{\left(s^+(n) - \hat{s}^+(n)\right)^2\right\}$. Please note that (8) holds *independently* of the type of the propagation error pdf.

2.3 The likelihood

The *a priori* knowledge about the (acoustic or transmission) channel is contained in the likelihood. This means that the additive noise $d(n)$, which distorts the clean speech $s(n)$ (cf. Figure 1), is modeled by the likelihood $p\left(y(n)\left|s, \mathbf{y}_0^{n-1}\right.\right)$. Assuming a memoryless (acoustic or transmission) channel, meaning that the noise samples $d(n)$, $d(n-1)$, ... are statistically independent of each other, the likelihood is [29]

$$p(y(n)|s(n), \mathbf{y}_0^{n-1}) = p(y(n)|s(n)) \tag{9}$$

being the pdf of $y(n)$ given a specific speech sample $s(n)$. Therefore, this is the pdf of the additive noise $d(n) = y(n) - s(n)$ as [3]

$$p\left(y(n)\left|s(n)\right.\right) = p_d\left(d(n) = y(n) - s(n)\right). \tag{10}$$

Since the variance of the noise $\sigma_d^2(n) = E\left\{d^2(n)\right\}$ cannot be measured directly in practice, it is usually estimated by a noise power estimator or a channel quality estimator.

3 Application to speech enhancement

Assuming that the speech and the noise processes are at least quasi-stationary along the time frames ℓ in each frequency bin k and statistically independent of each other, the signal model in Figure 1 is also valid in the STFT domain. Accordingly, the aim of recursive MMSE estimation in speech enhancement is to estimate the clean speech DFT coefficients $S_\ell(k)$ which are generated by an autoregressive process as sketched in Section 2: The clean speech $S_\ell(k)$ is modeled by the sum of the predicted speech $S_\ell^+(k)$ and a statistically independent, zero-mean prediction error $E_\ell(k)$, i. e., $S_\ell(k) = S_\ell^+(k) + E_\ell(k)$. $S_\ell^+(k)$ is calculated by a predictor of the order L_p as $S_\ell^+(k) = \mathbf{A}^H(k) \cdot \mathbf{S}_{\ell-L_p}^{\ell-1}(k)$ with $\mathbf{A}(k) = \left[A_{L_p}(k), A_{L_p-1}(k), \ldots, A_1(k)\right]^T$ being the complex-valued prediction coefficients and $\mathbf{S}_{\ell-L_p}^{\ell-1}(k) = \left[S_{\ell-L_p}(k), S_{\ell-L_p+1}(k), \ldots, S_{\ell-1}(k)\right]^T$. It is assumed that the prediction coefficients change very slowly along the frames ℓ and are, therefore, modeled as constants for the moment. The acoustic channel distorts the speech spectrum by superimposing some statistically independent acoustic noise $D_\ell(k)$, so that $Y_\ell(k) = S_\ell(k) + D_\ell(k)$ with $Y_\ell(k)$ being the noisy speech DFT coefficients. The estimated speech DFT coefficients $\widehat{S}_\ell(k)$ are calculated by the recursive MMSE estimator employing the underlying observations $\mathbf{Y}_0^\ell(k) = [Y_0(k), Y_1(k), \ldots, Y_\ell(k)]^T$ as (cf. (2))

$$\widehat{S}_\ell(k) = E\left\{S_\ell(k)\left|Y_\ell(k), \mathbf{Y}_0^{\ell-1}(k)\right.\right\}. \tag{11}$$

A block diagram of recursive MMSE estimation for speech enhancement assuming a memoryless acoustic channel is given in Figure 2. Please note that the upper signal path is related to the prior computation, the block in the center is the MMSE estimator (11), and the lower signal path refers to the likelihood computation. Starting in the lower left-hand corner, windowed segments of the noisy speech signal $y(n) = s(n) + d(n)$ are transformed into the DFT domain, followed by the likelihood computation.

3.1 The likelihood

Since a memoryless acoustic channel is assumed, the likelihood turns out to be (cf. (9) and (10))

$$\begin{aligned} p\left(Y_\ell(k)\left|S_\ell(k), \mathbf{Y}_0^{\ell-1}(k)\right.\right) &= p\left(Y_\ell(k)\left|S_\ell(k)\right.\right) \\ &= p_D\left(D_\ell(k) = Y_\ell(k) - S_\ell(k)\right). \end{aligned} \tag{12}$$

As can be seen, the likelihood is a function of the noisy speech (cf. Figure 2). Moreover, after computing the likelihood by means of the current observation $Y_\ell(k)$, the likelihood remains a function $g(S)$ of the unknown speech DFT coefficient S (cf. output of 'Likelihood Computation' in Figure 2). Furthermore, as we will see later, S will be the integration variable of the MMSE estimator (cf. (11)).

Assuming that the (complex-valued) additive noise $D_\ell(k)$ is a zero-mean Gaussian process with independent and identically distributed (i. i. d.) real and imaginary parts, the likelihood turns out to be [3]

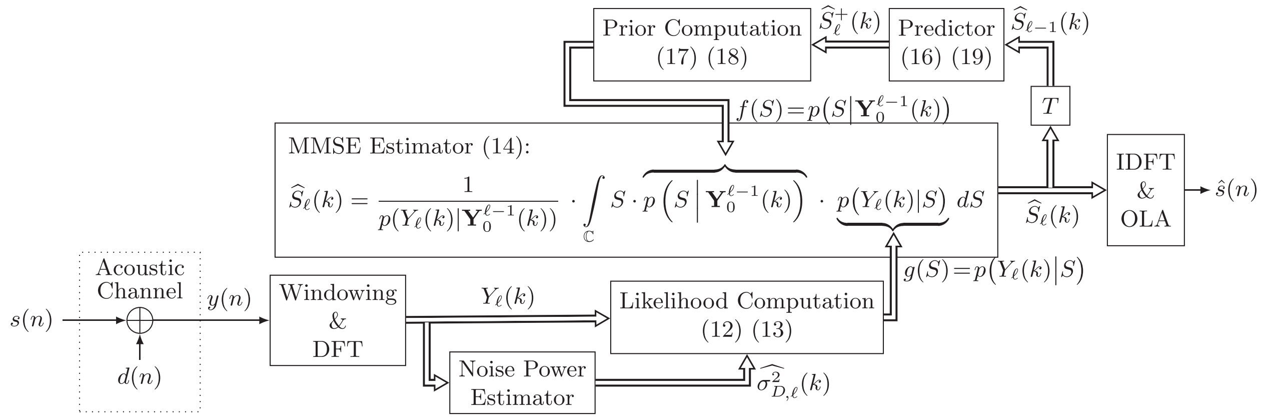
Prior Computation (17) (18)
$\widehat{S}_\ell^+(k)$
Predictor (16) (19)
$\widehat{S}_{\ell-1}(k)$
T
$f(S) = p(S|\mathbf{Y}_0^{\ell-1}(k))$
MMSE Estimator (14):
$\widehat{S}_\ell(k) = \frac{1}{p(Y_\ell(k)|\mathbf{Y}_0^{\ell-1}(k))} \cdot \int_{\mathbb{C}} S \cdot \overbrace{p\left(S \middle| \mathbf{Y}_0^{\ell-1}(k)\right)} \cdot \underbrace{p(Y_\ell(k)|S)} \, dS$
$\widehat{S}_\ell(k)$
IDFT & OLA
$\hat{s}(n)$
Acoustic Channel
$s(n)$
$y(n)$
$d(n)$
Windowing & DFT
$Y_\ell(k)$
Likelihood Computation (12) (13)
$g(S) = p(Y_\ell(k)|S)$
Noise Power Estimator
$\widehat{\sigma_{D,\ell}^2}(k)$

Figure 2 STFT domain recursive MMSE estimation for speech enhancement assuming a memoryless acoustic channel.

$$p(Y_\ell(k)|S_\ell(k)) = \frac{1}{\pi\sigma_{D,\ell}^2(k)} \cdot \exp\left(-\frac{|Y_\ell(k) - S_\ell(k)|^2}{\sigma_{D,\ell}^2(k)}\right) \quad (13)$$

with $\sigma_{D,\ell}^2(k)$ being the variance of the quasi-stationary noise process D. Please note that also for non-Gaussian noise pdfs, the likelihood is always a function of the noise power $\sigma_{D,\ell}^2(k)$ (cf. connection between 'Noise Power Estimator' and 'Likelihood Computation' in Figure 2). Therefore, in practice, its estimate $\widehat{\sigma_{D,\ell}^2}(k)$ is calculated by a noise power estimator using the noisy speech DFT coefficients [30-32].

