EUSIPCO 2013 1569730573

IMPROVED AMR WIDEBAND ERROR CONCEALMENT FOR MOBILE COMMUNICATIONS

Sai Han, Florian Pflug, and Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig
Schleinitzstr. 22, 38106 Braunschweig, Germany
{s.han,f.pflug,t.fingscheidt}@tu-bs.de

ABSTRACT

Mobile wideband speech communication (*HD Voice*) is more and more available in the past years, primarily in 3G networks. The specifics of mobile communication — even if it is packet-switched — is that received frames with residual bit errors after channel decoding must not necessarily be marked as *lost*, instead they may be marked as *bad* (bad frame indicator, BFI). In this work we present how to exploit the information of a soft input (i. e., a log-likelihood ratio input) within the Adaptive Multirate Wideband (AMR-WB) speech decoder, allowing a more robust error concealment as compared to the 3GPP Recommendation. Log-likelihood ratios may be taken from a soft-output channel decoder, or, as in our generic simulation, directly from the demodulator, without the need of a BFI. Since error concealment is non-mandatory, chipset manufacturers are free to implement this alternative speech decoding scheme still in a standard-compliant fashion.

Index Terms— AMR-WB, error concealment, soft-decision decoding, parameter estimation

1. INTRODUCTION

Roughly 150 years after the invention of the telephone, we are still widely used to narrowband speech communication with frequencies in the range of about 300 . . . 3400 Hz. With the standardization of the Adaptive Multirate Wideband (AMR-WB) speech coder in 3GPP [1] in 2001, the path was paved for a much better speech quality and intelligibility in mobile speech communication by transmission of the wideband speech frequency range of 50 . . . 7000 Hz — both in GSM, and in 3G WCDMA networks. Known as *HD Voice*, wideband speech services have now been introduced in roughly 50 mobile networks in about 40 countries [2, 3]. While users in these networks may indeed enjoy the improved speech quality, error robustness in adverse transmission conditions again comes into the focus: Higher quality expectations in consequence also call for improved coverage in these cases. However, compared to the respective narrowband speech service supported by the Adaptive Multirate (AMR) speech coder, error robustness is roughly the same, when comparing the widely employed 12.65 kbps mode of AMR-WB [1] with the widely used 12.2 kbps mode of AMR [4].

Several error concealment schemes have been proposed for plain speech waveform transmission [5], for ADPCM-coded speech [6, 7], and for hybrid speech coders such as GSM [8, 9], GSM Enhanced Fullrate [10], AMR [11], and AMR-WB [12]. The aforementioned schemes do only rely on a bad frame indicator (BFI), while other schemes already use some kind of soft information from the previous stage (either demodulator or channel decoder) [13–15]. In the 90's, Fingscheidt and Vary introduced the soft-decision speech decoding paradigm [16, 17] allowing to exploit so-called log-likelihood ratios (LLRs) obtained from the demodulator or from some soft-output channel decoding scheme [13, 18]. It made use of soft information in the whole chain of receiver blocks including speech decoding/error concealment. This concept was applied to G.726 ADPCM [19], A-law PCM and GSM Fullrate speech coding [17], high-quality PCM audio [20], up to distributed automatic speech recognition (DSR) [21, 22]. Recently much research focused on joint or iterative source-channel coding with soft-decision reconstruction, for example, MELP speech coding [23] or image and video coding in JPEG 2000 [24] and H.264/AVC [25], respectively.

Intraframe and interframe residual redundancy is exploited in an MMSE-based decoder utilizing soft decisions within GSM-AMR [26], however, only the line spectral frequency (LSF) parameters have been regarded (compare to [27]). In 2007, an AMR-WB speech decoder was proposed utilizing soft input and providing soft output [28, 29], the latter calculated as extrinsic LLRs employing the approach presented in [30]. The soft-output (i. e., extrinsic LLRs) is also required for the channel decoder in [28, 29]. These iterative source-channel decoding approaches, however, do not exploit 1st-order a priori knowledge for the reconstruction of the pitch delay parameter. Moreover, in [28, 29] results are presented only for the less-used modes 15.85 kbps and 23.05 kbps.

In this paper, we adopt the soft-decision speech decoding framework from [17] and apply it to the AMR-WB speech codec. The parameters consisting of immittance spectral pair (ISP) vectors, vector-quantized gains, and the adaptive codebook index (pitch delay) are reconstructed with appropriate parameter estimators in a soft-decision decoding framework. Moreover, we show how to use 1st-order *a priori* knowledge (AK1) for parameters with an alternating number of bits and for split-multistage vector-quantized parameters (ISPs).

