Steganographic Packet Loss Concealment for Wireless VoIP

Bernd Geiser, Frank Mertz, and Peter Vary[1]

Institute of Communication Systems and Data Processing (ind), RWTH Aachen University, 52056 Aachen, Germany
E-Mail: {geiser,mertz,vary}@ind.rwth-aachen.de
Web: www.ind.rwth-aachen.de

Abstract

We propose a method for the concealment of frame erasures in packet switched speech communication systems such as Voice over IP that is particularly tailored to the case of wireless transmission. At the receiver, the concealment algorithm is assisted by specific side information that is communicated via a steganographic channel within the bitstream of the employed speech codec. An exemplary implementation for the AMR codec is described and the obtained results are discussed.

1 Introduction

Currently, the infrastructure of voice communication networks is facing a major technological change towards *packet switched* transmission based on the Internet protocol stack (Voice over IP, VoIP). Since such networks, in contrast to traditional circuit switched connections, can only provide a "best effort" service, voice packets may be entirely lost or discarded due to an excessively delayed transmission time. This motivates the deployment of algorithms for receiver based packet loss concealment (PLC). Such algorithms, e.g., predict missing packets (corresponding to signal frames) using parameter-based inter- and extrapolation techniques. However, CELP speech codecs that are widely used for bandwidth-efficient communication are typically quite sensitive to frame erasures because of the predictive encoding of the speech parameters, in particular the adaptive codebook (ACB) parameters. This often leads to strong error propagation in case of erasures. Therefore, the transmission of suitable *side information* (SI) in subsequent packets can improve the concealment of a lost frame and limit the error propagation at the cost of additional system delay, see e.g. [6, 7, 3, 10, 9].

In this paper we consider the specific case that there is a *radio link* (e.g., WLAN or GSM/UMTS) within the transmission chain. In this case, the transmitted signal is additionally degraded by *bit errors*. If these errors cannot be corrected by the implemented channel coding scheme, typically the entire packet or frame is discarded. So, obviously, receiver based PLC is also required in this case. Moreover, if unequal error protection is employed, e.g., on circuit-switched GSM channels or in wireless packet networks using UDP-Lite (see, e.g., [8]), there may be *residual bit errors* within the less important bits of the bitstream of the encoded speech signal. Usually, this can be tolerated by the speech decoder. Considering the overall transmission chain in a heterogeneous network, e.g., for a call from an IP phone to a GSM cellphone (see Fig. 1), it is clear that the transmission is degraded by *both* packet losses and residual bit errors. To deal with such situations, we propose to employ a receiver based PLC algorithm that exploits suitable side information (SI). For this paper, we focus on the AMR [1] as the used speech codec and describe a PLC scheme which can:

a) correct (or detect) residual bit errors within the transmitted SI bitstream,

b) transmit the SI in a backwards compatible manner. This backwards compatibility is achieved by embedding the side information into the encoded speech signal using steganographic techniques.

The concept of transmitting SI via a *steganographic channel* within the bitstream of the employed speech codec provides several advantages over the conventional (separate) transmission of SI: (1) no additional bit rate is required, (2) standard payload formats can be used for the media packets, no particular payload format is required for piggybacking the SI, (3) the system is fully backwards compatible, i.e., the SI bitstream remains undetected by "legacy" terminals and is therefore ignored, and (4) with this technique, the SI can also be transmitted over conventional circuit-switched parts of the transmission chain, e.g. GSM or UMTS, if it can be ensured that no transcoding or tandeming operations occur during transmission.

The remainder of this paper presents an exemplary implementation of the described techniques for the AMR codec at a bit rate of 12.2 kbit/s. Section 2 introduces the implemented PLC mechanism as well as the additional SI that will be transmitted using the steganographic channel. The respective data hiding technique for the AMR codec is reviewed in Sec. 3. Finally, the entire system is evaluated and the results are discussed in Sec. 4.

2 PLC with Side Information

In [9], we proposed different low complexity extra- and interpolation techniques for the estimation of the AMR Wideband [2] parameters. From this set, the optimal technique for individual frames is determined at the sender and transmitted as SI (type 1). Knowing which estimation technique the receiver will use in case of loss, the coarsely quantized estimation error can be optionally transmitted as additional information to further improve the concealment (type 2). These types of SI require a lower additional bit rate than approaches that transmit copies of speech parameters. For the current paper, the mentioned approaches have been transferred to the narrowband AMR codec [1] and are shortly reviewed in this section.