3.2 The estimator

Using the likelihood (12), the clean speech DFT coefficients $\widehat{S}_\ell(k)$ are estimated as (cf. (2), (3), and (11))

$$\widehat{S}_\ell(k) = \frac{\int_{\mathbb{C}} S \cdot p(S|\mathbf{Y}_0^{\ell-1}(k)) \cdot p(Y_\ell(k)|S)\, dS}{p(Y_\ell(k)|\mathbf{Y}_0^{\ell-1}(k))} \quad (14)$$

with the evidence

$$p(Y_\ell(k)|\mathbf{Y}_0^{\ell-1}(k)) = \int_{\mathbb{C}} p(S|\mathbf{Y}_0^{\ell-1}(k)) \cdot p(Y_\ell(k)|S)\, dS \quad (15)$$

and the prior $p\left(S \middle| \mathbf{Y}_0^{\ell-1}(k)\right)$. Please note that both the prior and the likelihood are a function of the integration variable S, namely $f(S)$ (cf. upper signal path in Figure 2) and $g(S)$ (cf. lower signal path in Figure 2), respectively. The estimated clean speech signal $\hat{s}(n)$ is obtained by taking the inverse DFT (IDFT) of $\widehat{S}_\ell(k)$ from (14) and performing, e. g., an overlap-add (OLA) step.

3.3 The prior

As discussed in Section 2, the *a priori* speech estimate is calculated as (cf. (5))

$$\widehat{S}_\ell^+(k) = \mathbf{A}^H(k) \cdot \widehat{\mathbf{S}}_{\ell-L_p}^{\ell-1}(k). \quad (16)$$

This step is denoted as 'Predictor' in the upper signal path in Figure 2. Please note that the predictor employs previous speech estimates which is reflected by the delay unit denoted by 'T' in Figure 2. Assuming that the *a priori* speech estimate $\widehat{S}_\ell^+(k)$ is a sufficient statistic for the speech $S_\ell(k)$, the prior turns out to be the pdf of the propagation error $\bar{E}_\ell(k) = S_\ell(k) - \widehat{S}_\ell^+(k)$ (cf. (6) and (7))

$$p(S_\ell(k)|\mathbf{Y}_0^{\ell-1}(k)) = p_{\bar{E}}(\bar{E}_\ell(k) = S_\ell(k) - \widehat{S}_\ell^+(k)). \quad (17)$$

Employing a specific *a priori* speech estimate $\widehat{S}_\ell^+(k)$, the prior remains a function of the speech $f(S)$ and can be fed into the MMSE estimator (14) (cf. connection between 'Prior Computation' and 'MMSE Estimator' in Figure 2).

Assuming that the (complex-valued) propagation error is a zero-mean Gaussian process with i. i. d. real and imaginary parts, the prior is calculated [15,16]

$$p_{\bar{E}}\left(\bar{E}_\ell(k)\right) = \frac{1}{\pi\sigma_{\bar{E},\ell}^2(k)} \cdot \exp\left(-\frac{|\bar{E}_\ell(k)|^2}{\sigma_{\bar{E},\ell}^2(k)}\right) \quad (18)$$

with $\sigma_{\bar{E},\ell}^2(k)$ being the propagation error variance which cannot be measured in practice, thus, it has to be estimated [15,16].

Since the prediction coefficients $\mathbf{A}(k)$ in (16) are not accessible in practice, they need to be estimated as well, e. g., by the widely used normalized least-mean-squares (NLMS) algorithm [7]. Introducing again time variability, the prediction coefficients for the next frame are calculated recursively as

$$\widehat{\mathbf{A}}_{\ell+1}(k) = \widehat{\mathbf{A}}_\ell(k) + \mu \cdot \frac{\widehat{E}_\ell^*(k)}{||\widehat{\mathbf{S}}_{\ell-L_p}^{\ell-1}(k)||^2 + \Delta} \cdot \widehat{\mathbf{S}}_{\ell-L_p}^{\ell-1}(k) \quad (19)$$

with $\widehat{E}_\ell(k) = \widehat{S}_\ell(k) - \widehat{S}_\ell^+(k)$, as well as with μ, $(\cdot)^*$, Δ, and $||\cdot||$ denoting the step size constant, the complex conjugate, the regularization parameter, and the Euclidean norm, respectively.

3.4 The Kalman filter

The estimator (14) requires calculating two integrals over the whole complex plane for each time-frequency unit (ℓ, k). Fortunately, this estimator can be obtained in closed form by solving these integrals, reducing the computational complexity for practical implementations. Assuming a Gaussian distribution for both the propagation error (18) and the acoustic noise (13), the MMSE estimator (14) turns out to be a sum of the *a priori* estimate and the update (the derivation can be found in the Appendix) in the form of the Kalman filter equations (cf. [15], Equation 12)

$$\widehat{S}_\ell(k) = \widehat{S}_\ell^+(k) + \widehat{E}_\ell(k) \tag{20}$$

with

$$\widehat{E}_\ell(k) = K_\ell(k) \cdot R_\ell(k), \tag{21}$$

$$K_\ell(k) = \frac{\zeta_\ell(k)}{1 + \zeta_\ell(k)}, \tag{22}$$

$$R_\ell(k) = Y_\ell(k) - \widehat{S}_\ell^+(k) \tag{23}$$

where $K_\ell(k)$ is the so-called Kalman gain and $\zeta_\ell(k) = \sigma_{E,\ell}^2(k)/\sigma_{D,\ell}^2(k)$ (cf. the *a priori* SNR in [3]). The latter can be estimated by [16]

$$\hat{\zeta}_\ell(k) = \beta \frac{|\widehat{E}_{\ell-1}(k)|^2}{\widehat{\sigma_{D,\ell-1}^2}(k)} + (1-\beta) \max\left\{ \frac{|R_\ell(k)|^2}{\widehat{\sigma_{D,\ell-1}^2}(k)} - 1, 0 \right\} \tag{24}$$

with β being a smoothing factor, typically chosen close to one.

Please note that the recursive nature of MMSE estimation is well reflected by (20): The *a priori* estimate $\widehat{S}_\ell^+(k)$ utilizing the previous observations is corrected by the term $\widehat{E}_\ell(k)$ employing the current observation $Y_\ell(k)$, resulting in the *a posteriori* speech estimate $\widehat{S}_\ell(k)$.

4 Application to error concealment

The aim of error concealment is to estimate transmitter-sided (speech) samples, e. g., by a recursive MMSE estimator, in order to conceal distortions due to residual bit errors after demodulation or channel decoding. Bit error concealment of PCM audio or speech could theoretically be carried out by the equations in Section 2 using the hard-decoded receiver-sided samples. However, in order to exploit more information for improved estimation results, it is often advantageous to employ error concealment using reliability information on a bit level as in [19,26]. A block diagram of such a soft-decision decoding scheme based on recursive MMSE estimation is given in Figure 3. Similar to Figure 2, the likelihood computation is related to the lower signal path, the estimator can be found in the center, and the prior computation is performed in the upper signal path.

