ADVANCED SIGNAL PROCESSING IN SPEECH COMMUNICATION

Peter Vary

Institute of Communication Systems and Data Processing (ind)
Aachen University (RWTH), Templergraben 55, D-52056 Aachen, Germany
E-mail: vary@ind.rwth-aachen.de

ABSTRACT

In digital mobile communications several measures are taken beyond speech coding to enhance the perceived quality in the presence of acoustic background noise and transmission errors. In premium mobile phones meanwhile advanced algorithms for noise suppression, error concealment, and finally artificial bandwidth extension come to practical application. It will be shown in this contribution that theses three different concepts of speech enhancement are actually based on the same common principle of conditional estimation, taking statistical a priori knowledge into account. Recent developments in these three areas are presented.

1. INTRODUCTION

In digital cellular radio systems using state of the art speech encoding, the speech quality suffers mainly from three different sources of degradation:

- acoustical background noise
- bandpass limitation of the speech signal to the telephone frequency band: $0.3 \ldots 3.4$ kHz
- residual bit errors after channel decoding.

These degradations can be combated by three independent countermeasures which have been evolved independently and will be subsumed here under the generic term *speech enhancement*.

Fig. 1 shows a block diagram of the typical speech communication system. A microphone captures the speech disturbed by acoustical background noise. The samples $y = s + n$ are obtained by using a telephone bandpass ($0.3 \ldots 3.4$ kHz) and an A/D-converter running at a sampling frequency of $f_s = 8$ kHz.

Noise Reduction (NR) is the first stage of enhancement which delivers a signal $\hat{s}$ with a reduced background noise level to the speech encoder. Noise reduction requires

- the knowledge of the noisy signal $y(k)$ and
- statistical a priori knowledge about speech and noise.

We assume that a state of the art speech encoder such as the GSM Enhanced Full Rate Codec (GSM-EFR) is used and that thus the level of the coding distortions is acceptable low. The samples $\hat{s}$ are transformed frame by frame into parameters v by the model based speech encoder. The parameters v are represented by vectors $\mathbf{x}$ of bits x. The transmission over the noisy channel is described by the so-called equivalent channel which includes modulation and demodulation as well as inner channel encoding and channel decoding. In adverse transmission conditions, residual bit errors may remain after channel decoding. Therefore, error concealment is required to reduce the resulting subjectively annoying effects.

Error Concealment (EC), the second stage of enhancement, is based on

- the decoded and possibly disturbed bits $\hat{x}$,
- bit-reliability information and
- a priori knowledge about parameters v.

The channel decoder delivers for groups $\hat{\mathbf{x}}$ of bits or even for individual bits $\hat{x}$ a reliability measure, the Decoder Reliability Indicator (DRI). The error concealment stage delivers estimated parameters $\hat{v}$ which are applied to the model based speech decoder.

Finally the decoded signal $\tilde{s}$ is applied to the third stage of speech enhancement, which performs the artificial extension of narrowband telephone quality ($0.3 \ldots 3.4$ kHz) to wideband telephone quality ($0.05 \ldots 7.0$ kHz). This step is of special interest as soon as network operators introduce true wideband speech coding [1] into the networks. For a long transition period narrowband and wideband speech terminals will coexist. In case of a sending narrowband terminal, the speech quality at the receiving end can be improved by artificial bandwidth extension.

Bandwidth Extension (BWE), the third stage of enhancement needs

- speech signal $\tilde{s}$ degraded (at least by) by telephone bandpass filtering
- a priori knowledge about the spectral envelope of wideband speech.

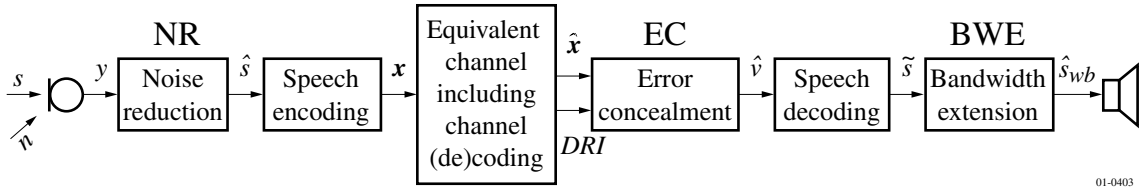

Figure 1: Speech communication system.

We will see, that the speech enhancement blocks NR, EC, and BWE of Fig. 1 are based on the same common principle of conditional estimation of signal parameters using statistical a priori knowledge.

The paper is organized as follows: In Section 2, the task of conditional estimation is introduced in a general form. In Section 3, the technique of single microphone noise reduction (NR) is addressed. Here, the conditional estimation is performed in the discrete Fourier domain independently for each frequency bin. In Section 4, the issue of error concealment (EC) by softbit source decoding is presented which is performed in the domain of the speech codec parameters, using a priori knowledge on parameter level. Finally, in Section 5, an algorithm for bandwidth extension (BWE) is described, which applies a state model of speech to conditionally estimate the wideband spectral envelope.

2. CONDITIONAL ESTIMATION

In this section, the procedure of conditional estimation is introduced in general terms. The speech enhancement algorithms, to be described later, are based on conditional estimation of speech parameters such as DFT-coefficients or predictor coefficients.
In Fig. 2 two different setups are illustrated.

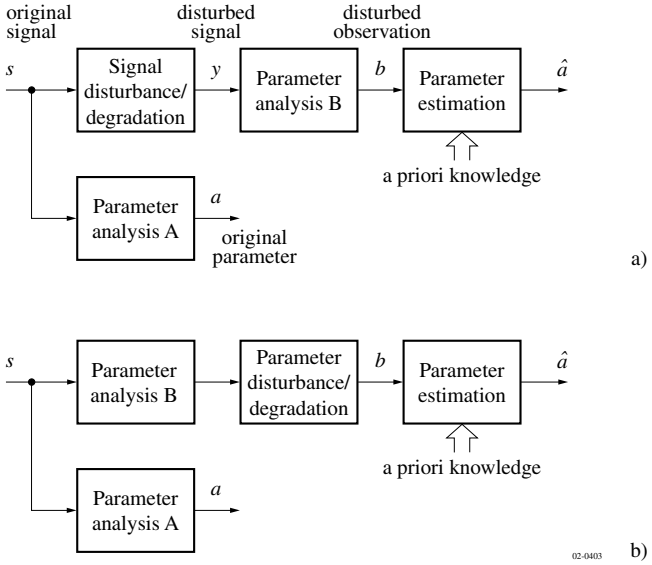

Figure 2: Conditional estimation in a parameter domain using a priori knowledge
a) signal disturbance
b) parameter disturbance

In both cases of this theoretical model, parameters a of the original signal s are obtained by a first analysis procedure A. In practice, the parameters a are not accessible, but instead of this we have disturbed/degraded observations b, which are gained by a second analysis procedure B. The analysis algorithms A and B must not necessarily be the same. The difference between the two situations consists in the place, where the disturbance is introduced: either on the signal level (Fig. 2a) or on the parameter level (Fig. 2b).