The paper is organized as follows: In Section 2, we will present our soft-decision decoding approach applied to AMR-WB, Section 3 describes the reference AMR-WB error concealment. Simulation results for the prominent 12.65 kbps mode are discussed in Section 4. Finally, some conclusions are drawn in Section 5.

2. AMR-WB SOFT-DECISION ERROR CONCEALMENT

2.1. Overview

In this paper, the channel model for speech transmission is described as an equivalent channel, which comprises modulation and (soft) demodulation with or without channel coding. The block diagram of the entire simulation setup depicted in Fig. 1 generically represents a mobile speech communication system such as GSM, UMTS,

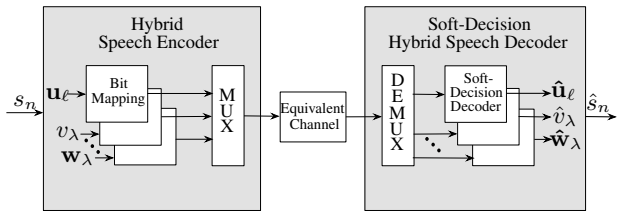


Fig. 1: Block diagram with the hybrid speech encoder, equivalent channel, and soft-decision speech decoder.

or LTE. The 16-bit linear encoded PCM speech samples s_n with sample index $n \in \{0, 1, \ldots\}$ are analyzed to extract a parameter set $\mathbf{u}_\ell = (u_{\ell,1}, u_{\ell,2}, \ldots, u_{\ell,N_p})$ with N_p being the number of parameters in this parameter set (e. g., ISP vector), or certain parameters v_λ, $\mathbf{w}_\lambda$ (e. g., pitch delay and vector-quantized gains) in the encoder, with $\ell \in \{1, 2, \ldots\}$ being the speech frame index and $\lambda \in \{1, 2, \ldots\}$ being a global subframe index. A parameter can be scalar- or vector-quantized according to a corresponding quantization codebook and can be expressed by the corresponding codebook index $i \in \{0, 1, \ldots, 2^M - 1\}$ using M bits. After quantization, the quantized parameter is represented by the bit combination[1] $\mathbf{x}_\ell = (x_\ell(0), x_\ell(1), \ldots, x_\ell(m), \ldots, x_\ell(M-1))$ with $x_\ell(m) \in \{-1, +1\}$. Thereafter, the bit combinations of all codec parameters are multiplexed and transmitted. The log-likelihood ratios (LLRs) of each received bit $\hat{x}_\ell(m) \in \{-1, +1\}$ are

$$L(\hat{x}_\ell(m)) = 4 \cdot \frac{E_b}{N_0} \cdot a \cdot y_\ell(m)\,, \tag{1}$$

with a being the fading factor of the channel, $y_\ell(m)$ being the received real-valued sample (i. e., $x_\ell(m)$ is distorted by additive white Gaussian noise), and E_b/N_0 being the ratio of the bit energy to the spectral noise density, $\hat{x}_\ell(m) = \text{sign}(L(\hat{x}_\ell(m))) = \text{sign}(y_\ell(m))$, respectively. As a result, a bit error probability can be obtained for each parameter bit m in frame ℓ according to

$$p_{e,\ell}(m) = \frac{1}{1 + \exp(|L(\hat{x}_\ell(m))|)}\,. \tag{2}$$

Furthermore, without too much loss in practice [17], a memoryless channel is assumed in our equivalent channel. The *transition probabilities*, which determine the probabilities of a received bit combination $\hat{\mathbf{x}}_\ell$, given a possibly transmitted bit combination $\mathbf{x}_\ell^{(i)}$ with $i = 0, 1, \ldots, 2^M - 1$, can be formulated by

$$\mathrm{P}(\hat{\mathbf{x}}_\ell | \mathbf{x}_\ell^{(i)}) = \prod_{m=0}^{M-1} \mathrm{P}(\hat{x}_\ell(m) | x_\ell^{(i)}(m))\,, \tag{3}$$

with $x_\ell^{(i)}(m)$ being a possibly transmitted bit. Furthermore, using (1) and (2), $\mathrm{P}(\hat{x}_\ell(m)|x_\ell^{(i)}(m))$ can be computed as

$$\mathrm{P}(\hat{x}_\ell(m)|x_\ell^{(i)}(m)) = \begin{cases} 1 - p_{e,\ell}(m), & \text{if } \hat{x}_\ell(m) = x_\ell^{(i)}(m), \\ p_{e,\ell}(m), & \text{else.} \end{cases}$$