4.1 The likelihood

In error concealment, it is assumed that each transmitter-sided sample $s(n)$, being processed as introduced in Section 2, is quantized with M bit and, therefore, can bijectively be mapped to a natural-binary bit combination $\mathbf{x}(n) = [x_0(n), x_1(n), \ldots, x_m(n), \ldots, x_{M-1}(n)]$ (see 'Quantization and Bit Mapping' in Figure 3). Assuming further binary phase-shift keying (BPSK) modulation, each transmitted bit (BPSK symbol) $x_m(n) \in \{-1, 1\}$ is more or less distorted by the channel, modeled by the real-valued channel noise $d_m(n)$ (cf. 'Transmission Channel' in Figure 3). For the demodulation of the received real-valued noisy symbols $y_m(n)$, the so-called energy per bit to noise power spectral density ratio E_b/N_0 is needed which is calculated by the channel estimator (cf. connection between 'Channel Estimator' and 'Demodulator' in Figure 3). The demodulator then calculates the LLR which is defined as

$$L(\hat{x}_m(n)) = \ln \frac{P(\hat{x}_m(n)|x_m(n) = +1)}{P(\hat{x}_m(n)|x_m(n) = -1)} \tag{25}$$

with the hard-decided receiver-sided bit $\hat{x}_m(n) = \operatorname{sign}(y_m(n))$ representing the receiver-sided observation. In practice, the LLR is calculated by means of E_b/N_0 and $y_m(n)$ (cf. the two inputs of the 'Demodulator' in Figure 3)[b] which reflects the likelihood of a possibly transmitted bit $x_m(n)$. The bit-error probabilities $\mathrm{BER}_m(n)$ describe the probability that a transmitted bit was distorted through the channel and can be calculated by means of the LLRs as [33]

$$\mathrm{BER}_m(n) = \frac{1}{1 + e^{|L(\hat{x}_m(n))|}}. \tag{26}$$

Once each transmitter-sided sample $s(n)$ is quantized, it assumes a discrete value $s^{(i)}$ with $i \in \{0, 1, \ldots, 2^M - 1\}$. Moreover, each $s^{(i)}$ can be mapped to one corresponding bit combination $\mathbf{x}^{(i)}$. The bit-wise transition probabilities define the *bit* likelihood given bit $x_m^{(i)}$ of the ith quantization table entry [19]

$$P\left(\hat{x}_m(n) \,\middle|\, x_m^{(i)}\right) = \begin{cases} \mathrm{BER}_m(n), & \text{if } \hat{x}_m(n) \neq x_m^{(i)}, \\ 1 - \mathrm{BER}_m(n), & \text{else.} \end{cases} \tag{27}$$

Assuming that the transmission channel is memoryless and that the bit distortions $d_m(n)$ are statistically independent of each other along the bit indices m, the *sample* likelihood is computed as (cf. 'Likelihood Computation' in Figure 3 and, e. g., [19])

$$P\left(\hat{\mathbf{x}}(n) \,\middle|\, \mathbf{x}^{(i)}\right) = \prod_{m=0}^{M-1} P\left(\hat{x}_m(n) \,\middle|\, x_m^{(i)}\right). \tag{28}$$

Please note that until this step, the recursive MMSE estimator operates on bit level. After computing the *sample*

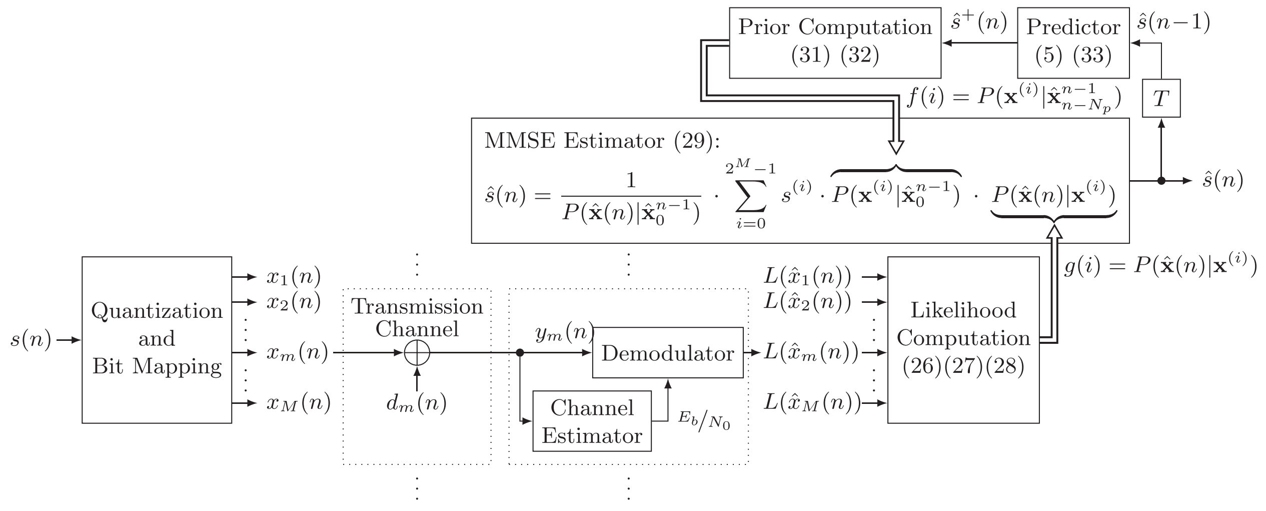
Prior Computation (31) (32)
$\hat{s}^+(n)$
Predictor (5) (33)
$\hat{s}(n-1)$
$f(i) = P(\mathbf{x}^{(i)}|\hat{\mathbf{x}}_{n-N_p}^{n-1})$
T
MMSE Estimator (29):
$\hat{s}(n) = \frac{1}{P(\hat{\mathbf{x}}(n)|\hat{\mathbf{x}}_0^{n-1})} \cdot \sum_{i=0}^{2^M-1} s^{(i)} \cdot \overbrace{P(\mathbf{x}^{(i)}|\hat{\mathbf{x}}_0^{n-1})} \cdot \underbrace{P(\hat{\mathbf{x}}(n)|\mathbf{x}^{(i)})}$
$\hat{s}(n)$
$g(i) = P(\hat{\mathbf{x}}(n)|\mathbf{x}^{(i)})$
$s(n)$
Quantization and Bit Mapping
$x_1(n)$
$x_2(n)$
$x_m(n)$
$x_M(n)$
Transmission Channel
$d_m(n)$
$y_m(n)$
Demodulator
Channel Estimator
E_b/N_0
$L(\hat{x}_1(n))$
$L(\hat{x}_2(n))$
$L(\hat{x}_m(n))$
$L(\hat{x}_M(n))$
Likelihood Computation (26)(27)(28)

Figure 3 Time domain recursive MMSE estimation for error concealment assuming a memoryless transmission channel.

likelihood (28) and for any further processing, however, the estimator deals with samples again.

4.2 The estimator

The recursive MMSE estimator (3) turns out to be a sum

$$\hat{s}(n) = \frac{\sum_{i=0}^{2^M-1} s^{(i)} \cdot P(\mathbf{x}^{(i)}|\hat{\mathbf{x}}_0^{n-1}) \cdot P(\hat{\mathbf{x}}(n)|\mathbf{x}^{(i)})}{P(\hat{\mathbf{x}}(n)|\hat{\mathbf{x}}_0^{n-1})} \tag{29}$$

with the evidence

$$P\left(\hat{\mathbf{x}}(n)\,\middle|\hat{\mathbf{x}}_0^{n-1}\right) = \sum_{i=0}^{2^M-1} P\left(\mathbf{x}^{(i)}\,\middle|\hat{\mathbf{x}}_0^{n-1}\right) \cdot P\left(\hat{\mathbf{x}}(n)\,\middle|\mathbf{x}^{(i)}\right), \tag{30}$$

the *sample* likelihood (28), and the prior $P\left(\mathbf{x}^{(i)}\,\middle|\hat{\mathbf{x}}_0^{n-1}\right)$. Please note that both the prior and the *sample* likelihood are a function of the summation index i in (29), namely $f(i)$ (cf. upper signal path in Figure 3) and $g(i)$ (cf. lower signal path in Figure 3), respectively. The result of the summation is the speech estimate $\hat{s}(n)$ (cf. right-hand side of the MMSE estimator in Figure 3).