The task of the conditional estimator is then to form an estimate $\hat{a}$ for each individual parameter a by using the disturbed observation b, a priori knowledge in terms of the statistics of a (discrete probabilities $P(a)$ or probability density functions (PDFs) $p(a)$) and even statistical knowledge about the disturbance/degradation in terms of transition probabilities $P(b|a)$ or conditional PDFs $p(b|a)$. If information about a has been lost due to the disturbance, the original value can not be reconstructed without errors. Thus, the estimation relies on finding the best possible estimate $\hat{a}$ in a statistical sense, i.e., such that the average estimation error should be minimized. For this purpose the "a posteriori" probability density function $p(a|b)$ of the original value a conditioned on the instantaneous observation b is exploited.
A cost function $C(a,\hat{a})$ is introduced [2], which assigns a value to each combination of undisturbed a and estimated $\hat{a}$ signal and thus weights the estimation error for each given $(a, \hat{a})$.

An estimation rule $\hat{a} = f(b)$, which minimizes the expectation of the cost function, is asked for. The average costs or expectation of $C(a,\hat{a})$ can be formulated by integration over the joint PDF of the undisturbed and disturbed value

$$\rho_0 = E\{C(a,\hat{a})\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C(a,\hat{a}) \cdot p(a,b)\; da\; db. \tag{1}$$

The estimation rule $\hat{a} = f(b)$ can be found by minimizing ρ_0. After applying Bayes' theorem, equation (1) can be converted as follows:

$$\rho_0 = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} C(a,\hat{a}) \cdot p(a|b)\; da \right) p(b)\; db. \tag{2}$$

As $p(b)$ is non-negative the minimum of ρ_0 can be found by minimizing the inner integral for every possible observation b [2].

$$\rho_1 = E\{C(a,\hat{a})|b\} = \int_{-\infty}^{\infty} C(a,\hat{a}) \cdot p(a|b)\; da. \tag{3}$$

2.1 Conditional Minimum Mean Square Error Estimation

Choosing a square cost function, i.e., $C(a,\hat{a}) = (a-\hat{a})^2$, minimization of the inner integral of (2) w.r.t. $\hat{a}$

$$\frac{d}{d\hat{a}} \left[\int_{-\infty}^{\infty} (a-\hat{a})^2 \cdot p(a|b)\; da \right] = -\int_{-\infty}^{\infty} 2(a-\hat{a}) \cdot p(a|b)\; da \stackrel{!}{=} 0 \tag{4}$$

leads with $\int_{-\infty}^{\infty} p(a|b)\; da = 1$ to the minimum mean square error (MMSE) or conditional mean estimator:

$$\boxed{\hat{a} = E\{a|b\} = \int_{-\infty}^{\infty} a \cdot p(a|b)\; da.} \tag{5}$$

The a posteriori probability density $p(a|b)$ is unknown, but by using Bayes theorem once more, (5) can be rewritten as

$$\boxed{\hat{a} = \frac{\int_{-\infty}^{\infty} a \cdot p(b|a) \cdot p(a)\; da}{p(b)} = \frac{\int_{-\infty}^{\infty} a \cdot p(b|a) \cdot p(a)\; da}{\int_{-\infty}^{\infty} p(b|a) \cdot p(a)\; da}.} \tag{6}$$

Both (5) and (6) can be derived as well for discrete probabilities, if a and b take discrete values (e.g. due to quantization). The integrals have to be replaced by summations and the PDFs by probabilities.

Even a mixed form is possible where the statistics of only one quantity is discrete. In this case we need the "mixed form" of the Bayes' theorem.

Equation (5) is the theoretical solution, whereas (6) leads to the real implementation. Under certain constraints, which are fulfilled in the noise reduction application, closed analytical solutions of (6) can be derived (see Section 3).

2.2 Conditional Maximum a Posteriori Estimation

Another useful function to weight the estimation error for (2) is the uniform cost

$$C = \begin{cases} 0 & ; \quad |a-\hat{a}| < \varepsilon \\ 1 & ; \quad \text{else} \end{cases} . \tag{7}$$

To minimize the integral of (3) with this cost function the maximum of $p(a|b)$ must be in the area where $C = 0$. Thus the estimate $\hat{a}$ is obtained as the maximum of the a posteriori probability density function.

$$\boxed{\hat{a} = \arg\max_a p(a|b),} \tag{8}$$

which can also be reformulated via Bayes rule towards

$$\boxed{\hat{a} = \arg\max_a \frac{p(b|a) \cdot p(a)}{p(b)}.} \tag{9}$$

If the a posteriori probability density is symmetric and unimodal the MMSE estimate equals the maximum a posteriori (MAP) estimate (see e.g. [3]).

3. NOISE REDUCTION (NR)

As a first application of conditional estimation, the concept of single microphone noise suppression by *spectral subtraction* or more general by *spectral weighting* techniques is described and recent developments exploiting improved a priori knowledge are presented.

A block diagram of a typical implementation is illustrated in Fig. 3. Due to the linearity of the DFT, the noisy spectral components

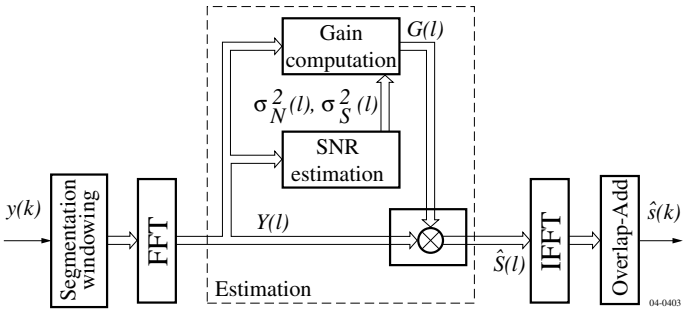

Figure 3: Single microphone noise reduction system.

can be described in the parameter domain by

$$Y(m,l) = R(m,l)e^{j\vartheta(m,l)} = S(m,l) + N(m,l), \tag{10}$$

where m is the frame index and l denotes the frequency index.