Spectral Envelope (LSFs) — The Line Spectral Frequencies (LSFs) of the lost frame $\mathbf{q}(n)$ are estimated by interpolation from the previous and following frame using the interpolation factor $\alpha \in [0,1]$, which is quantized and transmitted as SI together with the following frame, such that the interpolation becomes:

$$\hat{\mathbf{q}}(\mathbf{n}) = \hat{\alpha}\,\mathbf{q}(n-1) + (1-\hat{\alpha})\,\hat{\mathbf{q}}(n+1) \tag{1}$$

Since the LSFs are encoded predictively, $\mathbf{q}(n+1)$ is not explicitly known, but can be estimated as shown in [9], leading to the closed form interpolation

$$\hat{\mathbf{q}}(n) = \overline{\mathbf{q}} + a\cdot\mathbf{r}(n-2) + b\cdot\mathbf{r}(n-1) + c\cdot\mathbf{r}(n+1)\,; \tag{2}$$

with

$$a = \frac{\alpha}{\alpha+2},\; b = \frac{10\alpha-1}{3\alpha+6},\; c = \frac{3-3\alpha}{2+\alpha}, \tag{3}$$

the received residual LSF vector $\mathbf{r}$, and the mean LSF vector $\overline{\mathbf{q}}$. For the interpolation factor α, a choice of 4 values ($\hat{\alpha} \in \{0.3, 0.5, 0.7, 0.9\}$) has been proven sufficient. The SI for the spectral envelope therefore requires 2 bit/frame, i.e. 100 bit/s.

Pitch Lag — The optimal choice among the following estimation techniques is transmitted as SI, requiring 3 bit per frame: (i) standard concealment, (ii) linear interpolation, and (iii) a replacement approach, where the first N sub-frames will be estimated by the preceding value and the remaining sub-frames by the first pitch value of the following frame. In voiced segments, where a correct pitch lag is crucial, the estimation using the method indicated by the SI is often correct, sometimes only

[1] Authors appear in alphabetical order.

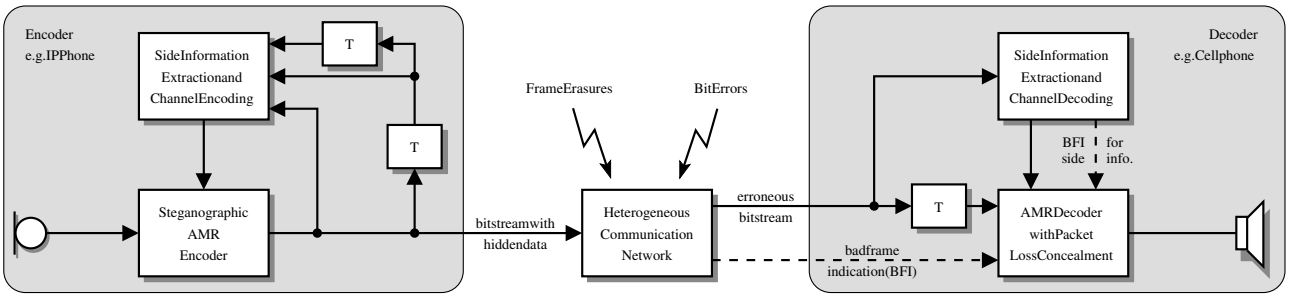

Figure 1: System concept: Transmission of side information for packet loss concealment as hidden data within the bitstream of a speech codec (e.g. AMR). Transmission channel: heterogeneous communication network with frames erasures (e.g. due to packet losses or bit errors) and residual bit errors in the codec bits. Exemplary application scenario: voice call from IP phone to GSM cellphone.

differing by a small value. Further improvement can be achieved by transmitting a correction $\hat{e}_{T_0} \in \{-3, -2, -1, 0, 1, 2, 3, 4\}$ for the pitch lag of each subframe, encoded with 3 bit each. If the correct value cannot be achieved by the adjustment, no correction is used ($\hat{e}_{T_0} = 0$). The quantized estimation error can be transmitted with an additional 12 bit/frame (i.e., 600 bit/s), requiring a total of 750 bit/s for the pitch lag SI.