4.3 The prior

As discussed in Section 2, the *a priori* speech estimate is calculated as $\hat{s}^+(n) = \mathbf{a}^T \cdot \hat{\mathbf{s}}_{n-N_p}^{n-1} = f\left(\hat{\mathbf{x}}_0^{n-1}\right)$ (cf. (5)). This step is carried out in the block 'Predictor' in Figure 3. Just as in Section 3, the predictor incorporates the previously estimated speech samples, reflected by the delay unit 'T' in Figure 3. Assuming that the *a priori* speech estimate $\hat{s}^+(n)$ is a sufficient statistic for $s^{(i)}$ and using that each bit combination $\mathbf{x}^{(i)}$ can bijectively be mapped to one corresponding $s^{(i)}$, the prior turns out to be

$$P\left(\mathbf{x}^{(i)}\,\middle|\hat{\mathbf{x}}_0^{n-1}\right) = P\left(s^{(i)}\,\middle|\hat{s}^+(n)\right). \tag{31}$$

However, since $s^{(i)}$ is a quantized quantity, the probability of its ith value can be calculated as [34]

$$P(s^{(i)}|\hat{s}^+(n)) = \int_{I_i} p_{\tilde{e}}(s - \hat{s}^+(n))\, ds \tag{32}$$

with $p_{\tilde{e}}(\cdot)$ being the propagation error pdf and with the PCM sample quantization intervals I_i, $i = 0, 1, \ldots, 2^M-1$. Employing a specific *a priori* speech estimate $\hat{s}^+(n)$, the prior remains a function of summation index i as $f(i)$ and can be fed into the MMSE estimator (29) (cf. connection between 'Prior Computation' and 'MMSE Estimator' in Figure 3).

Please note that in [26], the quantity $E\left\{\left(s^+(n) - \hat{s}^+(n)\right)^2\right\}$ in (8) was assumed to be zero. Furthermore, assuming that the prediction error $e(n) = s(n) - s^+(n)$ is stationary, the propagation error pdf can be determined by a histogram measurement in a training process. Moreover, online integration of $p_{\tilde{e}}(\cdot)$ is not necessary if the integrations over I_i intervals are performed beforehand and $\hat{s}^+(n)$ is quantized with M bits, leading to discrete probabilities $P_{\tilde{e}}(\cdot)$ [26]. Thus, employing a lookup table containing the precomputed $P_{\tilde{e}}(\cdot)$ values in the block 'Prior Computation' in Figure 3, the table entries can be indexed by the quantized *a priori* speech estimate $\hat{s}^{+(i)}(n)$. Hence, the resulting prior $P(s^{(i)}\,|\hat{s}^{+(i)}(n))$ being a function of the summation index i can be obtained in a computationally efficient way.

Introducing again time variability, the prediction coefficients $\mathbf{a}$ in (5) have to be estimated which can be done recursively, e. g., by means of the NLMS algorithm [7]

$$\hat{\mathbf{a}}(n+1) = \hat{\mathbf{a}}(n) + \mu \cdot \frac{\hat{e}(n)}{||\hat{\mathbf{s}}_{n-N_p}^{n-1}||^2 + \Delta} \cdot \hat{\mathbf{s}}_{n-N_p}^{n-1} \quad (33)$$

with $\hat{e}(n) = \hat{s}(n) - \hat{s}^{+}(n)$ as well as with μ and Δ being the step size constant and the regularization parameter, respectively. Please note that the prediction coefficients can alternatively be obtained by a slightly modified NLMS algorithm [26,35].

5 Links between speech enhancement and error concealment

So far, we have introduced an application example of recursive MMSE estimation for both speech enhancement and error concealment. In this section, we aim at showing links between the presented estimators.

As can be seen above, while the estimator (14) used for speech enhancement (cf. Figure 2) utilizes continuous distributions, the estimator (29) employed for error concealment deals with discrete ones (cf. Figure 3). In error concealment, the samples of the signal to be transmitted are *quantized* typically by 16 bit or 24 bit and thus are from a finite set of elements, whereas in speech enhancement the digital signals are assumed to be quantized fine enough (even though quantization with 64, 32, or 16 bit may take place), therefore, the codomain of the samples is assumed to be *continuous*.

5.1 The likelihood

This also influences the channel model. While in speech enhancement, the noise is modeled on a *sample* (or *coefficient*) level (1), in digital transmission and error concealment the noise occurs on a modulation symbol level or, for BPSK equivalently, on bit level (26), (27): The transmitted binary source bits (BPSK: symbols) $x_m(n)$ are distorted by the real-valued channel noise $d_m(n)$. Using the received value (symbol) $y_m(n)$, the demodulator calculates a corresponding LLR $L(\hat{x}_m(n))$ by means of E_b/N_0, which is a normalized SNR measure being a function of the channel noise power. Therefore, the likelihood (28) being computed by means of the LLRs (cf. (26), (27), and (28)) is a function of the channel noise power as it is the case in speech enhancement (cf. (13)). Thus, in both speech enhancement and error concealment, the current observation ($Y_\ell(k)$ or $y_m(n)$) and the channel noise power (density) ($\sigma^2_{D,\ell}(k)$ or N_0) are needed for likelihood computation (cf. Figures 2 and 3).

However, there remains a distinct difference between speech enhancement and error concealment concerning the estimation of the noise power. In speech enhancement, the noise power is estimated by means of the noisy speech often assuming that noise is more stationary than speech. Typical noise power estimators are, e. g., approaches based on minimum statistics (MS) [30], (improved) minima-controlled recursive averaging ((I)MCRA) [31,36], or approaches based on speech presence probability [32]. In digital transmission or error concealment, however, the amount of the noise is dependent on the distance between the received symbol and all possibly transmitted symbols in the constellation diagram, the latter having fixed positions depending on the modulation scheme. Thus, implicitly, in error concealment one has more information about possible channel inputs which is a clear advantage over speech enhancement.

The likelihood in speech enhancement (13) is usually modeled by a common pdf, typically by a Gaussian which is more or less justified by the central limit theorem [3]. In error concealment, however, the noise pdf does not follow any typical distribution and depends on the employed bit mapping. For the further discussion, we define the transmission channel noise in error concealment similar to the noise in speech enhancement: The noise $d(n)$ in error concealment will be the difference between the hard-decided speech samples and the transmitted (quantized) one. The histogram of the transmission channel noise turns out to be spiky as can be seen on the right-hand side of Figure 4 for 16 bit uniform PCM quantization, natural-binary bit mapping, and BPSK transmission. The bottom and top spikes in the histograms and the higher amplitudes in the waveforms are evoked by bit errors at bit positions close to the most significant bit (MSB).

In the case of acoustic noise, increasing the noise power (decreasing the SNR) naturally results in higher noise signal levels and an increasing width of the time-domain noise histogram (cf. car noise example on the left hand-side of Figure 4). In the case of the transmission channel noise, with increasing noise power (decreasing SNR), more bit errors occur, resulting in higher peaks belonging to $d \neq 0$ in the histogram on the right-hand side of Figure 4. However, this increase of noise power scales the height of all high spikes in the histogram (referring to single bit errors) belonging to $d \neq 0$ approximately equally (cf. right-hand side of Figure 4). Thus, while in the case of acoustic noise the noise power is typically associated with the *width* of the noise pdf, in error concealment, the noise power can be related to the *height* of spikes in the noise histogram belonging to $d \neq 0$ (here: for uniform PCM quantization, natural-binary bit mapping, and BPSK modulation).

5.2 The estimator

Although the structure of the estimators in Sections 3 and 4 is very similar (cf. Figures 2 and 3), their implementation may considerably differ. In speech enhancement, the clean speech is estimated by an integral over the whole complex plane (14). The online numerical computation of this integral is typically hard to manage in practice due to the two-dimensional pdfs, however, employing a common