The complex components of speech and noise

$$\begin{aligned} S &= S_{\mathrm{Re}} + jS_{\mathrm{Im}} \\ N &= N_{\mathrm{Re}} + jN_{\mathrm{Im}}, \end{aligned}$$

with $S_{\mathrm{Re}} = \mathrm{Re}\{S\}$ and $S_{\mathrm{Im}} = \mathrm{Im}\{S\}$, etc. can also be described by their amplitudes (A,R) and their phases (α,β) according to,

$$S(m,l) = A(m,l)e^{j\alpha(m,l)} \text{ and } N(m,l) = B(m,l)e^{j\beta(m,l)}.$$

For simplicity, the frame index m is omitted in Fig 3 and the following. The sub-block for SNR estimation calculates the frequency dependent variances of the speech and noise DFT coefficients. Widely used methods for estimating the noise spectral variance σ_N^2 and the speech variances σ_S^2 are the *Minimum Statistics*-algorithm [4] proposed by Martin and and the decision directed approach proposed by Ephraim and Malah [5].

The (conditional) speech estimator is based on statistical models for speech and noise (a priori knowledge) and uses either the MMSE or the MAP criterion. Under certain assumptions about the PDFs of the speech and the noise components, the equations (6) and/or (9) can be solved analytically. Often, the estimate $\hat{S}(l)$ is obtained by simply applying real-valued spectral weights $G(l),\ 0 \le G \le 1$ to the noisy DFT coefficients $Y(l)$ according to

$$\hat{S}(l) \ = \ G(l) \cdot Y(l). \tag{11}$$

Fig. 4 shows a theoretical model of such a noise reduction algorithm in relation to the conditional estimation problem of Fig. 2a. The signal degradation consists in the additive background noise $n(k)$. In both analysis blocks A and B of Fig. 2a the Discrete Fourier Transform (DFT) is used.

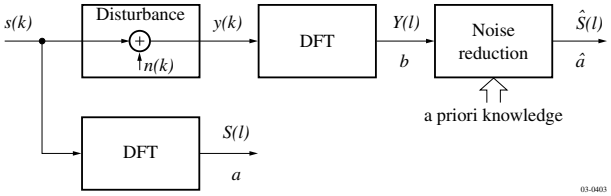

Figure 4: Model of noise suppression by conditional estimation. Correspondence with respect to Fig. 2a:
$S(l) = a$, $Y(l) = b$, $\hat{S}(l) = \hat{a}$.

3.1 Statistical Models for Noise Reduction

The formulation of appropriate transition PDFs $p(b|a) = p(Y|S)$ usually relies on the assumption, that real and imaginary part of the noise DFT coefficient $N(l)$ are zero mean independent Gaussian [5] with equal variance, which is justified by the central limit theorem. For many relevant acoustic noises this assumption approximates the real distribution very well. Thus, the transition PDFs $p(b|a) = p(Y_{\mathrm{Re}}|S_{\mathrm{Re}})$ can be written for each frequency index l separately for the real (and imaginary) part as

$$p(Y_{\mathrm{Re}}|S_{\mathrm{Re}}) = p(N_{\mathrm{Re}}) = \frac{1}{\sqrt{\pi}\sigma_N} \exp\left\{ -\frac{(Y_{\mathrm{Re}} - S_{\mathrm{Re}})^2}{\sigma_N^2} \right\}. \tag{12}$$

On the other hand the transition PDF $p(b|a) = p(Y|S)$ of the complex noisy DFT coefficient Y conditioned on the speech amplitude A and the phase α can then be written as joint Gaussian and the PDF of the noisy amplitude R given the speech amplitude A as Rician.

$$p(Y|A,\alpha) \ = \ \frac{1}{\pi\sigma_N^2} \exp\left\{ -\frac{|Y - Ae^{j\alpha}|^2}{\sigma_N^2} \right\} \tag{13}$$

$$p(R|A) \ = \ \frac{2R}{\sigma_N^2} \exp\left\{ -\frac{R^2 + A^2}{\sigma_N^2} \right\} I_0\left(\frac{2AR}{\sigma_N^2} \right). \tag{14}$$

I_0 denotes the modified Bessel function of zero-th order.
The statistical model of the real and imaginary parts of the DFT coefficients $S(l)$ of speech have been considered traditionally to be Gaussian distributed and consequently, the spectral amplitude A was assumed to be Rayleigh distributed

$$p(S_{\mathrm{Re}}) = \frac{1}{\sqrt{\pi}\sigma_S} \exp\left\{-\frac{S_{\mathrm{Re}}^2}{\sigma_S^2}\right\}, \quad p(A) = \frac{2A}{\sigma_S^2} \exp\left\{-\frac{A^2}{\sigma_S^2}\right\}. \quad (15)$$

Instead of a Gaussian model, Martin [6],[7] has proposed to use so-called super-Gaussian models, such as a Laplace or Gamma model for statistical independent real and imaginary parts of the speech coefficients.

An even more flexible super-Gaussian model which includes the Gaussian and the Gamma model as special cases has been proposed recently in [8],[9] as parametric approximation :

$$p(A) = \frac{\mu^{\nu+1}}{\Gamma(\nu+1)} \frac{A^{\nu}}{\sigma_S^{\nu+1}} \exp\left\{-\mu \frac{A}{\sigma_S}\right\}. \quad (16)$$

The parameters ν, μ determine the shape of the PDF and thus allow to adapt the underlying PDF of the conditional estimator optimally to the real distribution.