Adaptive Codebook Gain — In the standard concealment method, the gains are estimated by attenuated values from the previous sub-frames. While signal muting is necessary in cases of several consecutive frame losses, for short losses of 1-2 frames this attenuation leads to noticeable and unnecessary amplitude fluctuations in voiced speech segments. The optimal choice among the following estimation techniques is therefore transmitted as SI for the adaptive codebook gain, encoded with 3 bit/frame: (i) linear interpolation, (ii) sub-frame replacement by previous or following gains, or (iii) replacement by the mean gains of the previous or following frame. Further improvement can be achieved with additional 6 bit/frame to transmit the quantized estimation error vector for the 4 gains of a frame.

Fixed Codebook Vector and Gain — As in the standard concealment method, a random entry is chosen from the fixed codebook. Although this vector is usually not the correct one, it still serves well as noise signal for synthesizing unvoiced speech segments. However, the fixed codebook also contributes significantly at transitions and voiced onsets. In order to prevent artifacts when the fixed codebook contribution is estimated by a random codebook entry, the fixed codebook gain is estimated as in the standard method, i.e. attenuated, and no SI is transmitted for this parameter.

The codebooks for the quantized estimation errors of pitch and gain parameters have been trained on a large speech database different from the one used in the simulations presented in Sec. 4.

Considered SI setups — The transmission of the optimal estimation technique for each of these parameters (SI type 1) requires 8 bit/frame and results in an SI bitstream of 400 bit/s (SI setup i)). The quality can be further improved by transmitting the quantized estimation errors for the pitch lag and adaptive codebook gains (SI type 2), leading to 26 bit/frame, or 1.3 kbit/s in total (SI setup ii)). This rate for assisting PLC is comparable to that used in other approaches, e.g. the 0.7 kbit/s employed in the ITU G.729.1 codec.

3 Data Hiding Scheme

This section reviews the ACELP (algebraic CELP) data hiding mechanism from [4] which allows to hide steganographic data with 2 kbit/s = 40 bit/frame in the bitstream of the 12.2 kbit/s mode of the AMR codec.

3.1 ACELP Steganography

In order to maintain the speech quality of the coder, the steganographic bits are in general embedded in less important parts of the encoded bitstream, e.g. in the fixed codebook (FCB) contribution of the AMR codec. The impact of the hidden bits on the speech quality is minimized by a *joint* implementation of the speech encoding and data hiding operations, cf. [11]. The key to this "ACELP steganography" is a modified search strategy for the ACELP codebook. Therefore, we first need to define the "message" m that shall be embedded into a 5 ms subframe. Here, m is given as a 10 bit binary sequence which is split into five sub-messages with two bits each. The sub-messages are denoted by, e.g. $(m)_{0,1}$ for the first two bits of m. To enable the transmission of $N = 10$ steganographic bits, the ACELP codebook (or fixed codebook, FCB) is partitioned into $M = 2^{10}$ *sub-codebooks* that uniquely identify the selected message m. Based on the standard ACELP search method from [1], the proposed steganographic algorithm has been derived in two steps:

1. Codebook Partitioning — First, M disjoint sub-codebooks are established by appropriately restricting the set of admissible codevectors. In particular, a specific parity condition is imposed on certain parts of the AMR bitstream:

$$(m)_{2k,2k+1} = \left[\mathscr{G}\left(\left\lfloor \frac{i_k}{5} \right\rfloor\right) \oplus \mathscr{G}\left(\left\lfloor \frac{i_{k+5}}{5} \right\rfloor\right)\right] \mod 4, \quad (4)$$

for the ACELP pulse positions i_k with $k \in \{0, \ldots, 4\}$. $X \oplus Y$ is the bitwise exclusive disjunction (XOR) of two binary strings and $\mathscr{G}$ represents the standardized Gray encoding of the ACELP pulse position codewords. At the decoder, the hidden information can be retrieved directly from the AMR bitstream using (4).

2. Search Space Expansion — Based on the chosen codebook partitioning, an FCB search strategy is devised that provides a good trade-off between speech quality and computational complexity. Thereby, the "admissible values" for the pulse positions i_{k+5} can be computed by solving (4) for i_{k+5}. More details on this steganographic FCB search can be found in [4].

3.2 Impact on Speech Quality

It appears plausible that the embedding of 2 kbit/s into a 12.2 kbit/s bitstream ($\approx 16\%$ of the codec rate) should have some impact on the quality of the coded speech. At first sight, an objective analysis with the PESQ speech quality measurement tool [5] seems to confirm that there is a certain quality impairment (see results below). Yet, in the subjective ABX experiment from [4] (11 experienced listeners, quiet environment, diotic presentation, multiple playback allowed, four trials per sample), only 162 out of 264 votings (61%) correctly identified X as either A or B. The options A and B have been randomly assigned to "standard AMR speech" and "speech with 2 kbit/s of hidden data".