2.2. *A Posteriori* Probabilities

Conventional hard-decision decoding adopts inverse bit mapping in the speech decoder only utilizing the bit combination $\hat{\mathbf{x}}_\ell$ that is received by the decoder. No soft information obtained from the channel is exploited during hard-decision decoding. In contrast, for soft-decision decoding, bit error probabilities are employed to compute *a posteriori* probabilities $\mathrm{P}(\mathbf{x}_\ell^{(i)}|\hat{\mathbf{x}}_\ell, \hat{\mathbf{x}}_1^{\ell-1})$, which describe probably transmitted bit combinations $\mathbf{x}_\ell^{(i)}$ given the bit combinations $\hat{\mathbf{x}}_1^\ell = (\hat{\mathbf{x}}_\ell, \hat{\mathbf{x}}_1^{\ell-1}) = (\hat{\mathbf{x}}_\ell, \hat{\mathbf{x}}_{\ell-1}, \ldots, \hat{\mathbf{x}}_1)$ received in frames $1, 2, \ldots, \ell$. Generally, the quantized parameter can be regarded as an output of an Nth-order Markov process. Accordingly, a 0th-order Markov process leads to 0th-order *a priori* knowledge $\mathrm{P}(\mathbf{x}_\ell^{(i)})$ (AK0), a 1st-order Markov process leads to 1st-order *a priori* knowledge $\mathrm{P}(\mathbf{x}_\ell^{(i)}, \mathbf{x}_{\ell-1}^{(j)})$ or $\mathrm{P}(\mathbf{x}_\ell^{(i)}|\mathbf{x}_{\ell-1}^{(j)})$ (AK1), where $\mathbf{x}_\ell^{(i)}$ and $\mathbf{x}_{\ell-1}^{(j)}$ denote the bit combinations from the current and previous frames, respectively (the term j has the same range as i). In order to obtain this *a priori* knowledge, a large speech database is required to be processed by the speech encoder beforehand. Thereafter, the occurrence frequency distribution of different pairs of output symbols is counted and normalized. The AK0 term can then be obtained from

$$\mathrm{P}(\mathbf{x}_\ell^{(i)}) = \sum_{j=0}^{2^M-1} \mathrm{P}(\mathbf{x}_\ell^{(i)}, \mathbf{x}_{\ell-1}^{(j)})\,. \tag{4}$$

According to the chain rule, the AK1 term can be computed as

$$\mathrm{P}(\mathbf{x}_\ell^{(i)}|\mathbf{x}_{\ell-1}^{(j)}) = \frac{\mathrm{P}(\mathbf{x}_\ell^{(i)}, \mathbf{x}_{\ell-1}^{(j)})}{\sum_{k=0}^{2^M-1} \mathrm{P}(\mathbf{x}_\ell^{(k)}, \mathbf{x}_{\ell-1}^{(j)})}\,. \tag{5}$$

Accordingly, applying AK0 knowledge as calculated in (4), the *a posteriori* probabilities can be computed as

$$\mathrm{P}(\mathbf{x}_\ell^{(i)}|\hat{\mathbf{x}}_\ell, \hat{\mathbf{x}}_1^{\ell-1}) = \frac{1}{C} \cdot \mathrm{P}(\hat{\mathbf{x}}_\ell|\mathbf{x}_\ell^{(i)}) \cdot \mathrm{P}(\mathbf{x}_\ell^{(i)})\,, \tag{6}$$

with C normalizing the sum over the *a posteriori* probabilities to one. Correspondingly, applying AK1 as calculated in (5), the *a posteriori* probabilities can be computed, respectively, as [17]

$$\begin{aligned} \mathrm{P}(\mathbf{x}_\ell^{(i)}|\hat{\mathbf{x}}_\ell, \hat{\mathbf{x}}_1^{\ell-1}) = & \frac{1}{C} \cdot \mathrm{P}(\hat{\mathbf{x}}_\ell|\mathbf{x}_\ell^{(i)}) \cdot \sum_{j=0}^{2^M-1} \mathrm{P}(\mathbf{x}_\ell^{(i)}|\mathbf{x}_{\ell-1}^{(j)}) \\ & \cdot \mathrm{P}(\mathbf{x}_{\ell-1}^{(j)}|\hat{\mathbf{x}}_{\ell-1}, \hat{\mathbf{x}}_1^{\ell-2})\,. \end{aligned} \tag{7}$$

2.3. Parameter Estimation in AMR-WB

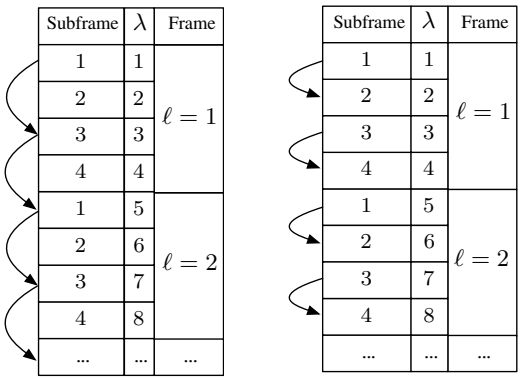
After the *a posteriori* probabilities for each received bit combination in every frame or subframe have been determined, each parameter can be estimated according to the corresponding quantization codebook either by minimum mean-square error (MMSE) estimation or by maximum *a posteriori* (MAP) estimation [31]. There are nine speech coding modes in AMR-WB [1] with bit rates ranging from 6.6 kbps to 23.85 kbps. We focus our investigations on the 12.65 kbps mode. As will be shown, soft-decision decoding of other modes can be implemented straightforwardly.