3.2 Conditional Estimation for Noise Reduction

Based on the given statistical models $p(a)$, conditional MMSE or MAP speech estimators (see $\hat{a}$ according to (5) and (9)) can be derived. Emphasis will be laid here on recent improvements concerning the exploitation of super-Gaussian a priori knowledge according to (16). The conditional estimators can generally be designed for either complex parameters, i.e., for S_{Re} and S_{Im}, or for the real valued spectral amplitudes, i.e., for A.
MMSE estimation (5) can be performed according to

$$\hat{S}_{\mathrm{Re}} = \int_{-\infty}^{\infty} S_{\mathrm{Re}} \cdot p(S_{\mathrm{Re}}|Y_{\mathrm{Re}})\, dS_{\mathrm{Re}} = \frac{\int_{-\infty}^{\infty} S_{\mathrm{Re}} \cdot p(Y_{\mathrm{Re}}|S_{\mathrm{Re}}) \cdot p(S_{\mathrm{Re}})\, dS_{\mathrm{Re}}}{\int_{-\infty}^{\infty} p(Y_{\mathrm{Re}}|S_{\mathrm{Re}}) \cdot p(S_{\mathrm{Re}})\, dS_{\mathrm{Re}}}. \quad (17)$$

Assuming a Gaussian distributions both of speech (15) and noise components, i.e., (12), equation (17) can be solved explicitly and leads to the so-called Wiener filter [10]

$$\hat{S}(l) \;=\; G(l) \cdot Y(l) \;=\; \frac{\sigma_S^2(l)}{\sigma_S^2(l) \;+\; \sigma_N^2(l)} \cdot Y(l). \quad (18)$$

Recently, improved MMSE estimators have been developed with Laplace or Gamma modeling of the real and imaginary parts of the speech DFT coefficients [6], [7].

From a perceptual point of view, it is more desirable to estimate the speech spectral amplitude than the complex spectrum due to the perceptual unimportance of the phase. The probably best known algorithm of Ephraim-Malah [5] is an MMSE estimator for the speech spectral amplitude A, i.e.,

$$\hat{A} = E\{A|Y\} = \int_0^{\infty} A \cdot p(A|Y)\, dA = \frac{\int_0^{\infty} A \cdot p(Y|A) \cdot p(A)\, dA}{\int_0^{\infty} p(Y|A) \cdot p(A) dA}. \quad (19)$$

Using (15), (13), the integration results in a spectral amplitude estimation rule according to (11). Later [11], the same authors introduced a minimum mean square error log spectral amplitude (MMSE-LSA) estimator, that minimizes the estimation error w.r.t. the logarithmic spectrum $\hat{A} = \exp\{E\{\log A|Y\}\}$.
Wolfe and Godsill [12] introduced alternatives to the Ephraim-Malah spectral amplitude estimator based on the maximum a posteriori estimation rule MAP (9):

$$\hat{A} = \arg\max_A p(A|R) = \arg\max_A \frac{p(R|A) \cdot p(A)}{p(R)}. \quad (20)$$

The MAP spectral amplitude estimator exploits the a posteriori density $p(a|b) = p(A|R)$, conditioned on the observed noisy amplitude. Another alternative was introduced by Wolfe and Godsill [12] in form of a joint MAP amplitude and phase estimator which results in a very similar weighting function.
In [8] and [9], the super-Gaussian model (16) has been applied in combination with the MAP or joint MAP approach of Wolfe and Godsill. Here the resulting efficient weighting rule allows an adaptation of the underlying super-Gaussian statistical model to the real distribution of the speech spectral amplitude of a given system. Under the assumption of a real-valued weight $G(l)$ (i.e. that the noisy phase of $Y(l)$ is the phase of the estimate $\hat{S}(l)$) the maximum of

$$p(A, \alpha|Y), \text{ resp. } \log\left(p(A, \alpha|Y)\right)$$

can be found by partial derivation with respect to A and α ([9]), leading with (16) to

$$G(l) \;=\; u + \sqrt{u^2 \;+\; \frac{\nu}{2\gamma(l)}}, \text{ with }, \; u = \frac{1}{2} - \frac{\mu}{4\sqrt{\gamma(l) \cdot \xi(l)}}, \quad (21)$$

where ν and μ are constants and ξ and γ are the a priori and the a posteriori SNRs

$$\xi(l) = \frac{\sigma_S^2(l)}{\sigma_N^2(l)}; \qquad \gamma(l) = \frac{R^2(l)}{\sigma_N^2(l)}.$$

In informal subjective listening tests the super-Gaussian models are clearly preferred by the test persons.

4. ERROR CONCEALMENT (EC)

Digital speech, audio, and video communication over noisy channels is usually based on source and channel coding. The source encoder delivers source parameters such as, e.g., A-law coded speech samples, or filter coefficients of the digital vocal tract model. The achievable speech, audio, or video quality is determined by the model, the quantizers and the resulting net bit rate of the source coding algorithm. For error protection channel coding is applied to the corresponding bit patterns of these parameters, to preserve the quality level over a wide range of channel characteristics. Nevertheless, even with channel coding residual bit errors occur in case of (temporarily) adverse channel conditions that may lead to a severe degradation of the signal quality. These annoying effects can be reduced or even be eliminated by *error concealment* (e.g., [13], [14]).

In this section we will discuss a concept of conditional parameter estimation that can be applied at the receiving end without any modifications of the transmitter. It is assumed that a parametric source encoder delivers quantized parameters v. Each parameter value is transmitted over the noisy channel as a bit pattern (bit vector) $\mathbf{x}$. At the receiving end a SISO channel decoder (Soft Input - Soft Output) is assumed, which produces soft information. This information

consists of bipolar bits $\hat{\mathbf{x}}$ and a reliability measure (instantaneous error probability) per received bit. This joint information can equivalently be described by so-called L-values or by real valued softbits $\tilde{x}$, with $-1 \le \tilde{x} \le +1$. A detailed discussion of these representations is beyond the scope of this paper.

The essential point of error concealment by exploiting this soft information is, that within the source decoding process reliability information from the channel decoder and a priori knowledge about the source is taken into consideration.

In the softbit approach we replace the table lookup module by a conditional parameter estimator.

The actual overall transmission system is depicted in Fig. 5.

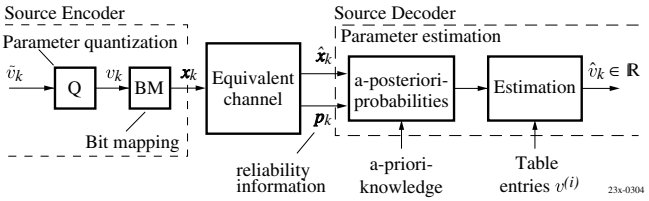

Figure 5: Error concealment by softbit source decoding .