3.3 Impact of Bit Errors and Channel Coding

If the (modified) AMR bitstream is transmitted over a network which tolerates bit errors in less important bits of the codec bitstream (unequal error protection), e.g., a circuit-switched GSM network or a wireless packet network employing UDP-Lite (see, e.g., [8]), such residual bit errors will inevitably lead to errors in the extracted SI bitstream at the receiver. Hence, a dedicated error protection for the SI bits is advisable, in particular because the SI bitstream exhibits an increased sensitivity to bit errors, resulting from the combination of two AMR bits according to (4).

Depending on how many steganographic bits per frame are remaining, the error protection can either be a simple cyclic redundancy check (CRC) for error detection, or, e.g. a channel code for error correction with an additional CRC for detecting residual errors. The implemented channel coding scheme actually follows this two-step approach. The main component is a suitably shortened BCH block code to protect the SI bits in each speech frame. Furthermore, before BCH encoding, a CRC is added to the information bits to detect residual errors which could not be corrected or detected by the BCH decoder. Suitable (n,k)-BCH codes with codeword length $n = 40$ exist for an information length of $k \in \{1,7,13,16,22,28,34\}$. For example, in case of an SI length of $l_{\mathrm{SI}} = 8$ bit, the (40, 13)-code would allow to allocate $k - l_{\mathrm{SI}} = (13-8)$ bit $= 5$ bit to the CRC. For $l_{\mathrm{SI}} = 26$, the (40, 28)-code offers the best error correction capabilities, but there is only room for 2 CRC bits.

The receiver extracts 40 bit of encoded SI per frame from the received AMR bitstream. Then, the BCH decoder, using Berlekamp's algorithm, attempts to correct any bit errors in the SI bitstream. In case the Berlekamp algorithm fails to decode the message, the flag $\mathrm{BFI}_{\mathrm{BCH}}$ is set. After BCH decoding, the CRC is recomputed and compared with the received CRC bits. The flag $\mathrm{BFI}_{\mathrm{CRC}}$ is set if the CRC fails. In addition to the decoded SI bits, the decoder outputs a bad frame indicator (BFI) which is related to the SI bits. To minimize the "false acceptance rate", this BFI is computed by a logical OR of $\mathrm{BFI}_{\mathrm{BCH}}$ and $\mathrm{BFI}_{\mathrm{CRC}}$.

4 Results

The proposed transmission of side information for PLC using a steganographic channel within the codec bitstream has been evaluated by the following simulations.

4.1 Simulation Setup

For the simulations we considered the following two scenarios:

a) The speech signal is transmitted from a Voice over IP (VoIP) terminal over a packet based network — wireline (e.g. DSL, LAN) or wireless (e.g. WLAN, UMTS) — with a certain packet loss rate to another VoIP terminal.

b) The speech signal is transmitted from the VoIP terminal over a packet network to a gateway where the encoded speech frames are extracted from the IP packets and without de- and re-encoding handed to a cellular GSM network for transmission over a circuit-switched channel to a mobile terminal.

GSM bit level simulations — The GSM channel has been simulated with a bit level reference implementation in the Synopsys System Studio Software, assuming a transmission frequency of 900 MHz and a user terminal moving with 50 km/h. The "typical urban" channel model has been chosen with an additional AWGN impairment with different E_b/N_0 ratios. In GSM, the speech codec bits are sorted into classes of sensitivity before channel coding and transmission. If residual bit errors are detected in the most important bits, the frame is marked unusable by setting the BFI (bad frame indication) flag. BFI and bit error patterns have been generated for E_b/N_0 ratios from 8-11 dB. The resulting frame erasure and residual bit error rates are shown in Fig. 2.

Steganographic VoIP simulations — The simulations were made with a speech database of 8 different English speakers (4 male, 4 female). There were 192 files in total, each about 8 s long, i.e., about 25 min of speech. Each file was encoded by the 12.2 kbit/s mode of the AMR speech codec. In this process, the SI for the concealment was generated according to setup i) and ii) described in Sec. 2, protected by the channel coding scheme presented in Sec. 3.3, and embedded into the bitstream as explained in Sec. 3.1. Random packet losses were introduced with a packet loss rate of 0 %, 3 %, or 6 %, respectively. For scenario b), the bit error patterns generated from GSM bit level simulations were added and those frames were marked unusable which had bit errors in the class of most sensitive codec bits (i.e., for which the BFI from the GSM channel decoder was set). The received speech frames, i.e., those not lost in the packet transmission and not marked unusable because of corrupted sensitive bits, are then decoded by the AMR decoder. The lost or unusable frames have been estimated by the respective concealment routine. The concealment utilizes the SI extracted from the following frame if available and not corrupted, otherwise the standard concealment routine of the AMR codec is used. The resulting speech quality has been determined by the PESQ speech quality measurement tool [5] and subjective evaluations by the authors.