Each 20-ms frame consists of four subframes. As mentioned before, speech signals are analyzed in the AMR-WB encoder to extract various parameters. These parameters include the immittance spectral pair (ISP) vectors $\mathbf{u}_\ell$ and a voice activity detector (VAD) flag for each frame, and adaptive codebook indices (pitch delay) v_λ, an LTP filtering flag (except for the two lowest bitrate modes), fixed (algebraic) codebook indices, the vector-quantized adaptive codebook gain (pitch gain) and the fixed codebook gain $\mathbf{w}_\lambda$ for each subframe.

2.3.1. Immittance Spectral Pair (ISP) Vector

For quantization purposes, the linear predictive (LP) filter coefficients are transformed to immittance spectral pairs (ISPs). A combination of multistage vector quantization (MSVQ) and split vector quantization (SVQ) is used to quantize the residual ISP vector.

[1] Without loss of generality, in the Section 2.1 and Section 2.2, we assume a frame-based parameter (index ℓ) being transmitted.



(a) 512×512 AK1 matrix (b) 512×64 AK1 matrix

Fig. 2: Training of the pitch delay *a priori* knowledge: Generation of the AK1 matrix for odd (a) and even (b) subframes.

Quantization is done as follows: The residual ISP vector consisting of 16 samples is split into two subvectors of 9 and 7 dimensions, respectively. Each subvector is quantized in two different stages, both with 8 bits in the first stage. In the second stage, the quantization error vectors are split into three subvectors (6 bits, 7 bits, and 7 bits, each with 3 dimensions, respectively) and two subvectors (3 and 4 dimensions, both with 5 bits) [1]. As a result, according to (6), (7), we have in total seven subvectors with 256, 256, 64, 128, 128, 32, 32 *a posteriori* probabilities, respectively. The simulation is done for all seven subvectors in parallel, therefore (6), (7) are computed seven times in each frame with $M \in \{8, 8, 6, 7, 7, 5, 5\}$, respectively. The results from several simulations using MMSE and MAP estimation showed that the appropriate estimator for a certain ISP subvector $\mathbf{u}'_\ell$ is the MMSE estimator, which is given as:

$$\hat{\mathbf{u}}'_\ell = \sum_{i=0}^{2^M-1} \mathbf{u}'^{(i)}_\ell \cdot \mathrm{P}(\mathbf{x}^{(i)}_\ell | \hat{\mathbf{x}}_\ell, \hat{\mathbf{x}}_1^{\ell-1}) \,. \tag{8}$$

2.3.2. Adaptive Codebook Index (Pitch Delay)

For the pitch delay in the AMR-WB 12.65 kbps mode, a number of 9 bits and 6 bits, respectively, are used for *odd* and *even* subframes. In *odd* subframes, a fractional pitch delay with a total of 512 values in the range of $[34, 231]$ is used.

In *even* subframes, the relative difference to the previous frame (i. e., an odd subframe) is encoded, with the pitch delay being in the range $[T_1 - 8, T_1 + 7\frac{1}{2}]$, with an interval of $\frac{1}{2}$, and T_1 being the nearest integer to the pitch delay of the previous subframe, respectively [1]. Accordingly, 512 and 64 *a posteriori* probabilities are required for odd and even frames, respectively. In order to achieve better quality, the right-hand-side past *a posteriori* probability term in (7) is taken from *the last odd (9 bits) subframe*, for computing the left-hand-side *a posteriori* probability in (7) for both odd and even subframes. As a result, a rectangular AK1 matrix with the size of 512×64 is used for even subframes, while the size of the AK1 matrix in odd subframes remains 512×512. Correspondingly, the 1st-order *a priori* knowledge $\mathrm{P}(\mathbf{x}^{(i)}_\ell | \mathbf{x}^{(j)}_{\ell-1})$ from (7) is required to be computed separately in a training process as presented in Fig. 2: The *odd* subframes are counted in pairs ($\lambda = 1$ and $\lambda = 3$; $\lambda = 3$ and $\lambda = 5$; ...) for the calculation of the 512×512 AK1 matrix; the *even* subframes are counted in pairs ($\lambda = 1$ and $\lambda = 2$; $\lambda = 3$ and $\lambda = 4$; ...) for the calculation of the 512×64 AK1 matrix.

