Voicing Controlled Frame Loss Concealment for Adaptive Multi-Rate (AMR) Speech Frames in Voice-over-IP

Frank Mertz [*], *Hervé Taddei* [†], *Imre Varga* [†], *Peter Vary* [*]

[*] Institute of Communication Systems and Data Processing (IND),
Aachen University (RWTH), Germany
[†] Siemens AG, Information and Communication Mobile, Munich, Germany
{mertz|vary}@ind.rwth-aachen.de, {herve.taddei|imre.varga}@siemens.com

Abstract

In this paper we present a voicing controlled, speech parameter based frame loss concealment for frames that have been encoded with the Adaptive Multi-Rate (AMR) speech codec. The missing parameters are estimated by interpolation and extrapolation techniques that are chosen in dependence of the voicing state of the speech frames preceding and following the lost frames. The voicing controlled concealment outperforms the conventional extrapolation/muting based approach and it shows a consistent improvement over interpolation techniques that do not distinguish between voiced and unvoiced speech. The quality can be further improved if additional information about the predictor states of predictively encoded parameters is available from a redundant transmission in future packets.

1. Introduction

In *Voice-over-IP* (VoIP) applications packet loss leads to the loss of one or more successive speech frames and thereby to a degradation of the speech quality. The amount and length of these losses may be controlled to some extent by sender driven approaches, e.g. a transmission of redundant information in following packets and interleaving of speech frames. In addition, an effective concealment routine will have to be implemented in the receiver to deal with losses of speech frames. In this paper we present results of our studies on the concealment of missing speech frames that have been encoded with the Adaptive Multi-Rate (AMR) codec. The inherent ACELP codec structure is very sensitive to frame losses, as the predictive encoding of parameters like the LSF coefficients and the fixed codebook gain leads to error propagation. Additionally, an inaccurate estimation of a lost frame will affect the decoding of following frames due to incorrect adaptive codebook entries.

Because of varying transmission delays, VoIP applications require the use of a receiver buffer (so-called jitter buffer). If, in case of frame losses, the frame following the lost frames has already been received, it may be utilized by a frame loss concealment routine. The utilization of frames succeeding a loss for interpolation techniques has already been discussed, e.g. in [1], [2], and [3]. However, so far the proposed methods were applied regardless of the current signal structure. Our studies have shown that the achievable quality of the concealment routine highly depends on the properties of the lost signal segment. Therefore, we use a state controlled loss concealment that depends on the voicing state of the speech frame preceding and following the lost frames. The missing parameters are estimated by interpolation and extrapolation techniques that are chosen in dependence of the voicing state, i.e., whether the lost frames lie within a voiced, an unvoiced, or a transitional speech segment. Since the pitch lag parameter exhibits the level of periodicity in the current signal segment, the classification is solely based on this parameter, and it is thus of low complexity.

The problem of different sensitivities of encoded speech frames to loss has also been addressed in [3], where the use of a *Forward Error Correction* (FEC) scheme is proposed to limit the effects of a loss of these frames. This is achieved by transmitting redundant information in following packets. In this paper we concentrate on methods to conceal the effect of frame loss by estimating the missing speech parameters without additional redundant information on lost frames. Nevertheless, any additional (partial) information on the missing speech frames may be utilized to improve the performance of the proposed concealment methods. For the predictively encoded parameters of the AMR codec, i.e. the line spectral frequencies (LSF coefficients) and the fixed codebook gain, we will show that the knowledge of the missing predictor states may improve the performance.

This paper is structured as follows. Section 2 describes the method used for classification of frames into voiced and unvoiced speech. In Section 3 the concealment methods are detailed that we propose to use in dependence of the voicing transition in the speech signal. Section 4 presents simulation results that show the improved performance of the proposed methods, and finally, Section 5 concludes the paper.

2. Voicing classification

The choice of an appropriate concealment method strongly depends on the current periodicity of the speech signal around the lost frames. Therefore, the T_0 parameter, describing the pitch lag and thereby the fundamental speech frequency, is used as a measure of the periodicity for a classification into voiced and unvoiced speech frames. As shown in Figure 1, the parameter T_0 undergoes only slight variations in voiced regions of the speech, whereas in unvoiced segments T_0 has an unpredictable and rather random behavior, showing great value differences between successive subframes. This results from the missing periodicity in unvoiced speech segments.