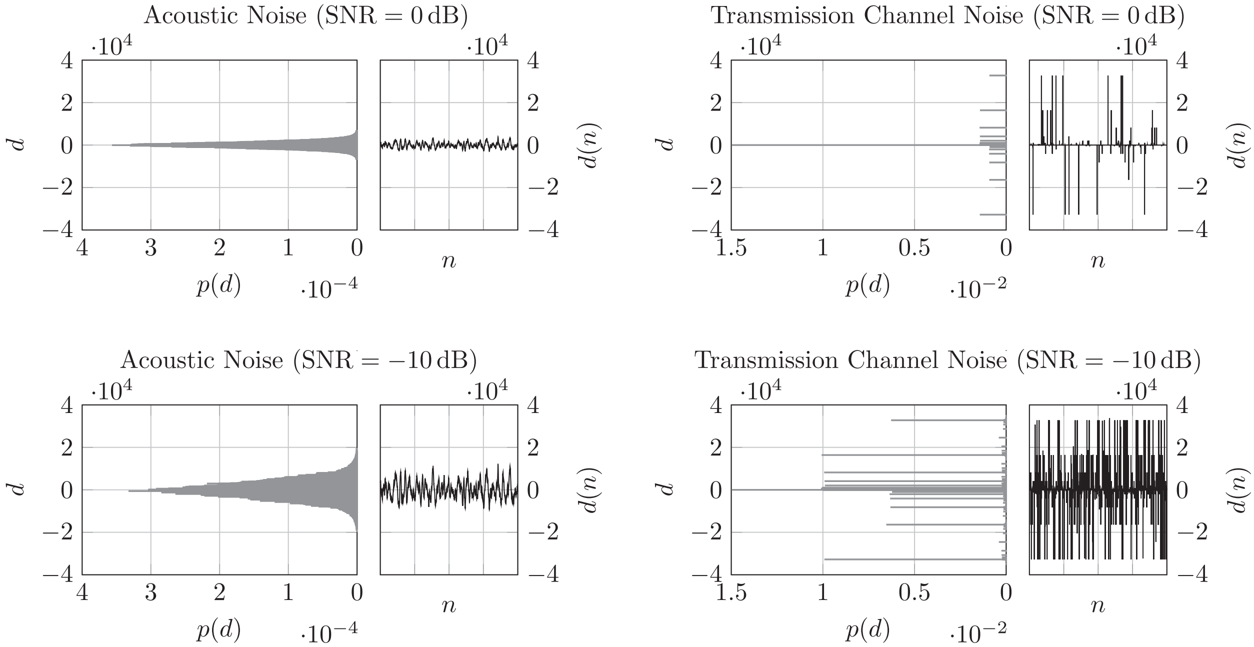

Figure 4 Normalized time domain histograms and waveforms of car noise and transmission channel noise. Left: normalized time domain histograms and waveforms of car noise; right: normalized time domain histograms and waveforms of transmission channel noise applied to 16 bit speech samples with underlying natural-binary bit mapping. The SNR values were calculated by means of a fixed speech signal level of −26 dB$_{\text{ov}}$ and respective noise signal levels, both measured according to ITU-T P.56 [37].

distribution for the prior and the likelihood allows for a closed-form solution. Accordingly, a Gaussian assumption for both the prior and the likelihood results in the Kalman filter Eqs. (20), (21), (22), (23) (cf. Section 3).

In error concealment, the clean speech estimator (29) is a sum due to quantization and the respective finite number of transmittable bit combinations. This estimator is computationally complex but manageable in practice due to the one-dimensional pdfs (cf. discrete terms (28), (31)). Thus, the sum (29) can explicitly be computed at runtime [26]. Interestingly, a closed-form solution of the sum cannot be achieved due to the likelihood which cannot be approximated by a common pdf, as outlined in Section 5.1.

5.3 The prior

As can be seen in Section 2, the prior is the pdf of the propagation error (7). In Section 3, it was assumed that the propagation error is Gaussian distributed, therefore, the prior is modeled by a bivariate Gaussian (18) with a complex-valued argument. In [15], the Gaussian assumption is justified with the tradeoff between mathematical manageability and pdf model mismatch. It was shown in [16] that the histogram of the propagation error DFT coefficients differ from a Gaussian (cf. left-hand side of Figure 5). Therefore, in [16] the histogram of the propagation error was measured and a parametric pdf was trained and then employed as prior. Furthermore, since the speech estimates strongly depend on the channel, the propagation error also depends on the channel. Accordingly, in [38] the SNR dependency of the propagation error histogram was reported and an SNR-dependent estimator was proposed.

Although in [26] the variance $E\left\{\left(s^+ - \hat{s}^+\right)^2\right\}$ is assumed to be zero, the non-Gaussianity of the propagation error pdf $p_{\bar{e}}\left(\bar{e} = s - \hat{s}^+\right)$ was also observed in the context of error concealment. As can be seen on the right-hand side of Figure 5, $p_{\bar{e}}$ in (32), being measured in a training process, turns out to be rather super-Gaussian. Furthermore, in [26] the dependency of $p_{\bar{e}}\left(\bar{e} = s - \hat{s}^+\right)$ on $\hat{s}^+$ was investigated revealing different shapes and variances dependent on the amplitude of $\hat{s}^+$, while $\hat{s}^+ = s^+$ was assumed there.

The propagation error in error concealment is quasi-continuous, while the prior is discrete. Thus, an integration step (32) is needed in order to discretize the propagation error pdf to obtain the prior. Fortunately, this discretization step can be done during a training process, resulting in a lookup table which nicely reduces the computational complexity [26].

Please note that usually the NLMS algorithm is employed for calculating the prediction coefficients both in speech enhancement and in error concealment due to its robustness and low computational complexity. However, there are other algorithms which can also be utilized,

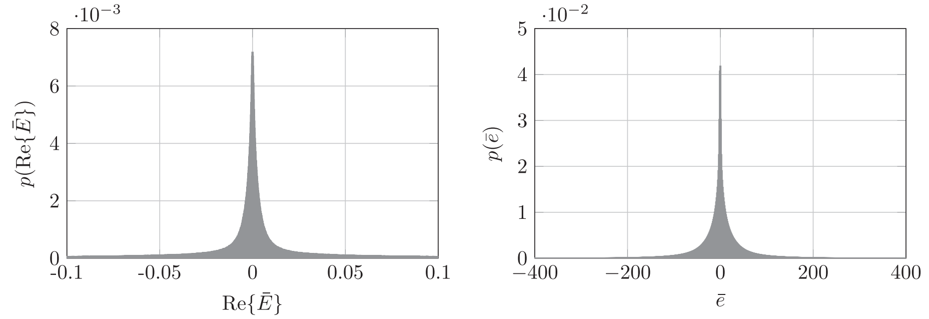

Figure 5 Propagation error histograms measured in a speech enhancement system and in an error concealment system. Left: normalized frequency domain histogram of the propagation error $\bar{E}_\ell(k) = S_\ell(k) - \widehat{S}_\ell^+(k)$ measured in a speech enhancement system from Section 3; right: normalized time domain histogram of the propagation error $\bar{e}(n) = s(n) - \hat{s}^+(n)$ measured in an error concealment system from Section 4 assuming that $E\{(s^+ - \hat{s}^+)^2\}$ is zero [26].

as reported in [16]. Please note that the propagation error is also dependent on the algorithm for determining the prediction coefficients, therefore, the change of the algorithm involves a new propagation error pdf training in both disciplines.

5.4 Outlook

In this section, we briefly sketch further possible research directions for speech enhancement inspired by error concealment. Since one of the key success factors in error concealment is that bit reliability information is exploited, the speech enhancement approach from Section 3 could benefit from using bit likelihoods. This means that instead of (13) the DFT coefficient likelihood (12) is calculated by means of bit likelihoods as in error concealment (cf. (27) and (28)).

In the following, we will estimate the real and imaginary parts of the complex-valued speech DFT coefficients separately, as in, e. g., [10]. Assuming that the resulting real-valued real and imaginary parts of STFT-domain quantities are quantized, their natural-binary representation is possible. In the following, we will introduce the processing steps for the real part only, denoted by superscript 're'. Of course, the imaginary part can be treated in the same way employing the same processing steps.