The codec parameter $\tilde{v}_k$ at time instant k is quantized according to $\mathbf{Q}[\tilde{v}_k] = v_k$ with $v_k \in \{v^{(i)},\ i = 0, 1, \ldots 2^M - 1\}$ = QT (QT: quantization table) and can be represented by the quantization table index i. At the time instant k a bit combination

$$\mathbf{x}_k = (x_k(0), x_k(1), \ldots x_k(M-1)) \tag{22}$$

consisting of M bits is assigned via bit mapping (BM) to each quantized parameter v_k (or quantization table index i). There is a unique mapping between the quantizer levels v_k and the bit patterns $\mathbf{x}_k \in \{\mathbf{x}^{(i)},\ i = 0, 1, \ldots 2^M - 1\}$ The bits are assumed to be bipolar, i.e., $x_k \in \{-1, +1\}$. Due to the channel noise the received bit combination $\hat{\mathbf{x}}_k$ is possibly not identical to the transmitted one. In the conventional hardbit decoding scheme the received bit combination $\hat{\mathbf{x}}_k$ is applied to table look up decoding (inverse bit mapping scheme (BM^{-1})). Thereafter, the decoded parameter $\hat{v}_k$ is used within the specific parametric source decoder algorithm to reconstruct samples $\hat{s}$ of the speech signal (see also Fig. 1).

The concept of error concealment by softbit source decoding (SD) as depicted in Fig. 5, requires reliability information in terms of estimated instantaneous bit error probabilities

$$\mathbf{P}_k = (P_k(0), P_k(1), \ldots P_k(M-1)) \tag{23}$$

of the hardbit combination $\hat{\mathbf{x}}_k$.

The kernel of the SD-algorithm consists of

- step 1: calculation of 2^M a posteriori probabilities
  $P(v^{(i)} \mid \hat{\mathbf{x}}_k) = P(\mathbf{x}^{(i)} \mid \hat{\mathbf{x}}_k)$ with $i \in \{0, 1, \ldots 2^M - 1\}$
- step 2: estimation of a real-valued parameter $\hat{v}_k$.

Fig. 6 shows the theoretical model of this approach. With regard to Fig. 2b the analysis block A delivers a quantized parameter v, e.g. a predictor coefficient of a speech codec (see also Fig. 5). In contrast to that, the analysis block B produces the quantized version v of this parameter in terms of the bit pattern $\mathbf{x}$. This bit pattern is transmitted over the equivalent noisy channel, which introduces disturbance (in addition to the quantizer). At the receiving end, we have a possibly degraded bit pattern plus some reliability information, represented by the softbits $\tilde{x}$. The task of the estimator is to

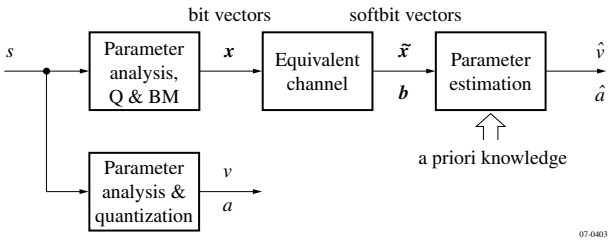

Figure 6: Theoretical model of softbit error concealment by conditional estimation. Correspondence with respect to Fig. 2b: $v = a$ (quantized parameter), $\tilde{\mathbf{x}} = \mathbf{b}$ (softbit vector), $\hat{v} = \hat{a}$.

determine an estimate $\hat{v}$ according to the MMSE or MAP criterion, taking the reliability information from the channel decoder and the a priori knowledge about the source into account.

4.1 Statistical Models for Error Concealment

In specifying the required a priori knowledge there are some degrees of freedom. We need a priori knowledge about the quantized parameter in terms of the 2^M probabilities $P(\mathbf{x}^{(i)}) = P(v^{(i)}),\ i = 0, 1, \ldots 2^M - 1$, i.e., the histogram of the quantized parameter v.

In the general case we can model the quantized parameter as a Markov process. To find out an appropriate Markov order it is convenient to measure terms such as $P(\mathbf{x}_k)$, $P(\mathbf{x}_k \mid \mathbf{x}_{k-1})$, or $P(\mathbf{x}_k, \mathbf{x}_{k-1})$ or even higher order conditional and joint probabilities. This can be achieved by applying a large signal database to the source encoder and by counting how often the different quantizer output symbols, or different pairs of output symbols, occur. We call $P(\mathbf{x}_k)$ 0th order a priori knowledge (AK0) because it gives a statistical description of a 0th order Markov process, i.e., a memoryless process. Accordingly, we call $P(\mathbf{x}_k \mid \mathbf{x}_{k-1})$ or $P(\mathbf{x}_k, \mathbf{x}_{k-1})$ 1st order a priori knowledge (AK1) because it refers to a 1st order Markov process. The decision which model should be taken is a matter of the

- observed redundancy
- allowed complexity of the softbit source decoder
- tradeoff between performance and complexity.

For simplicity we restrict here to the case of a 0th order Markov process. Then the statistical model of the parameter consists of the measured histogram of the quantized parameter, i.e. the probabilities $P(\mathbf{x}^{(i)}) = P(v^{(i)}),\ i = 0, 1, \ldots 2^M - 1$. With the entropy defined as

$$H(\mathbf{x}_k) \;=\; -\sum_{i=0}^{2^M-1} P\big(\mathbf{x}^{(i)}\big) \log_2 P\big(\mathbf{x}^{(i)}\big)\,. \tag{24}$$

the redundancy of $\Delta R = M - H(\mathbf{x}_k)$ can be exploited for error concealment.

4.2 Conditional Estimation for Error Concealment

For parameter estimation we can use once more either the MMSE or the MAP criterion. The right decision depends on the specific parameter. In speech coding, the MAP criterion is appropriate e.g. for the pitch information, while for filter parameters and gain factors the MMSE criterion gives subjectively better results.

Let us assume that the channel related transition probabilities $P(\hat{\mathbf{x}}_k \mid \mathbf{x}^{(i)})$ on bit vector level can be computed from the P_k on bit level. This is true, if we can derive the (estimated) instantaneous bit error rate from the soft output $\tilde{x}$ (Fig.6), respectively from the

decoder reliability DRI (Fig. 5) of the equivalent channel. Using Bayes' theorem the a posteriori probability $P(a|b) = P(\mathbf{x}^{(i)} \mid \hat{\mathbf{x}}_k)$ can be calculated as

$$P(\mathbf{x}^{(i)} \mid \hat{\mathbf{x}}_k) = \frac{P(\hat{\mathbf{x}}_k \mid \mathbf{x}^{(i)}) P(\mathbf{x}^{(i)})}{\sum_{j=0}^{2^M-1} P(\hat{\mathbf{x}}_k \mid \mathbf{x}^{(j)}) P(\mathbf{x}^{(j)})} \tag{25}$$

If we have received a certain bit pattern $\hat{\mathbf{x}}_k$, then the probability $P(\mathbf{x}^{(i)} \mid \hat{\mathbf{x}}_k)$ quantifies the reliability of the decision that the pattern $\mathbf{x}^{(i)}$ and thus the quantized parameter value $v^{(i)}$ was transmitted at time k.