Thus, the absolute values of the differences of T_0 in consecutive subframes are used as an indication for voiced/unvoiced speech. A decision function

$$V(n) = \sum_{i=1}^{3} |T_0^{(i)}(n) - T_0^{(i+1)}(n)| \qquad (1)$$

is computed for each frame n, with $T_0^{(i)}(n)$ the pitch lag of

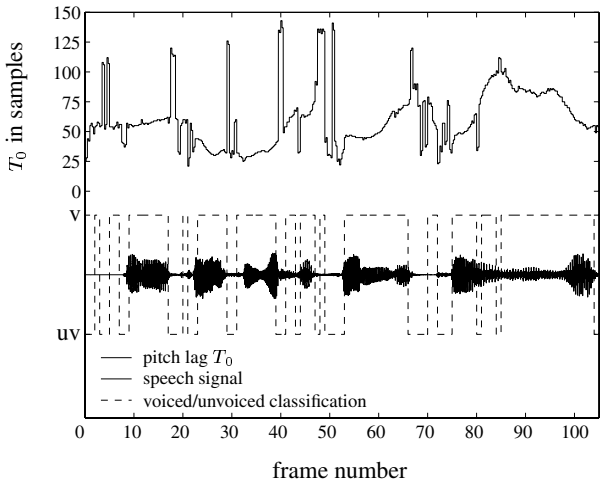

Figure 1: *Illustration of voiced/unvoiced classification. The pitch lags T_0 (upper solid line) are used to classify the frames (20 ms) of the given speech signal (lower solid line) into voiced (v) and unvoiced (uv) segments (illustrated by the dashed line).*

subframe i ($i = 1, \ldots 4$). With an appropriate threshold V_{th} for $V(n)$, a classification in voiced and unvoiced frames is done as follows:

$$\text{Frame is } \begin{cases} \text{voiced} & \text{for } V(n) \leq V_{th} \\ \text{unvoiced} & \text{for } V(n) > V_{th} \end{cases} . \qquad (2)$$

For narrowband speech (sampling frequency 8 kHz) a threshold of $V_{th} = 10$ proved to be suitable.

The classification sometimes detects an unvoiced speech frame within a voiced region. In these cases the speech structure undergoes a significant change within a voiced sound, expressed in a jump of the T_0 parameter that causes $V(n)$ to exceed the threshold. Since the voicing controlled choice of an appropriate concealment method, that will be presented in Section 3, is based on the periodicity of the speech signal, these cases have not to be considered as misclassifications, but they in fact support the concealment of lost frames in that speech segment.

3. Frame loss concealment

The concealment methods described in this section utilize received frames on both sides of the lost frames. Depending on the classification of both the preceding and the following speech frame, a particular concealment method is chosen for each different voicing transition. The concealment is based on the codec parameters. The parameters of missing frames, LSF coefficients, pitch lag, gain factors, and innovation vector (i.e. fixed codebook entry), are estimated by extra- and interpolation techniques.

From simulations it has been found to be beneficial to linearly interpolate the LSF coefficients as proposed in [2], regardless of the voicing state. This method shall be briefly reviewed in the following section, before the voicing controlled estimation of the remaining parameters is discussed.

3.1. LSF Interpolation

The AMR codec encodes the LSF coefficients predictively and transmits only the mean-removed relative coefficients. The AMR decoder calculates the absolute LSF coefficients from the received relative coefficients of the current and the previous frame. Therefore, error propagation is limited to one frame.

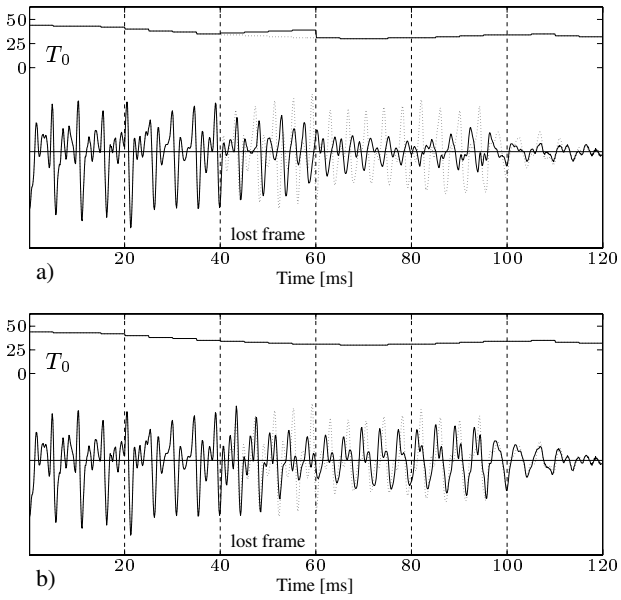

Figure 2: *Frame loss at voiced-voiced transition: decoded speech signals and respective pitch lags (solid line - lossy signal; dotted line - error-free signal) a) BFI concealment b) voicing controlled concealment*