In order to be able to employ a bit-level model for the acoustic channel, we assume that the real part of the speech and the noisy speech DFT coefficients are quantized by M bit. Therefore, those quantities can bijectively be mapped into the bit combinations $\mathbf{S}_\ell^{\mathrm{re}}(k) = \left[S_{0,\ell}^{\mathrm{re}}(k), S_{1,\ell}^{\mathrm{re}}(k), \ldots, S_{m,\ell}^{\mathrm{re}}(k), \ldots, S_{M-1,\ell}^{\mathrm{re}}(k)\right]$ and $\mathbf{Y}_\ell^{\mathrm{re}}(k) = \left[Y_{0,\ell}^{\mathrm{re}}(k), Y_{1,\ell}^{\mathrm{re}}(k), \ldots, Y_{m,\ell}^{\mathrm{re}}(k), \ldots, Y_{M-1,\ell}^{\mathrm{re}}(k)\right]$, respectively. Due to the fact that speech is distorted by acoustic noise while passing through the acoustic channel, the observed bit combinations at the acoustic channel output $\mathbf{Y}_\ell^{\mathrm{re}}(k)$ may differ from those at the channel input $\mathbf{S}_\ell^{\mathrm{re}}(k)$. Accordingly, a bit error at bit position $m \in \{0, 1, \ldots, M{-}1\}$ occurs if the received bit $Y_{m,\ell}^{\mathrm{re}}(k)$ is not equal to the transmitted one $S_{m,\ell}^{\mathrm{re}}(k)$. The bit error rate $\mathrm{BER}_m^{\mathrm{re}}(k)$ can be measured within a training process by comparing $S_{m,\ell}^{\mathrm{re}}(k)$ to $Y_{m,\ell}^{\mathrm{re}}(k)$ for all bit positions m, individually for each frequency bin k. Please note that the bit errors also depend on the local SNR in the current time-frequency unit (ℓ, k), therefore, the SNR has to be taken into account during the training process. Accordingly, the training steps can be summarized as follows: Using speech and car noise data we generated noisy speech signals at different signal SNR levels. Then, we calculated the short-time spectra of the clean speech, the noise, and the noisy speech signals resulting in $S_\ell(k)$, $D_\ell(k)$, and $Y_\ell(k) = S_\ell(k) + D_\ell(k)$, respectively. Using the resulting DFT coefficients, we calculated the true speech power $\sigma_{S,\ell}^2(k) = E\left\{|S_\ell(k)|^2\right\}$ and the true noise power $\sigma_{D,\ell}^2(k) = E\left\{|D_\ell(k)|^2\right\}$. Using those two power spectra, we obtained the true *a priori* SNR as $\xi_\ell(k) = \sigma_{S,\ell}^2(k)/\sigma_{D,\ell}^2(k)$. Then, we quantized the real part of the clean speech and noisy speech DFT coefficients by 16 bit resulting in the bit combinations $\mathbf{S}_\ell^{\mathrm{re}}(k)$ and $\mathbf{Y}_\ell^{\mathrm{re}}(k)$, respectively. By this means, the whole training data was processed and $S_{m,\ell}^{\mathrm{re}}(k) \in \mathbf{S}_\ell^{\mathrm{re}}(k)$ and $Y_{m,\ell}^{\mathrm{re}}(k) \in \mathbf{Y}_\ell^{\mathrm{re}}(k)$ were compared to each other. Bit errors $S_{m,\ell}^{\mathrm{re}}(k) \neq Y_{m,\ell}^{\mathrm{re}}(k)$ were counted at bit position m, frequency bin k, and in dependence of the *a priori* SNR. For the latter, the ideal *a priori* SNR $\xi_\ell(k)$ was quantized resulting in discrete *a priori* SNR values ξ_q.

The resulting bit error rates $\mathrm{BER}_m^{\mathrm{re}}(k, \xi_q)$ were stored in a lookup table; examples can be seen in Figure 6: As expected, at higher SNRs less bits are in error. Typical for car noise, at higher frequencies less bits are in error as compared to lower frequencies.

A speech enhancement approach as described in Section 3 can be modified to include bit likelihoods which can be calculated by the bit error rates from the previous training step. The resulting frequency-dependent

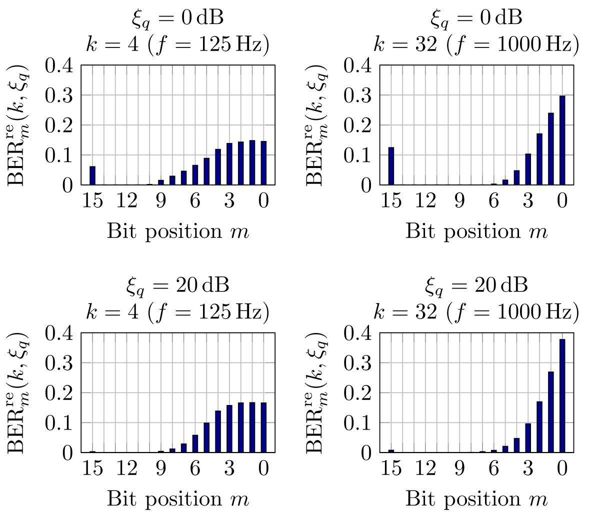

Figure 6 Bit error probabilities $\mathrm{BER}^{\mathrm{re}}_m(k, \xi_q)$ for different *a priori* SNRs ξ_q and frequencies k. Natural-binary bit representation with $M = 16$ bit quantization and the MSB (bit position $M-1 = 15$) being the sign bit is used.

lookup tables are then addressed by a quantized *a priori* SNR estimate $\hat{\xi}_{q,\ell}(k)$, computed, e. g., by the well-known decision-directed *a priori* SNR estimator [3] and the same quantization intervals as for the training process, resulting in $\mathrm{BER}^{\mathrm{re}}_m(k, \hat{\xi}_{q,\ell}(k))$ values. Since each transmitter-sided DFT coefficient $S_\ell(k)$ is quantized, it assumes a discrete value $S^{\mathrm{re},(i)}$ with $i \in \{0, 1, \ldots, 2^M-1\}$. Furthermore, each $S^{\mathrm{re},(i)}$ can be mapped to one corresponding bit combination $\mathbf{S}^{\mathrm{re},(i)}$. Then, using the resulting bit error rate $\mathrm{BER}^{\mathrm{re}}_m\left(k, \hat{\xi}_{q,\ell}(k)\right)$, the bit likelihood with the given bit $S^{\mathrm{re},(i)}_m \in \mathbf{S}^{\mathrm{re},(i)}$ of the ith quantization table entry can be obtained similar to (27) as

$$P\left(Y^{\mathrm{re}}_{m,\ell}(k)\middle|S^{\mathrm{re},(i)}_m\right)=\begin{cases}\mathrm{BER}^{\mathrm{re}}_{m,\ell}\left(k,\hat{\xi}_{q,\ell}(k)\right), & \text{if } Y^{\mathrm{re}}_{m,\ell}(k)\neq S^{\mathrm{re},(i)}_m,\\ 1-\mathrm{BER}^{\mathrm{re}}_{m,\ell}\left(k,\hat{\xi}_{q,\ell}(k)\right), & \text{else.}\end{cases} \tag{34}$$

The bit likelihood describes the probability of an observed bit $Y^{\mathrm{re}}_{m,\ell}(k)$ given a possible channel input $S^{\mathrm{re},(i)}_m$. The coefficient likelihood can be obtained using the bit likelihoods similar to (28) as

$$P\left(\mathbf{Y}^{\mathrm{re}}_\ell(k)\,\middle|\,\mathbf{S}^{\mathrm{re},(i)}\right)=\prod_{m=0}^{M-1}P\left(Y^{\mathrm{re}}_{m,\ell}(k)\,\middle|\,S^{\mathrm{re},(i)}_m\right). \tag{35}$$

By this means, the simple Gaussian assumption for the noise DFT coefficients (13) can be replaced by such a new approach which allows for an environment-specific processing (cf. [39]). This is also illustrated in Figure 7 for an *a priori* SNR of 20 dB and the frequency bin $k = 32$ (corresponding to $f = 1\,\mathrm{kHz}$): As can be seen, the resulting likelihood is a sharp pdf unlike a Gaussian pdf. The Gaussian assumption for noise is typically justified with the central limit theorem assuming that the span of correlation of the noise samples is sufficiently short compared to the frame length [40]. Although this assumption is better fulfilled by a wide range of noise types than by speech signals, it can generally be said that it is not fulfilled perfectly in practice by noise signals. Accordingly, the likelihood turns out to be a sharper pdf than a Gaussian such as in Figure 7. Please note that the non-Gaussianity of noise DFT coefficients was also reported in [41]. Furthermore, there are publications dealing with speech estimators based on a non-Gaussian assumption for the noise, e. g., [10,12]. Moreover, it was shown in these papers that a more realistic likelihood function (just as the proposed one obtained by training) offers a more precise (acoustic) channel model which can improve estimation results.