The MAP estimator follows the criterion

$$\hat{v}_k = v^{(j)} \quad \text{with} \quad j = \arg\max_i P\big(\mathbf{x}^{(i)} \mid \hat{\mathbf{x}}_k\big). \tag{26}$$

MAP estimation minimizes the probability of an erroneous decoded parameter. The decoded parameter $\hat{v}_k$ equals one of the codebook/quantization table entries. In case of error-free transmission only one of the 2^M a priori probabilities takes the value 1, all the others are 0. In this situation the MAP-decoder selects the same table entry as the conventional table-look up decoder.

In applying the MMSE solution according to (5) we have to take into consideration, that the statistics of the parameters are described here by discrete probabilities. Therefore, the integrals have to be replaced by discrete summations. The optimum decoded parameter in a minimum mean square error sense equals

$$\hat{v}_k = \sum_{i=0}^{2^M-1} v^{(i)} \cdot P\big(\mathbf{x}_k^{(i)} \mid \hat{\mathbf{x}}_k\big). \tag{27}$$

According to the orthogonality principle of linear mean square estimation (see, e.g., [2]) the variance of the estimation error $e_0 = \hat{v}_k - v_k$ is $\sigma_e^2 = \sigma_v^2 - \sigma_{\hat{v}}^2$ with σ_v^2 being the variance of the undisturbed parameter v_k and $\sigma_{\hat{v}}^2$ denoting the variance of the estimated parameter $\hat{v}_k$. Because of $\sigma_e^2 \geq 0$ we can state that the variance of the estimated parameter is smaller than or equals the variance of the error free parameter.

For the worst case channel with $P_k = 0.5$ the a posteriori probabilities simplify to $P(\mathbf{x}^{(i)} \mid \hat{\mathbf{x}}_k) = P(\mathbf{x}^{(i)})$. If in this case the unquantized parameter $\tilde{v}_k$ as well as the quantization table entries $v_k^{(i)}$ are distributed symmetrically around zero the MMSE estimated parameter according to Eq. (27) is attenuated to zero (by weighted averaging). These symmetries are often found for gain factors (plus sign) in speech and audio encoders. Thus the MMSE estimation of gain factors results in an inherent muting mechanism providing a graceful degradation of the signal quality. This is one of the main advantages of softbit source decoding.

On the other hand, if the channel is free of errors ($p_e = 0$) and $\mathbf{x}^{(\kappa)}$ has been transmitted, then all the parameter transition probabilities are zero except $P(\hat{\mathbf{x}}_k \mid \mathbf{x}^{(\kappa)}) = 1$. This yields $P(\mathbf{x}^{(\kappa)} \mid \hat{\mathbf{x}}_k) = 1$ while all other a posteriori probabilities become zero. As a consequence, also the MMSE estimator yields the correct parameter value $\hat{v}_k = v_k$. This is equivalent to bit exactness in clear channel situations. In practical applications e.g. in the GSM transmission link, the subjective speech or audio quality can significantly be improved in the presence of residual errors at the output of the channel decoder.

5. BANDWIDTH EXTENSION (BWE)

In today's public telephone networks, the limitation to a frequency range of about 0.3 to 3.4 kHz causes the typical sound of the *narrowband* telephone speech. As long as there are still (sending) narrowband terminals in the network, artificial bandwidth extension is a very attractive feature for any receiving wideband terminal.

The basic concept of artificial bandwidth extension is to exploit implicit redundancy of the linear source-filter model, which is widely used in speech coding and recognition. This model consists of an *auto-regressive* (AR) filter (corresponding to the vocal tract) and a source producing a spectrally flat excitation. According to this model bandwidth extension is divided into two separate tasks [15]:

- the *extension of the spectral envelope* of the speech signal and
- the *extension of the excitation signal*.

A common feature of most of the algorithms proposed in literature is, that in a first step, the baseband of the excitation (0.3...3.4 kHz) is obtained from the narrowband speech signal by linear prediction (LP). The excitation signal is spectrally flat and can be extended to the frequency band 0.05...7.0 kHz by simple spectral folding or (pitch synchronous) modulation techniques (e.g. [15], [16], [17]).

In a second step, the spectral envelope of the wideband speech signal is estimated in terms of LP coefficients or in terms of the corresponding cepstral coefficients.

Finally in a third step, the artificial wideband speech signal is produced by applying the extended excitation signal to the extended AR-filter.

A simplified block diagram of such an approach is given in Fig. 7 [18],[17], where the wideband spectral envelope is estimated in terms of cepstral coefficients $\hat{\mathbf{c}}_{wb}$.

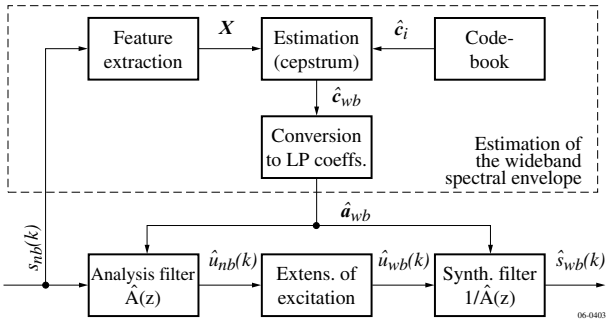

Figure 7: Block diagram of the BWE algorithm

The estimated cepstral coefficients $\hat{\mathbf{c}}_{wb}$ are converted to the wideband LP-coefficients $\hat{\mathbf{a}}_{\mathrm{wb}}$ which describe the all-pole (vocal tract) filter $1/\hat{A}(z)$ of the source-filter model. The estimation is based on the observation of a feature vector $\mathbf{X}$ that is extracted from the narrowband speech signal $s_{\mathrm{nb}}(k)$, which has been interpolated before to the sample rate of $f_s = 16$ kHz.