The extrapolation/muting based concealment unit [4] of the AMR codec, as recommended for GSM, uses an extrapolation technique to replace the LSF coefficients in case of transmission errors. The LSF values are extrapolated from the last correctly received frame and slightly shifted towards the mean LSF vector. The conventional concealment unit of the AMR codec will be called *BFI concealment* in the following, since it is activated by setting the *Bad Frame Indication* (BFI) flag.

The linear interpolation of LSF coefficients in case of frame losses has already been investigated in [2]. Prior to interpolation, the absolute LSF coefficients of the lost frames are estimated as in the BFI concealment method providing the basis for decoding the LSF coefficients of the frame following the lost segment. A superior performance of the linear interpolation compared to the standard BFI concealment for LSF coefficients has been shown in [2] using the spectral distortion measure.

3.2. Transition *voiced-voiced*

Within a voiced region of speech the signal exhibits a strong periodicity. Expressing this periodicity, the pitch lag follows a rather smooth curve with only small variations. In the AMR codec the *pitch lag* parameter T_0 is encoded predictively only within a frame, there is no dependency between the values of successive frames. Therefore, the missing T_0 values can be linearly interpolated between the fourth subframe of the last received frame and the first subframe of the first received frame behind the loss. By this interpolation the original T_0 curve is more precisely estimated than in case of the conventional error concealment unit (BFI concealment), where the missing T_0 values are extrapolated from the last received frame by repeating the value of the fourth subframe, incremented by 1 for each successive lost subframe. The results of both methods can be seen from the upper curves in Figure 2 a) and b).

In the conventional error concealment unit, the *codebook gains* are both extrapolated from the previous frame and attenuated to prevent possible artifacts, which can be seen clearly from an exemplary signal in Figure 2. Simulations have shown that this attenuation of the signal amplitude is not necessary if the missing frames lie within a voiced region of speech, at least not until the lost segment gets too large. Therefore, the gain factors of both the adaptive and the fixed codebook are linearly interpolated at voiced-voiced transitions, avoiding an unnecessary fluctuation in the signal amplitude (see Figure 2 b)).

In case of BFI concealment, a random entry is chosen from the fixed codebook as *innovation vector*. In simulations it has been found to be beneficial to search instead for a segment of the innovation in the preceding frames that is similar to the subframe directly preceding the lost frames. This is done by using a cross-correlation measure. The innovation vector of the lost frames is then set to the continuation of the similar segment, thereby utilizing residual periodicities that are left in the innovation vector.

Figure 2 visualizes the performance of the proposed method compared to the standard BFI concealment on an exemplary single frame loss. With the interpolation approach the signal resembles more closely that of the error-free case than with BFI concealment. To verify the improvements gained by the proposed concealment method, the following simulation has been carried out. In a short speech signal (2.5 s), encoded with AMR mode 10.2 kbit/s, 5% single frame losses have been introduced within purely voiced segments. During decoding the given methods were used to conceal the lost frames and the decoded speech file has been compared to the original speech (i.e. before encoding) using the objective PESQ [5] measure. The obtained results, that are given in Table 1, clearly show the improvement by the proposed voicing controlled method.

Table 1: *PESQ comparison for 5% single frame losses at voiced-voiced transitions; AMR-mode 10.2 kbit/s.*

concealment method	PESQ-MOS
BFI concealment	2.960
linear interpolation	3.183
voicing controlled concealment	3.413

3.3. Transition *voiced-unvoiced*

When estimating speech parameters of lost frames at transitions of voiced to unvoiced speech, the *pitch lag* T_0 must not be interpolated to avoid an unnatural change in the fundamental frequency. This would occur when the first pitch lag in the following unvoiced speech frame strongly differs from that of the preceding voiced frame (see T_0 curves in Figure 3). Better results can be accomplished by extrapolating the pitch lag from the preceding voiced speech frame. The contribution of the *fixed codebook* is set to a random codebook entry as in the conventional BFI concealment unit. To mitigate possible artifacts at this transitions both *codebook gains* are treated as in the BFI concealment, i.e. extrapolated and attenuated. This results in a better subjective speech quality, even if in simulations the PESQ-MOS value has been higher when interpolating the codebook gains, as can be seen in Table 2.