The equation of computing *a posteriori* probabilities of *odd* subframes ($\lambda = 1, 3, 5, \ldots$) is as follows (recursion as (7)!) :

$$\begin{aligned} \mathrm{P}(\mathbf{x}^{(i)}_\lambda | \hat{\mathbf{x}}_\lambda, \hat{\mathbf{x}}_1^{\lambda-2}) = & \frac{1}{C} \cdot \mathrm{P}(\hat{\mathbf{x}}_\lambda | \mathbf{x}^{(i)}_\lambda) \cdot \sum_{j=0}^{511} \mathrm{P}(\mathbf{x}^{(i)}_\lambda | \mathbf{x}^{(j)}_{\lambda-2}) \\ & \cdot \mathrm{P}(\mathbf{x}^{(j)}_{\lambda-2} | \hat{\mathbf{x}}_{\lambda-2}, \hat{\mathbf{x}}_1^{\lambda-4}) \,, \end{aligned} \tag{9}$$

with $i \in \{0, 1, \ldots, 511\}$ and $\hat{\mathbf{x}}_\lambda$ being the received bit combination of the *odd* subframe at the current global subframe index λ.

The equation of computing *a posteriori* probabilities of *even* subframes ($\lambda = 2, 4, 6, \ldots$) is as follows (no recursion!) :

$$\begin{aligned} \mathrm{P}(\mathbf{y}^{(i)}_\lambda | \hat{\mathbf{y}}_\lambda, \hat{\mathbf{x}}_1^{\lambda-1}) = & \frac{1}{C} \cdot \mathrm{P}(\hat{\mathbf{y}}_\lambda | \mathbf{y}^{(i)}_\lambda) \cdot \sum_{j=0}^{511} \mathrm{P}(\mathbf{y}^{(i)}_\lambda | \mathbf{x}^{(j)}_{\lambda-1}) \\ & \cdot \mathrm{P}(\mathbf{x}^{(j)}_{\lambda-1} | \hat{\mathbf{x}}_{\lambda-1}, \hat{\mathbf{x}}_1^{\lambda-3}) \,, \end{aligned} \tag{10}$$

with $i \in \{0, 1, \ldots, 63\}$ and $\hat{\mathbf{y}}_\lambda$ being the current received bit combination of the *even* subframe at index λ, and $\hat{\mathbf{x}}_{\lambda-1}$ being the bit combination of the previous *odd* subframe at index $\lambda-1$. Note that the right-hand-side *a posteriori* probability term in (10) equals the left-hand-side *a posteriori* probability term in (9) in the last subframe $\lambda-1$. Furthermore, the MAP estimator is adopted to compute the estimated pitch delay as:

$$\hat{\mathbf{v}}_\lambda = \mathbf{v}^{(i^{\mathrm{opt}})} \;\text{ with } i^{\mathrm{opt}} = \arg\max_i \mathrm{P}(i) \,, \tag{11}$$

with $\mathrm{P}(i)$ being the *a posteriori* probability either from (9) or (10).

2.3.3. Vector-Quantized Gain (VQ Gain)

The adaptive codebook gain and the correction factor, which is the ratio between an estimated algebraic codebook gain and the true algebraic codebook gain, are vector-quantized (two-dimensional vector quantization) using a 7-bit codebook in each subframe [1]. Therefore, 128 *a posteriori* probabilities are required to estimate $\hat{\mathbf{w}}_\lambda$ comprising both gain and correction factor. The estimation of $\hat{\mathbf{w}}_\lambda$ can be carried out using the MMSE estimator:

$$\hat{\mathbf{w}}_\lambda = \sum_{i=0}^{127} \mathbf{w}^{(i)}_\lambda \cdot \mathrm{P}(\mathbf{x}^{(i)}_\lambda | \hat{\mathbf{x}}_\lambda, \hat{\mathbf{x}}_1^{\lambda-1}) \,. \tag{12}$$

2.3.4. Other Parameters

Due to the high bit rate of the AMR-WB (algebraic) ACELP codebooks (12 ... 88 bits), it is too complex to compute *a posteriori* probabilities according to (6) or (7). Moreover, the AK0 and AK1 *a priori* knowledge are likely to be uniform due to the random nature of the fixed excitation, i. e., they do not exhibit a sufficient amount of residual parameter redundancy. Therefore, the fixed codebook index is simply hard-decision decoded as in [12]. The LTP filtering flag in each subframe and the voice activity detector (VAD) flag in each frame are not sensitive to errors. Therefore, we did not apply the soft-decision decoding approach to these parameters. However, if one wanted to apply our approach to these parameters as well, the MAP estimator (11) would be appropriate. Please note that for other AMR-WB codec modes, only the summation length in (8), (9), (10), and (12) must be adapted to the actual bit rate of the parameters.