By applying the corresponding (inverse) FIR analysis filter $\hat{A}(z)$ to the narrowband input signal $s_{\mathrm{nb}}(k)$, an estimate $\hat{u}_{\mathrm{nb}}(k)$ of the narrowband excitation signal (prediction residual) is derived, since the analysis filter is the inverse of the vocal tract (synthesis) filter. The *extension of the excitation signal* converts the narrowband excitation signal $\hat{u}_{\mathrm{nb}}(k)$ into an extended version $\hat{u}_{\mathrm{wb}}(k)$ by exploiting the spectral flatness. The extended wideband excitation signal $\hat{u}_{\mathrm{wb}}(k)$ is fed into the wideband all-pole synthesis filter $1/\hat{A}(z)$ to synthesize the enhanced output speech $\hat{s}_{\mathrm{wb}}(k)$.

In the bandwidth extension algorithm described here [18],[17], the method of conditional estimation is applied in a more sophisticated version than for acoustical background noise reduction, as the a priori knowledge is now based on a state model of speech production.

Here, the kernel task of extending the spectral envelope will be considered only. Fig. 8 describes this task in the context of conditional estimation according to Fig. 2a.

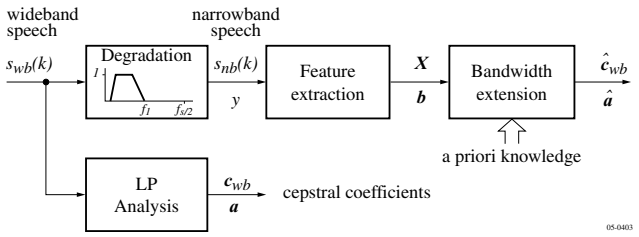

Figure 8: Model of bandwidth extension by conditional estimation
Correspondence with respect to Fig. 2a: $\mathbf{c}_{wb} = \mathbf{a}$, $\mathbf{X} = \mathbf{b}$, $\hat{\mathbf{c}}_{wb} = \hat{\mathbf{a}}$ (vectors); wb=wideband; nb=narrowband

With regard to Fig. 2a, the two analysis procedures A and B are different. By the analysis A, we calculate via linear prediction (LP) analysis, the vector of cepstral coefficients $\boldsymbol{c}_{\mathrm{wb}}$ of the wideband speech signal $s_{\mathrm{wb}}(k)$, whereas analysis B delivers a feature vector $\mathbf{X}$ which is extracted from the narrowband signal $s_{\mathrm{nb}}(k)$. The bandwidth extension algorithm estimates the cepstral coefficients using the feature vector and an underlying state model of speech production. Each speech frame of 20 ms with time or frame index m, can be characterized by a state $S_i,\ i = 1,\ldots N_s$, the typical vector of cepstral coefficients $\hat{\mathbf{c}}_i$ and the "measured" feature vector $\mathbf{X}$. For simplicity the frame index will be omitted in the sequel.

5.1 Statistical Model for Bandwidth Extension

5.1.1 State Model

Each state $S_i,\ i = 1,2,\ldots N_S$ of the model is assigned to a typical speech sound (frame of 20 ms) which is associated with a representative envelope $\hat{\mathbf{c}}_i$.

The states of the model are defined by the entries $\hat{\mathbf{c}}_i$ of a *vector quantizer* (VQ) of the spectral envelope representation $\mathbf{c}_{wb}$ (vector of cepstral coefficients of the wideband speech signal): each centroid $\hat{\mathbf{c}}_i$ of the vector quantizer represents the spectral envelope of a typical speech sound. However, wideband speech s_{wb} is available only in the training phase, whereas in the application phase of the BWE algorithm the states S_i have to be identified by classification of the narrowband speech signal s_{nb}.

For each signal frame a vector $\mathbf{X}$ of features which should deliver maximum information about the state S_i, is extracted from the narrowband signal. The vector $\mathbf{X}$ contains features like normalized autocorrelation function, zero crossing rate, normed frame energy, gradient index, local kurtosis and spectral centroid, for a detailed description refer to [17].

The connection between the observations $\mathbf{X}$ and the states S_i (and thus the corresponding codebook entries $\hat{\mathbf{c}}_i$) is contributed by a state-specific statistical model. For each state S_i the features $\mathbf{X}$ as well as the unknown spectral envelope $\mathbf{c}_{wb}$ exhibit characteristic statistical relations. The following statistical quantities can be measured during an offline training process with representative wideband speech signals $s_{wb}(k)$ and corresponding narrowband signals $s_{nb}(k)$:

- the codebook entries $\hat{\mathbf{c}}_i$ of the vector quantizer (e.g. by using the standard LBG training algorithm [19])
- the state probabilities $P(S_i)$
- the conditional feature PDFs $p(\mathbf{X}|S_i)$ (observation probabilities).

Note: In [18],[17] a hidden Markov model(HMM) is used. However, for explaining the basic concept, a simpler state model is considered here, which does not take into account the state transition probabilities.

The wideband speech is needed to calculate the true state sequence and the narrowband speech is used to determine the conditional observation PDFs of feature vectors $\mathbf{X}$.

As the *observation PDF* is conditioned to the state S_i there exists a separate PDF $p(\mathbf{X}|S_i)$ for each state. According to the definition of the state model, it is assumed that the observation $\mathbf{X}$ for each frame only depends on the particular frame.

A common way to model measured high-dimensional probability density functions is the approximation with *Gaussian mixture models* (GMM; see, e.g., [20], [21]).

5.2 Conditional Estimation for Bandwidth Extension

5.2.1 Minimum Mean-Square Error Estimation (MMSE)

By the MMSE estimation rule according to (5) a continuous estimation of the parameter vector $\mathbf{c}_{wb}$ shall be performed with the a posteriori PDF $p(\mathbf{a}|\mathbf{b}) = p(\mathbf{c}|\mathbf{X})$.