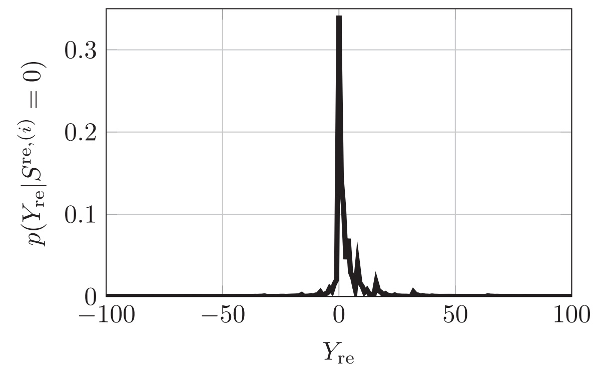

Figure 7 Coefficient likelihood example. Coefficient likelihood $P\left(\mathbf{Y}^{\mathrm{re}}_\ell(k)\middle|S^{\mathrm{re},(i)}\right)$ for the specific channel input $S^{\mathrm{re},(i)} = 0$ calculated by (34) and (35) using trained bit error probabilities $\mathrm{BER}^{\mathrm{re}}_m(k, \xi_q)$.

The speech prior $P\left(\mathbf{S}^{\mathrm{re},(i)}\mid S^{+,\mathrm{re}}_\ell(k)\right)$ (cf. (17)) can be obtained by integrating (18) according to the PCM quantization intervals (i) (cf. (32)). Here, $S^{+,\mathrm{re}}_\ell(k)$ is the real part of the *a priori* speech estimate $S^+_\ell(k)$ calculated by, e. g., (16). Using the coefficient likelihood (35) and a speech prior, the recursive MMSE estimation formula turns out to be (cf. (14) and (29))

$$\widehat{S}^{\mathrm{re}}_\ell(k)=\frac{\sum\limits_{i=0}^{2^M-1}S^{\mathrm{re},(i)}\cdot P\left(\mathbf{Y}^{\mathrm{re}}_\ell(k)\mid\mathbf{S}^{\mathrm{re},(i)}\right)\cdot P\left(\mathbf{S}^{\mathrm{re},(i)}\mid S^{+,\mathrm{re}}_\ell(k)\right)}{\sum\limits_{i=0}^{2^M-1}P\left(\mathbf{Y}^{\mathrm{re}}_\ell(k)\mid\mathbf{S}^{\mathrm{re},(i)}\right)\cdot P\left(\mathbf{S}^{\mathrm{re},(i)}\mid S^{+,\mathrm{re}}_\ell(k)\right)}. \tag{36}$$

The imaginary part of the speech estimate $\widehat{S}^{\mathrm{im}}_\ell(k)$ can be obtained in a similar way as $\widehat{S}^{\mathrm{re}}_\ell(k)$ and the final (complex-valued) speech estimate is calculated by $\widehat{S}_\ell(k) = \widehat{S}^{\mathrm{re}}_\ell(k) + j\widehat{S}^{\mathrm{im}}_\ell(k)$. The *a priori* speech estimate $S^+_\ell(k)$ is gained by (16) which is then used for the update step.

6 Conclusions

This paper provides new insights into links between speech enhancement and error concealment based on recursive MMSE estimation. It turns out that recent approaches to bit error concealment based on a predictor are well comparable to iterative approaches in speech enhancement, such as the Kalman filter approach. The main difference between both disciplines are the channel model and the noise pdf. In error concealment, powerful bit reliability information can be exploited in order to obtain robust estimation results, while in speech enhancement the channel is modeled on a sample level without reliable reference. On the other hand, the autoregressive model of speech and the prior computation are well comparable in both disciplines. Finally, some new research directions are identified for speech enhancement, inspired by error concealment.

Endnotes

[a] Please note that the presented signal model and the equations are valid in analogy also in the STFT domain.

[b] Please note that for binary phase-shift keying (BPSK) modulation, the LLR is obtained by $L(\hat{x}_m(n)) = 4 \cdot E_b/N_0 \cdot y_m(n)$ [33].

Appendix

For ease of readability, we will omit the indices ℓ and k in this section. In this appendix, we aim at showing that assuming a Gaussian distribution for the propagation error and the acoustic noise, the recursive MMSE estimator (14) turns out to be the Kalman filter as in (20), (21), (22), and (23). Employing (17) and (12), (14) turns out to be

$$\widehat{S} = \frac{\int\limits_{\mathbb{C}} S \cdot p_{\bar{E}}(S - \widehat{S}^{+}) \cdot p_D(Y - S)\, dS}{\int\limits_{\mathbb{C}} p_{\bar{E}}(S - \widehat{S}^{+}) \cdot p_D(Y - S)\, dS}. \tag{37}$$

Introducing a new integration variable $\bar{E} = S - \widehat{S}^{+}$, (37) can be rewritten as

$$\begin{aligned} \widehat{S} &= \widehat{S}^{+} + \frac{\int\limits_{\mathbb{C}} \bar{E} \cdot p_{\bar{E}}\left(\bar{E}\right) \cdot p_D\left(Y - \bar{E} - \widehat{S}^{+}\right)\, d\bar{E}}{\int\limits_{\mathbb{C}} p_{\bar{E}}\left(\bar{E}\right) \cdot p_D\left(Y - \bar{E} - \widehat{S}^{+}\right)\, d\bar{E}} \\ &= \widehat{S}^{+} + \widehat{E}. \end{aligned} \tag{38}$$

As can be seen, the recursive MMSE estimator turns out to be the sum of the *a priori* speech estimate $\widehat{S}^{+}$ and a fraction with each an integral in the numerator and denominator $\widehat{E} = f\left(Y - \widehat{S}^{+}\right)$. Please note that this fraction is a classical (non-recursive) MMSE estimator, however, with an extra term '$-\widehat{S}^{+}$' in the pdf $p_D(\cdot)$. Assuming a Gaussian distribution for $p_{\bar{E}}(\cdot)$ and $p_D(\cdot)$, this classical MMSE estimator turns out to be the Wiener filter as we will see later.

Employing (13) for $p_D(\cdot)$ and (18) for $p_{\bar{E}}(\cdot)$ as well as canceling the constant factors to the exponential functions, the fraction in (38) turns out to be

$$\widehat{E} = \frac{\int\limits_{\mathbb{C}} \bar{E} \cdot e^{-\frac{|\bar{E}|^2}{\sigma_{\bar{E}}^2}} \cdot e^{-\frac{|Y - \bar{E} - \widehat{S}^{+}|^2}{\sigma_D^2}}\, d\bar{E}}{\int\limits_{\mathbb{C}} e^{-\frac{|\bar{E}|^2}{\sigma_{\bar{E}}^2}} \cdot e^{-\frac{|Y - \bar{E} - \widehat{S}^{+}|^2}{\sigma_D^2}}\, d\bar{E}}. \tag{39}$$

Employing polar integration with $\bar{E} = |\bar{E}|e^{j\epsilon}$ and $d\bar{E} = |\bar{E}|\, d|\bar{E}|\, d\epsilon$, as well as employing $R = Y - \widehat{S}^{+} = |R|e^{j\rho}$, we obtain

$$\widehat{E} = \frac{\int\limits_0^\infty \int\limits_0^{2\pi} |\bar{E}|^2 e^{j\epsilon} \cdot e^{-\frac{|\bar{E}|^2}{\sigma_{\bar{E}}^2}} \cdot e^{-\frac{|\bar{E}|^2 + |R|^2 - 2|\bar{E}||R|\cos(\epsilon - \rho)}{\sigma_D^2}}\, d\epsilon\, d|\bar{E}|}{\int\limits_0^\infty \int\limits_0^{2\pi} |\bar{E}| \cdot e^{-\frac{|\bar{E}|^2}{\sigma_{\bar{E}}^2}} \cdot e^{-\frac{|\bar{E}|^2 + |R|^2 - 2|\bar{E}||R|\cos(\epsilon - \rho)}{\sigma_D^2}}\, d\epsilon\, d|\bar{E}|}. \tag{40}$$