3. REFERENCE AMR-WB ERROR CONCEALMENT

In order to reduce the effects of error-prone channels, error concealment is usually employed. The AMR-WB error concealment techniques proposed in 3GPP Recommendation TS 26.191 [12] mainly

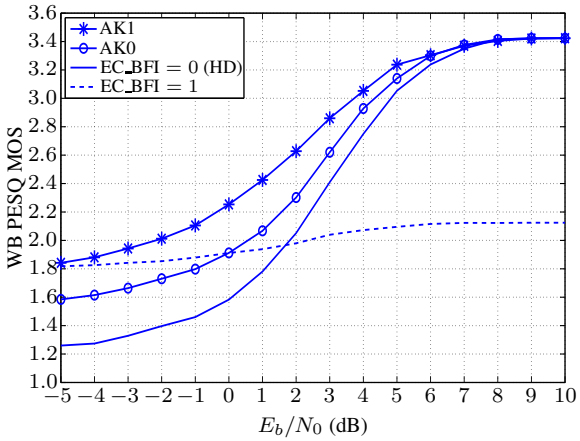


Fig. 3: Distortion of every 10th frame with the given E_b/N_0, $E_b/N_0 \to \infty$ for all other frames.

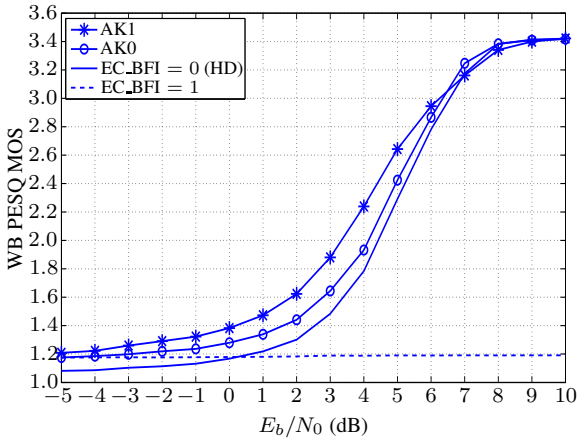


Fig. 4: Distortion of five consecutive frames with the given E_b/N_0, $E_b/N_0 \to \infty$ for the following five frames, in every ten frames.

rely on a bad frame indicator (BFI) to mark the current received frame either as good (BFI=0), or bad (BFI=1). If BFI is set to 1, the parameters of a speech frame are reconstructed according to previous good frames either using repetition or extrapolation. For the pitch delay and VQ gains, two relevant transmit frame types have to be distinguished: `SPEECH_BAD`, when the current frame is erroneous, and `SPEECH_LOST`, when the current frame is lost. Accordingly, different methods of substitution for the erroneous or lost case are applied, once BFI is set to 1. Moreover, whenever the current received frame is error-free but the BFI of the previous frame is set to 1, the fixed codebook gain of the last good frame will be used for the current frame at times. The fixed codebook index is employed as received (`SPEECH_BAD`), or randomly chosen (`SPEECH_LOST`). In short, based on the value of BFI, an erroneous frame, a lost frame, or an error-free frame can be distinguished. Therefore, the reliability of a frame is mainly determined by the BFI.

4. SIMULATIONS

4.1. Simulation Setup

Without too much loss of generality, we investigate the default 3GPP stream format [1] in the widely-used AMR-WB 12.65 kbps mode. The NTT wideband speech database with 16 kHz sampling rate is used for training and testing [32]. For training, 1926 speech signals each with a length of 8 s were applied, which covers 20 languages (except British English) each with 4 male speakers and 4 female speakers. A number of 96 British English speech signals including 4 male and 4 female speakers each of length 8 s are used for testing. The wideband extension to the perceptual evaluation of speech quality (WB PESQ) (P.862.2) [33] is selected as an instrumental measure for speech quality; the higher the MOS score of WB PESQ, the better the speech quality is. In order to generically demonstrate the robustness of our AMR-WB soft-decision decoding approach, we decided to simulate a *two-state* additive white Gaussian noise (AWGN, $a = 1$) channel and binary phase-shift keying (BPSK) without channel coding. The E_b/N_0 ratio in the bad channel state is varied between –5 dB to 10 dB. Two simulations were performed separately: In the first simulation, all bits in every 10th frame of the bit stream are distorted in our channel model, according to the chosen E_b/N_0 ratio; in all other frames $E_b/N_0 \to \infty$ (see Fig. 3). In the second simulation, five erroneous frames (E_b/N_0) followed by five error-free frames in every ten frames are simulated (see Fig. 4).