Thus the *minimum mean-square error* (MMSE) estimator for the cepstral coefficient vector is given by

$$\hat{\mathbf{c}}_{\mathrm{MMSE}} = E\{\mathbf{c}|\mathbf{X}\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{c} \cdot p(\mathbf{c}|\mathbf{X})\, d\mathbf{c}\ . \tag{28}$$

Because we do not have a model of the conditional PDF $p(\mathbf{c}|\mathbf{X})$ in closed-form, this quantity has to be expressed indirectly via the states of the model

$$p(\mathbf{c}|\mathbf{X}) = \sum_{i=1}^{N_S} p(\mathbf{c}, S_i|\mathbf{X})\ . \tag{29}$$

Insertion of $p(\mathbf{c}, S_i|\mathbf{X}) = p(\mathbf{c}|S_i, \mathbf{X}) \cdot P(S_i|\mathbf{X})$ into (28) yields

$$\hat{\mathbf{c}}_{\mathrm{MMSE}} = \sum_{i=1}^{N_S} P(S_i|\mathbf{X}) \cdot \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{c}\, p(\mathbf{c}|S_i, \mathbf{X})\, d\mathbf{c}\ , \tag{30}$$

which can be written as:

$$\hat{\mathbf{c}}_{\mathrm{MMSE}} = \sum_{i=1}^{N_S} \hat{\mathbf{c}}_i P(S_i|\mathbf{X})\ . \tag{31}$$

Hence, the estimated coefficient set $\hat{\mathbf{c}}_{\mathrm{MMSE}}$ is calculated by a weighted sum of the individual code book entries $\hat{\mathbf{c}}_i$, which are weighted by the respective a posteriori probabilities of the corresponding states. Accordingly, the described MMSE estimator can be interpreted in analogy to the error concealment algorithm described in Section 4 as a *soft classification*.

5.2.2 Calculation of A Posteriori State Probabilities $P(S_i|\mathbf{X})$

The a posteriori probability $P(S_i|\mathbf{X})$ can be formulated in terms of the measured state probabilities $P(S_i)$ and the measured conditional feature PDFs $p(\mathbf{X}|S_i)$ as follows:

$$P(S_i|\mathbf{X}) = \frac{p(S_i, \mathbf{X})}{p(\mathbf{x})} = \frac{p(\mathbf{X}|S_i) \cdot P(S_i)}{\sum_{j=1}^{N_S} p(\mathbf{X}|S_j) \cdot P(S_j)}\ . \tag{32}$$

In the denominator of (32) the hardly tractable PDF $p(\mathbf{X})$ of the observation sequence has been replaced by a summation over the marginal density of the joint PDF $p(S_j, \mathbf{X}) = p(\mathbf{X}|S_j) \cdot P(S_j)$.

6. CONCLUSION

If speech is transmitted in the presence of acoustical background noise over a disturbed digital telephone channel, the speech quality at the receiving end will be degraded. First of all, the speech quality is limited due to the telephone frequency characteristic (0.3 . . . 3.4 kHz) of A/D conversion. Secondly the performance of the speech codec will be reduced by the acoustical background noise. Finally, residual bit errors occur in practice, if the channel decoder is temporarily overloaded during adverse channel conditions.

These three sources of degradation can be combated by three different advanced approaches of speech enhancement, i.e.

- noise reduction (NR)
- error concealment (EC)
- bandwidth extension (BWE).

It has been shown in this contribution that the solutions found for these problems have the same mathematical roots in terms of conditional Bayesian estimation. From an algorithmic point of view, the main differences consist in the underlying statistical models based on probability density functions in the case of NR, on discrete probabilities in the EC-application and a mixture of probabilities and densities in the case of BWE.

For simplicity, the concepts have been explained without taking frame-to-frame correlation into account. However, this extension is straightforward and can be found in the cited literature.

For each of these three topics state of the art approaches and recent new solutions have been presented.

Acknowledgement: The author would like to thank Peter Jax and Thomas Lotter for many valuable discussions and contributions to this paper.

REFERENCES

[1] B. Bessette, R. Lefebvre, M. Jelínek, J. Rotola-Pukkila, H. Mikkola, and K. Järvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 8, pp. 620–636, Nov. 2002.

[2] J.L. Melsa and D.L. Cohn, *Decision and Estimation Theory*, McGraw-Hill, 1978.

[3] H.L. van Trees, *Detection, Estimation and Modulation Theory: Part 1, Detection Estimation and Linear Modulation Theory*, John Wiley and Sons, 1968.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, pp. 504–512, July 2001.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[6] R. Martin, "Speech Enhancement using MMSE short time spectral estimation with Gamma distributed priors," in *Proc. International Conf., on Acoustics, Speech and Signal Processing*, Orlando, USA, 2002, pp. 87–90.

[7] R. Martin and C. Breithaupt, "Speech Enhancement in the DFT Domain using Laplacian Speech Priors," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, September 2003, pp. 253–256.

[8] T. Lotter and P. Vary, "Noise Reduction by Maximum a Posteriori Spectral Amplitude Estimation with Supergaussian Speech Modelling," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, September 2003, pp. 83–86.

[9] T. Lotter and P. Vary, "Noise Reduction by Joint Maximum a Posteriori Spectral Amplitude and Phase Estimation with Super-Gaussian Speech Modelling," in *European Signal Processing Conference*, Vienna, Austria, September 2004.

[10] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*, Principles of Electrical Engineering Series, MIT Press, 1949.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, April 1985.

[12] P.J. Wolfe and S.J. Godsill, "Efficient Alternatives to the Ephraim-Malah Suppression Rule for Audio Signal Enhancement," *EURASIP Journal on Applied Signal Processing, Special Issue: Digital Audio for Multimedia Communications*, pp. 1043–1051, September 2003.

[13] T. Fingscheidt and P. Vary, "Softbit Speech Decoding: A new Approach to Error Concealment," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 3, pp. 240–251, 2001.

[14] C.G. Gerlach, *Beiträge zur Optimalität in der codierten Sprachübertragung*, Ph.D. thesis, Aachener Beiträge zu digitalen Nachrichtensystemen, Hrsg. P. Vary, Bd. 5, (ISBN 3-86073-434-2), 1996.

[15] H. Carl, *Untersuchung verschiedener Methoden der Sprachkodierung und eine Anwendung zur Bandbreitenvergrößerung von Schmalband-Sprachsignalen*, Ph.D. thesis, Ruhr-Universität Bochum, Bochum, Germany, 1994.

[16] U. Kornagel, "Spectral widening of the excitation signal for telephone-band speech enhancement," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Darmstadt, Germany, September 2001, pp. 215–218.

[17] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, Aug. 2003.

[18] P. Jax, *Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds*, Ph.D. thesis, Aachener Beiträge zu digitalen Nachrichtensystemen, Hrsg. P. Vary, (ISBN 3-86073834-8), 2002.

[19] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Communications*, vol. 1, pp. 84–95, 1980.

[20] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[21] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, Wiley, Teubner, 1996.