Integrating with respect to ϵ using [42], (40) turns out to be

$$\widehat{E} = \frac{e^{j\rho} \int\limits_0^\infty |\bar{E}|^2 \cdot e^{-|\bar{E}|^2\left[\frac{1}{\sigma_{\bar{E}}^2} + \frac{1}{\sigma_D^2}\right]} \cdot I_1\left(\frac{2|\bar{E}||R|}{\sigma_D^2}\right) d|\bar{E}|}{\int\limits_0^\infty |\bar{E}| \cdot e^{-|\bar{E}|^2\left[\frac{1}{\sigma_{\bar{E}}^2} + \frac{1}{\sigma_D^2}\right]} \cdot I_0\left(\frac{2|\bar{E}||R|}{\sigma_D^2}\right) d|\bar{E}|} \tag{41}$$

with $I_0(\cdot)$ and $I_1(\cdot)$ being the modified Bessel function of zeroth and first order, respectively. Integrating with respect to $|\bar{E}|$ using [42], (41) turns out to be

$$\widehat{E} = \frac{\sigma_{\bar{E}}^2}{\sigma_{\bar{E}}^2 + \sigma_D^2} \cdot |R|e^{j\rho} = \frac{\sigma_{\bar{E}}^2 \sigma_D^2}{\sigma_{\bar{E}}^2 \sigma_D^2 + 1} \cdot \left(Y - \widehat{S}^{+}\right). \tag{42}$$

Thus, substituting (42) for $\widehat{E}$ in (38) and defining $\zeta = \sigma_{\bar{E}}^2/\sigma_D^2$ results in (cf. (20)-(23))

$$\widehat{S} = \widehat{S}^{+} + \frac{\zeta}{1 + \zeta} \cdot \left(Y - \widehat{S}^{+}\right). \tag{43}$$

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank Gerald Enzner for inspiring initial discussions toward this paper.

Received: 17 July 2014 Accepted: 7 January 2015
Published online: 20 February 2015

References

1. Trees Van HL, *Detection, Estimation and Modulation Theory. Vol 1. Detection, Estimation and Linear Modulation Theory.* (Wiley, Hoboken, NJ, USA, 1968)
2. S Kay, *Fundamentals of Statistical Signal Processing. Vol 1. Estimation Theory.* (Prentice Hall, Upper Saddle River, NJ, USA, 1993)
3. Y Ephraim, D Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoustics Speech Signal Process. **32**(6), 1109–1121 (1984)

4. L Rabiner, R Schafer, *Digital Processing of Speech Signals*. (Prentice Hall, Upper Saddle River, NJ, USA, 1978)
5. P Vary, R Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. (Wiley, Hoboken, NJ, USA, 2006)
6. R Kalman, A new approach to linear filtering and prediction problems. Trans. ASME J. Basic Eng. **82**, 35–45 (1960)
7. S Haykin, *Adaptive Filter Theory*. (Prentice Hall, Upper Saddle River, NJ, USA, 2002)
8. R McAulay, M Malpass, Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoustics Speech Signal Process. **28**(2), 137–145 (1980)
9. Y Ephraim, D Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoustics Speech Signal Process. **33**(2), 443–445 (1985)
10. R Martin, in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors, vol. 1 (Orlando, FL, USA, 2002), pp. 253–256
11. I Andrianakis, PR White, in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. MMSE speech spectral amplitude estimators with chi and gamma speech priors, vol. 3 (Toulouse, France, 2006), pp. 1068–1071
12. JS Erkelens, RC Hendriks, R Heusdens, J Jensen, Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. IEEE Trans Audio Speech Lang. Process. **15**(6), 1741–1752 (2007)
13. P Scalart, JV Filho, in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Speech enhancement based on a priori signal to noise estimation, vol. 2 (Atlanta, GA, USA, 1996), pp. 629–632
14. KK Paliwal, A Basu, in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. A speech enhancement method based on Kalman filtering (Dallas, TX, USA, 1987), pp. 177–180
15. E Zavarehei, S Vaseghi, in *Proc. of ISCA INTERSPEECH 2005*. Speech enhancement in temporal DFT trajectories using Kalman filters (Lisbon, Portugal, 2005), pp. 2077–2080
16. T Esch, P Vary, in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Speech enhancement using a modified Kalman filter based on complex linear prediction and supergaussian priors (Las Vegas, NV, USA, 2008), pp. 4877–4880
17. T Esch, P Vary, in *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC)*. Modified Kalman filter exploiting interframe correlation of speech and noise magnitudes (Seattle, WA, USA, 2008), pp. 1–4
18. N Görtz, in *Proc. of IEEE International Symposium on Information Theory*. Joint source channel decoding using bit-reliability information and source statistics (Cambridge, MA, USA, 1998), p. 9
19. T Fingscheidt, P Vary, Softbit speech decoding: a new approach to error concealment. IEEE Trans. Speech Audio Process. **9**(3), 240–251 (2001)
20. F Lahouti, AK Khandani, Soft reconstruction of speech in the presence of noise and packet loss. IEEE Trans. Audio Speech Lang. Process. **15**(1), 44–56 (2007)
21. AM Pourmir, F Lahouti, in *Proc. of 16th International Conference on Software, Telecommunications, and Computer Networks (SoftCOM)*. Joint source channel speech decoding using long-term residual redundancy (Split, Croatia, 2008), pp. 329–333
22. AJ Jameel, H Adnan, Y Xiaohu, A Hussain, in *Proc. of Developments in eSystems Engineering (DeSE)*. Error concealment of EVRC speech decoder using residual redundancy (Abu Dhabi, United Arab Emirates, 2009), pp. 84–88
23. S Han, F Pflug, T Fingscheidt, in *Proc. of 21th European Signal Processing Conference (EUSIPCO)*. Improved AMR wideband error concealment for mobile communications (Marrakech, Morocco, 2013), pp. 1–5
24. M Adrat, J Spittka, S Heinen, P Vary, in *Proc. of IEEE Workshop on Speech Coding (SCW)*. Error concealment by near optimum MMSE-estimation of source codec parameters (Delavan, WI, USA, 2000), pp. 84–86
25. F Pflug, T Fingscheidt, in *Proc. of 134th International Audio Engineering Society (AES) Convention*. Delayless robust DPCM audio transmission for digital wireless microphones (Rome, Italy, 2013), pp. 1–8
26. F Pflug, T Fingscheidt, Robust ultra-low latency soft-decision decoding of linear PCM audio. IEEE Trans. Audio Speech Lang. Process. **21**(11), 2324–2336 (2013)
27. A Papoulis, U Pillai, *Probability, Random Variables and Stochastic Processes*, 4th edn. (McGraw-Hill, New York, NY, USA, 2002)
28. T Esch, Model-based speech enhancement exploiting temporal and spectral dependencies. PhD thesis, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany (2012). http://darwin.bth.rwth-aachen.de/opus3/volltexte/2012/4035/pdf/4035.pdf
29. R Martin, PU Heute, Eds Antweiler C, *Advances in Digital Speech Transmission*. (Wiley, Hoboken, NJ, USA, 2008)
30. R Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001)
31. I Cohen, Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003)
32. T Gerkmann, RC Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1383–1393 (2012)
33. J Hagenauer, Source-controlled channel decoding. IEEE Trans. Commun. **43**(9), 2449–2457 (1995)
34. T Fingscheidt, P Vary, in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Robust speech decoding: A universal approach to bit error concealment, vol. 3 (Munich, Germany, 1997), pp. 1667–1670
35. GDT Schuller, B Yu, D Huang, B Edler, Perceptual audio coding using adaptive pre- and post-filters and lossless compression. IEEE Trans. Speech Audio Process. **10**(6), 379–390 (2002)
36. I Cohen, B Berdugo, Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process. Lett. **9**(1), 12–15 (2002)
37. Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T). Recommendation ITU-T P.56, Objective Measurement of Active Speech Level (1993). http://www.itu.int/rec/T-REC-P/
38. T Esch, P Vary, in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Model-based speech enhancement using SNR dependent MMSE estimation (Prague, Czech Republic, 2011), pp. 4073–4076
39. T Fingscheidt, S Suhadi, S Stan, Environment-optimized speech enhancement. IEEE Trans. Audio Speech Lang. Process. **16**(4), 825–834 (2008)
40. T Lotter, P Vary, Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. EURASIP J. Appl. Signal Process. **7**, 1110–1126 (2005)
41. B Fodor, T Fingscheidt, in *Proc. of European Signal Processing Conference (EUSIPCO)*. MMSE Speech Spectral Amplitude Estimation Assuming Non-Gaussian Noise (Barcelona, Spain, 2011), pp. 2314–2318
42. IS Gradshteyn, IM Ryzhik, *Table of Integral, Series, and Products*, 4th edn. (Academic Press, New York, NY, USA, 1965)