As shown in Figs. 3 and 4, the EC_BFI=0 (HD) curves with BFI=0 in the bad channel state represent hard-decision decoding, which means error concealment is not applied in this case. The EC_BFI=1 curves with BFI=1 in the bad channel state refer to the `SPEECH_BAD` case in the reference error concealment from 3GPP Recommendation TS 26.191. The AK0 curves represent soft-decision decoding with 0th-order *a priori* knowledge, the AK1 curves represent soft-decision decoding with 1st-order *a priori* knowledge.

4.2. Discussion

First of all, the EC_BFI curves in both figures indicate that in this simulation the optimum switch between BFI=1 and BFI=0 is in the range $E_b/N_0 \in [0.5\,\text{dB}, 1.5\,\text{dB}]$. As shown in Fig. 3, the MOS score of WB PESQ for AK1 is increased by up to 0.6 MOS points compared to the reference error concealment with optimal BFI. The corresponding improvement in MOS score for Fig. 4 falls between 0.2 and 0.4. The E_b/N_0 gain between soft- and hard-decision decoding reaches 4 dB or 6 dB, respectively. Particularly, for very low E_b/N_0 ratios in Fig. 4, the AK1 soft-decision decoding gain appears to be lower than in Fig. 3. A reason may be that five consecutive bad frames always drive the codec parameters already towards their means (AK1 performance approaches AK0 performance). Supported by informal listening tests, these results—especially for Fig. 4—show significant improvement in speech quality of soft-decision error concealment (AK0, AK1), compared to the reference error concealment methods EC_BFI. Moreover, if only a few bits in the frame are erroneous, while all other bits are error-free, the gap between soft-decision decoding and reference error concealment even gets larger. In contrast, since bit-wise channel reliability information is utilized in soft-decision decoding, MMSE or MAP estimation will automatically lead to the correctly estimated value in case of an error-free frame.

5. CONCLUSIONS

In this paper we described how to build an AMR-WB speech decoder capable of using soft inputs from the channel (decoder). We provided estimation formulae for all relevant codec parameters and demonstrated gains in robustness of up to 0.6 PESQ MOS points or up to 6 dB in channel SNR. The proposed approach is applicable in a standard-compliant fashion to replace the non-mandatory GSM/UMTS/LTE error concealment.

6. REFERENCES

[1] "Speech Codec Speech Processing Functions: Adaptive Multi-Rate-Wideband (AMR-WB) Speech Codec; Transcoding Functions (3GPP TS 26.190)," 3GPP; TSG-SA, Dec. 2001.

[2] T. Fingscheidt, "The Silent Speech Bandwidth Revolution in Mobile Telephony," IEEE Speech and Language Processing Technical Committee's Newsletter, www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2012-08/, Aug. 2012.

[3] "Global Mobile Suppliers Association," www.gsacom.com.

[4] "Mandatory Speech Codec Speech Processing Functions: AMR Speech Codec; Transcoding Functions (3GPP TS 26.090)," 3GPP; TSG-SA, Dec. 1999.

[5] "ITU-T Recommendation G.711 Appendix I, A High Quality Low-Complexity Algorithm for Packet Loss Concealment with G.711," ITU-T, Sept. 1999.

[6] "ITU-T Recommendation G.722 Appendix III, A High-Quality Packet Loss Concealment Algorithm for G.722," ITU-T, Nov. 2006.

[7] M. Serizawa and Y. Nozawa, "A Packet Loss Concealment Method Using Pitch Waveform Repetition and Internal State Update on the Decoded Speech for the Sub-Band ADPCM Wideband Speech Codec," in *Proc. of IEEE Workshop Speech Coding*, Tsukuba City, Japan, Oct. 2002, pp. 68–70.

[8] "Substitution and Muting of Lost Frames for Full Rate Speech Traffic Channels (GSM 06.11)," ETSI TC-SMG, Feb. 1992.

[9] "European Digital Cellular Telecommunications System Half Rate Speech Part 3: Substitution and Muting of Lost Frames for Half Rate Speech Traffic Channels (GSM 06.21)," ETSI TM/TM5/TCH-HS, Jan. 1995.

[10] "Digital Cellular Telecommunications System: Substitution and Muting of Lost Frames for Enhanced Full Rate (EFR) Speech Traffic Channels (GSM 06.61)," ETSI TC-SMG, Feb. 1996.

[11] "Mandatory Speech Codec Speech Processing Functions: AMR Speech Codec; Error Concealment of Lost Frames (3GPP TS 26.091)," 3GPP; TSG-SA, Dec. 1999.

[12] "Adaptive Multi-Rate - Wideband (AMR-WB) Speech Codec; Error Concealment of Erroneous or Lost Frames; (3GPP TS 26.191)," 3GPP; TSG-SA, Mar. 2001.