3.4. Transition *unvoiced-voiced*

The transition from unvoiced to voiced speech is the most difficult position for the concealment of a lost frame. The transitional frames carry the very important information on how to

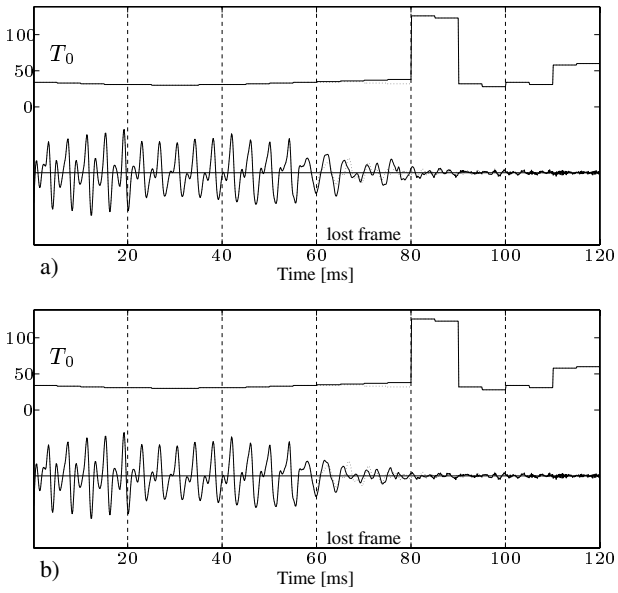

Figure 3: *Frame loss at voiced-unvoiced transition: decoded speech signals and respective pitch lags (solid line - lossy signal; dotted line - error-free signal) a) BFI concealment b) voicing controlled concealment with gain muting*

Table 2: *PESQ comparison for 5% single frame losses at voiced-unvoiced transitions; AMR-mode 10.2 kbit/s.*

concealment method	PESQ-MOS
BFI concealment	3.320
linear interpolation	3.317
voicing controlled & gain interpolation	3.398
voicing controlled & gain muting	3.324

build up the periodicity of the beginning voiced segment. In this respect the innovation vector (fixed codebook) is essential. Again, a linear interpolation of the *pitch lags* is not advisable because of a possibly large difference between the pitch lag of the beginning voiced frame and the more random values of the preceding unvoiced frame. For the same reason, the concept of the BFI concealment, i.e. repeating and incrementing the previous value, is also not recommended, since it would take the unvoiced pitch lag as basis. Therefore, we use a linear extrapolation technique that goes backward in time by using the following pitch lag to estimate the values of the lost frames. This will help to build up the periodicity of the voiced speech. The contribution of the *fixed codebook* is again set to a random codebook entry. When interpolating the *codebook gains*, the form and periodicity of the voiced signal is reached faster, as can be seen in Figure 4. However, for the subjective quality it is better to use an extrapolation and attenuation of the gains to mitigate possible artifacts at the transitions, even if the PESQ-MOS value is smaller (see Table 3).

Table 3: *PESQ comparison for 4% single frame losses at unvoiced-voiced transitions; AMR-mode 10.2 kbit/s.*

concealment method	PESQ-MOS
BFI concealment	2.262
linear interpolation	2.625
voicing controlled & gain interpolation	2.615
voicing controlled & gain muting	2.365

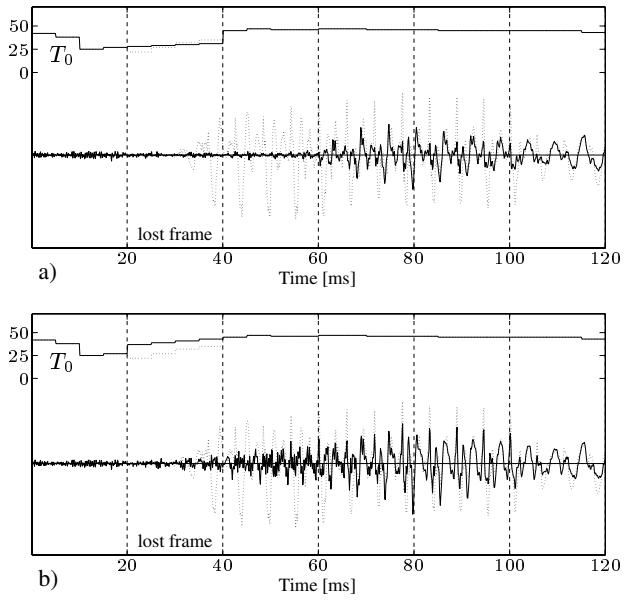

Figure 4: *Frame loss at unvoiced-voiced transition: decoded speech signals and respective pitch lags (solid line - lossy signal; dotted line - error-free signal) a) BFI concealment b) voicing controlled concealment with gain interpolation*