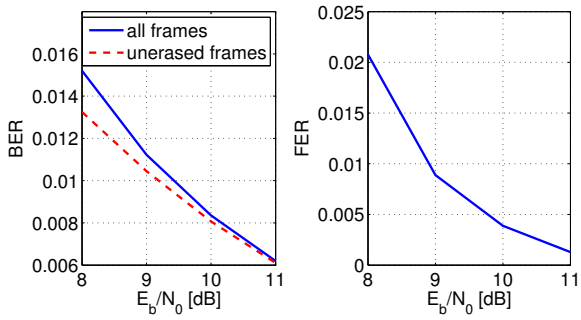

Figure 2: GSM bit level simulation: Frame erasure rate (FER) and residual bit error rate (BER) – in all frames and in the unerased frames only – on channels with different E_b/N_0 ratios.

4.2 Simulation Results

The simulation results depicted in Fig. 3 show the estimated MOS-LQO values measured by PESQ for different SI setups (line styles) in dependence on the quality of the GSM channel (given as E_b/N_0 in dB). In addition, different packet loss rates on the packet transmission channel have been considered (plr $= 0\%, 3\%,$ and 6%), distinguished by different line colors and markers in Fig. 3. The values obtained on a channel with only packet losses and no bit errors (scenario a)) are shown as horizontal lines without markers. These lines therefore serve as upper bounds for the results of scenario b).

The upper horizontal lines (blue) show the base quality achieved on a channel without bit errors and without packet losses. The MOS-LQO values determined by PESQ show a quality loss for the cases of embedding SI, although the subjective impression reveals no audible difference as discussed in Sec.3.2. There is no dependence on the amount of SI because it is always filled up by error protection bits to reach a total of 2 kbit/s for embedding. For increasing packet loss rates, as shown by the red (middle) and green (lower) horizontal lines, a considerable quality increase by the utilization of SI can be observed which is also confirmed by auditory impression. The higher the loss rate, the higher the quality gain has been. The use of a higher SI rate (26 bit/frame) leads to further quality improvement over the use of 8 bit/frame.

The increase of quality due to utilization of SI can also be consistently shown when considering a GSM bit error channel in addition to the packet losses. This is shown by the respective curves (red and square marker for 3% packet loss rate, green and circle marker for 6% packet loss rate). However, here a dependence on the amount of SI can be observed. The higher SI rate,

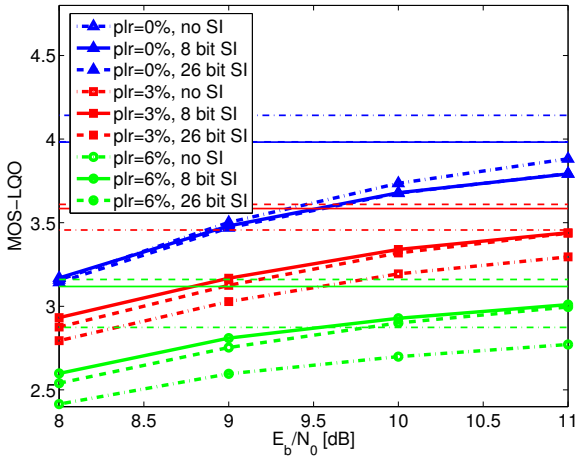

Figure 3: MOS-LQO values as measured by PESQ for different amounts of SI on heterogeneous channels with packet loss (IP network) and bit errors (GSM channel); SI setups i) with 8 bit/frame and ii) with 26 bit/frame; GSM channel: E_b/N_0 from 8 to 11 dB; IP network: packet loss rates of 0%, 3%, and 6%. For comparison: horizontal lines show values for channels with packet losses only, no bit errors.