[13] J. Hagenauer and P. Hoeher, "A Viterbi Algorithm with Soft-Decision Outputs and its Applications," in *Proc. of GLOBECOM*, Dallas, TX, USA, Nov. 1989, pp. 1680–1686.

[14] V. Cuperman, F.-H. Liu, and P. Ho, "Robust Vector Quantization for Noisy Channels Using Soft Decision and Sequential Decoding," *Europ. Trans. Telecomm.*, vol. 5, no. 5, pp. 7–18, Sept. 1994.

[15] N. Farvardin and V. Vaishampayan, "Optimal Quantizer Design for Noisy Channels: An Approach to Combined Source-Channel Coding," *IEEE Trans. Inf. Theory*, vol. 33, no. 6, pp. 827–838, Nov. 1987.

[16] T. Fingscheidt and P. Vary, "Error Concealment by Softbit Speech Decoding," in *Proc. of ITG-Fachtagung "Sprachkommunikation"*, Frankfurt a.M., Germany, Sept. 1996, pp. 7–10.

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

[18] J. Hagenauer, "Source- Controlled Channel Decoding," *IEEE Trans. Commun.*, vol. 43, no. 9, pp. 2449–2457, Sept. 1995.

[19] T. Fingscheidt, "Graceful Degradation in ADPCM Speech Transmission," in *Proc. of DAGA*, Aachen, Germany, Mar. 2003.

[20] F. Pflug and T. Fingscheidt, "Delayless Soft-Decision Decoding of High-Quality Audio Transmitted Over AWGN Channels," in *Proc. of ICASSP'2011*, Prague, Czech Republic, May 2011, pp. 489–492.

[21] V. Ion and R. Häb-Umbach, "Uncertainty Decoding for Distributed Speech Recognition over Error-Prone Networks," *Speech Commun.*, vol. 48, no. 11, pp. 1435–1446, Nov. 2006.

[22] A. M. Peinado, V. Sánchez, J. L. Pérez-Córdoba, and A. J. Rubio, "Efficient MMSE-Based Channel Error Mitigation Techniques. Application to Distributed Speech Recognition Over Wireless Channels," *IEEE Trans. Wireless Commun.*, vol. 4, no. 1, pp. 14–19, Jan. 2005.

[23] T. Fazel and T. Fuja, "Robust Transmission of MELP-Compressed Speech: An Illustrative Example of Joint Source-Channel Decoding," *IEEE Trans. Commun.*, vol. 51, no. 6, pp. 973–982, Jun. 2003.

[24] M. Fresia and G. Caire, "A Linear Encoding Approach to Index Assignment in Lossy Source-Channel Coding," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1322–1344, Mar. 2010.

[25] M. E. Nasruminallah and L. Hanzo, "Robust Transmission of H.264 Coded Video Using Three-Stage Iterative Joint Source and Channel Decoding," in *Proc. of GLOBECOM*, Honolulu, HI, USA, Dec. 2009, pp. 1–6.

[26] F. Lahouti and A. K. Khandani, "Soft Reconstruction of Speech in the Presence of Noise and Packet Loss," *IEEE Trans. Audio, Speech, Language. Process.*, vol. 15, no. 1, pp. 44–56, Jan. 2007.

[27] T. Fingscheidt, T. Hindelang, R. V. Cox, and N. Seshadri, "Joint Source-Channel (De-)Coding for Mobile Communications," *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 200–212, Feb. 2002.

[28] N. S. Othman, M. El-Hajjar, O. Alamri, and L. Hanzo, "Soft-bit Assisted Iterative AMR-WB Source-Decoding and Turbo-Detection of Channel-Coded Differential Space-Time Spreading Using Sphere Packing Modulation," in *Proc. of IEEE Veh. Technol. Conf.*, Dublin, Ireland, Apr. 2007, pp. 2010–2014.

[29] N. S. Othman, M. El-Hajjar, O. Alamri, S. X. Ng, and L. Hanzo, "Iterative AMR-WB Source and Channel Decoding Using Differential Space-Time Spreading-Assisted Sphere-Packing Modulation," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 484–490, Jan. 2009.

[30] M. Adrat, P. Vary, and J.Spittka, "Iterative Source-Channel Decoder Using Extrinsic Information from Softbit-Source Decoding ," in *Proc. of ICASSP'2001*, Salt Lake City, UT, USA, May. 2001.

[31] J. L. Melsa and D. L. Cohn, *Decision and Estimation Theory*, McGraw-Hill Kogakusha, Tokyo, Japan, 1978.

[32] "Multi-Lingual Speech Database for Telephonometry," NTT-AT, 1994.

[33] "ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," ITU-T, Nov. 2007.