3.5. Transition *unvoiced-unvoiced*

The loss of a frame inside an unvoiced speech segment is rather uncritical for the expected speech quality. The noise like nature of unvoiced speech allows to estimate a missing frame fairly easily by a random noise sequence. Because of the missing periodicity in unvoiced speech the *pitch lag* shows a random behavior. It should be avoided to produce an unnatural periodicity, as it might result from repeating a previous pitch lag as done in BFI concealment. Therefore, the T_0 parameter set (4 for each frame) is repeated as block from the previous frame to preserve the random-like behavior. Both *codebook gains* are linearly interpolated and a random fixed codebook entry is chosen for the *innovation vector*.

The results that can be seen in Table 4 have been obtained by a simulation of double frame losses, because the performance of the methods did not much differ when simulating single frame losses. Considering longer loss lengths, it becomes clear that the voicing dependent method for unvoiced regions leads to better results than the other methods.

Table 4: *PESQ comparison for 5.6% double frame losses at unvoiced-unvoiced transitions; AMR-mode 10.2 kbit/s.*

concealment method	PESQ-MOS
BFI concealment	3.174
linear interpolation	3.384
voicing controlled concealment	3.419

4. Overall performance results

For an evaluation of the overall performance of the proposed voicing controlled concealment technique we ran several simulations on a speech file of 2 minutes length. We introduced random frame loss of about 2%, 3%, and 5%, respectively, occurring at random voicing transitions. The results in Table 5 clearly show the superior performance of the proposed technique over the conventional BFI concealment technique. It also performs consistently better than a linear interpolation technique that does not distinguish between the voicing states. The tendencies in the measured PESQ values have been confirmed by subjective impressions in informal listening tests.

Recent results have shown that a knowledge of the predictor states, e.g. by a redundant transmission in future packets, can further improve the resulting speech quality (see Table 5).

Table 5: *PESQ comparison for different concealment methods and channel conditions; signal length 2 minutes.*

concealment method	frame loss rate		
	2 %	3 %	5 %
BFI concealment	3.538	3.393	3.101
linear interpolation	3.682	3.571	3.392
voicing controlled concealment	3.716	3.623	3.439
voicing controlled concealment & gain muting at uv-v and v-uv	3.698	3.597	3.360
voicing controlled concealment & predictor states available	3.739	3.652	3.494

5. Conclusions

We have shown that a voicing controlled choice of concealment methods, that have been in particular designed for each voicing transition, consistently improves the performance of the concealment unit. The proposed method avoids unneeded signal muting in voiced and unvoiced speech segments and it mitigates artifacts at transitions between voiced and unvoiced speech segments by attenuating the codebook gains in these cases. For the voiced/unvoiced classification of speech frames, a simple and low complex method depending solely on the pitch lag parameter has been presented, that detects the current level of periodicity in the speech signal.

We have shown that the performance may be improved by providing additional information on the missing predictor states of the predictively encoded parameters (LSF coefficients and fixed codebook gain), e.g. by transmitting these state information as redundant data in following packets.

6. References

[1] J. Wang and J. D. Gibson, "Parameter interpolation to enhance frame erasure robustness of CELP coders in packet networks," in *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, Utah, USA, May 2001.

[2] T. Fingscheidt and J. G. Perez, "An Interpolative Decoding Approach for Speech Streaming Services and Voice Over IP," in *Proceedings of 4th International ITG Conference on Source and Channel Coding*, Berlin, Germany, Jan. 2002.

[3] I. Johansson, T. Frankkila, and P. Synnergren, "Bandwidth Efficient AMR Operation for VoIP," in *IEEE Workshop on Speech Coding*, Tsukuba, Ibaraki, Japan, Oct. 2002.

[4] ETSI, *Spec. GSM 06.91: Substitution and muting of lost frames for Adaptive Multi Rate (AMR) speech traffic channels*. European Telecommunications Standards Institute.

[5] ITU-T Recommendation 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," Genf, 2001.