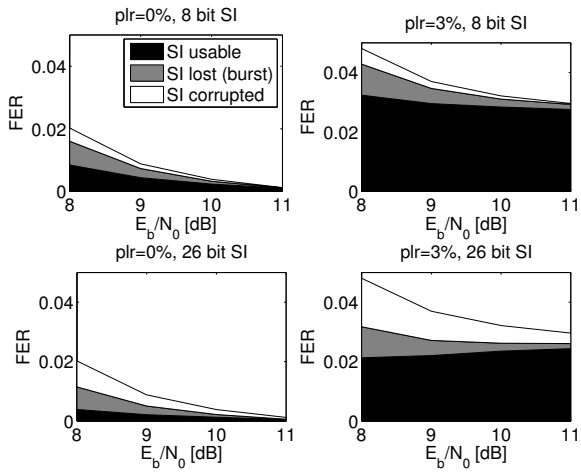

Figure 4: Total frame erasure rates (FER) including IP and GSM channel and percentage of usable SI and unusable SI for these frames; different E_b/N_0 on the GSM channel; additional packet loss rate of 0% (left column) and 3% (right column); comparing different amount of SI: 8 bit SI (upper graphs) and 26 bit SI (lower graphs).

which had shown a quality improvement for the pure packet loss channels, now leads to a lower quality than the low SI rate. The reason for this behavior is the higher sensitivity to residual bit errors, because in case of a higher SI rate less bits are available for the BCH-code and therefore less errors can be corrected. This is confirmed by the higher percentage of losses for which the SI from the following frame is unusable because of uncorrectable bit errors, as shown in Fig. 4 (compare the upper to the lower graphs).

When the packet loss rate approaches 0% (blue curves, triangle markers in Fig. 3), the transmission is mainly affected by bit errors on the GSM channel. Here, the benefit of utilizing SI decreases as, for the majority of losses, the SI has been unusable, either because of burst losses, i.e., loss of two or more successive frames, or because of uncorrectable bit errors, as shown in Fig. 4, left column.

5 Conclusions

This paper proposed the idea to transmit SI for packet loss concealment *within* the bitstream of the AMR speech codec by using steganographic techniques. In (wireless) packet-switched networks, this approach improves the robustness of the codec against packet losses without requiring additional bit rate. Furthermore, a transparent end-to-end transmission of such SI even over adjacent circuit-switched networks (GSM, UMTS) is possible as long as there is no transcoding involved. However, a sufficient protection of the SI bits by channel coding techniques is required if the end-to-end transmission may leave residual bit errors in the codec bitstream. The proposed concept is inherently backwards compatible because "legacy" terminals without support for the assisted PLC will automatically ignore the hidden information while the impact on speech quality remains negligible.

References

[1] 3GPP TS 26.090. AMR speech codec; transcoding functions, April 2000.

[2] 3GPP TS 26.190. AMR wideband speech codec; transcoding functions, December 2001.

[3] Yannis Agiomyrgiannakis and Yannis Stylianou. Coding with Side Information Techniques for LSF Reconstruction in Voice Over IP. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelphia, PA, USA, March 18-23 2005.

[4] Bernd Geiser and Peter Vary. High Rate Data Hiding in ACELP Speech Codecs. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, USA, March 2008.

[5] ITU-T Rec. P.862. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.

[6] I. Johansson, T. Frankkila, and P. Synnergren. Bandwidth efficient AMR operation for VoIP. In *IEEE Workshop on Speech Coding*, Tsukuba, Ibaraki, Japan, October 2002.

[7] R. Lefebvre, G.T. Philippe, and R. Salami. A study of design compromises for speech coders in packet networks. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 265–8, 17-21 May 2004.

[8] Frank Mertz, Ulrich Engelke, Peter Vary, Hervé Taddei, and Imre Varga. Applicability of UDP-Lite for Voice over IP in UMTS Networks. In *16th IEEE International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC 2005)*, Berlin, Germany, Sep. 11-14 2005.

[9] Frank Mertz and Peter Vary. Packet Loss Concealment with Side Information for Voice over IP in Cellular Networks. In *ITG-Fachtagung Sprachkommunikation*, Kiel, Germany, April 2006.

[10] Levent Tosun and Peter Kabal. Dynamically Adding Redundancy for Improved Error Concealment in Packet Voice Coding. In *13th European Signal processing Conference. EUSIPCO 2005*, Antalya, Turkey, Sep. 4-8 2005.

[11] Peter Vary and Bernd Geiser. Steganographic wideband telephony using narrowband speech codecs. In *Conference Record of Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, November 2007. invited talk.