

Speech Property-Based FEC for Internet Telephony Applications*

Henning Sanneck** and Nguyen Tuong Long Le

GMD Fokus, Kaiserin-Augusta-Allee 31, D-10589 Berlin, Germany

ABSTRACT

Recently we have seen research efforts on how to protect a real-time speech signal when transmitting over an unreliable packet-switched network like the Internet by open-loop error control. Research has covered the type of Forward Error Correction (generic or voice-specific), the protocol support needed and adaptivity to the current network congestion state.

However, the sender does not take into account that some segments of the signal are essential to the speech quality, while others can be extrapolated at the receiver from data received earlier in the event of a packet loss. This is especially true for modern frame-based codecs like the G.729 and G.723.1 which contain an internal loss concealment algorithm. Thus, the sender consumes additional bandwidth and aggravates the congestion in the Internet by sending unnecessary redundancy.

In this paper we first analyze the concealment performance of the G.729 decoder. We find that the loss of unvoiced frames can be concealed well. Also, the loss of voiced frames is concealed well once the decoder has obtained sufficient information on them. However the decoder fails to conceal the loss of voiced frames at an unvoiced/voiced transition because it extrapolates internal state (filter coefficients and excitation) for an unvoiced sound. Moreover, once the encoder has failed to build the appropriate linear prediction synthesis filter, it takes a long time for the decoder to resynchronize with the encoder.

Using this result, we then develop a new FEC scheme to support frame-based codecs, which adjusts the amount of added redundancy adaptively to the properties of the speech signal. Objective quality measures (ITU P.861A and EMBSD) show that our speech property-based FEC (SPB-FEC) scheme achieves almost the same speech quality as current FEC schemes while approximately halving the amount of necessary redundant data to adequately protect the voice flow.

Keywords: Internet Telephony, Forward Error Correction, Speech Properties, G.729, Objective Speech Quality Measurement

1. INTRODUCTION

In recent years, both the general public and the research community have been showing significant interest in interactive speech transmission over the Internet (Internet Telephony). Currently, the main incentive for Internet Telephony is the cheap flat-rate charge compared to usage-based charging for traditional telephony services. Although this cheap flat-rate charge might not persist in the future, speech transmission over the Internet is still very attractive because it can be integrated with other Internet applications to provide interactive multimedia communication services that are impossible (or at least very difficult) to deploy over the traditional telephone network. Furthermore, high complexity speech encoding and decoding can be performed with inexpensive hardware in the end systems at user premises. Examples are the two frame-based codecs G.723.1 ([9]) and G.729 ([10]), which are very attractive for Internet Telephony because they provide toll quality speech at much lower bit rates (5.3/6.3 kBit/s and 8 kBit/s respectively) than conventional PCM (64 kBit/s). Thus the network resource requirements for a large scale deployment can be reduced significantly.

However, today's packet-switched networks, like the Internet, are based on the "best effort" principle which does not guarantee a minimum packet loss rate and a minimum delay of packet transmission required for Internet Telephony. This results in adverse affects on the quality of Internet Telephony, e.g. speech packets can be discarded when routers or gateways are congested. Due to the real-time requirement for interactive speech transmission, it is usually impossible for the

* to appear in *Proceedings of the SPIE/ACM SIGMM Multimedia Computing and Networking Conference 2000 (MMCN 2000)*, San Jose, CA, January 2000

** Correspondence: Email: sanneck@fokus.gmd.de; WWW: <http://www.fokus.gmd.de/usr/sanneck>; Telephone: +49 30 3463 7175; Fax: +49 30 3463 8175

receivers to request the sender to retransmit the lost packets. Besides, voice packets that do not arrive before their playout time are considered lost and cannot be played when they are received. Furthermore, considering the backward-adaptive coding schemes of the G.723.1 and G.729 source coders, packet loss results in loss of synchronization between the encoder and the decoder. Thus, errors occur not only during the time period represented by the lost packet, but also propagate into following segments of the speech signal until the decoder is resynchronized with the encoder. To alleviate this problem, both G.723.1 and G.729 decoders contain an internal (codec-specific) loss concealment scheme.

To cope with the packet loss problem on an end-to-end basis, i.e. without modifying the network itself, much research has been done to develop schemes for open-loop error control for audio transmissions over the Internet ([3], [5], [14], [16]). Figure 1 illustrates the generic structure of audio tools with such mechanisms. In parallel to the conventional encoding and packetization process, an analysis module extracts redundant information, that is then sent either as a separate stream ([18]) or packetized together with the original data stream ("piggybacking", [15]). The redundant information can be generated either by generic Forward Error Correction (FEC) codes or by exploiting speech-specific properties. The amount of speech-specific information can range from a simple pitch period measurement in the AP/C scheme ([20])¹ over short-term energy and zero crossings to recover the basic envelope of the speech signal ([7]) up to running entire speech coders in parallel with an offset in time ([8]). Typically such schemes are complemented by concealment schemes, which try to recover missing speech segments from already decoded PCM samples ([19]).

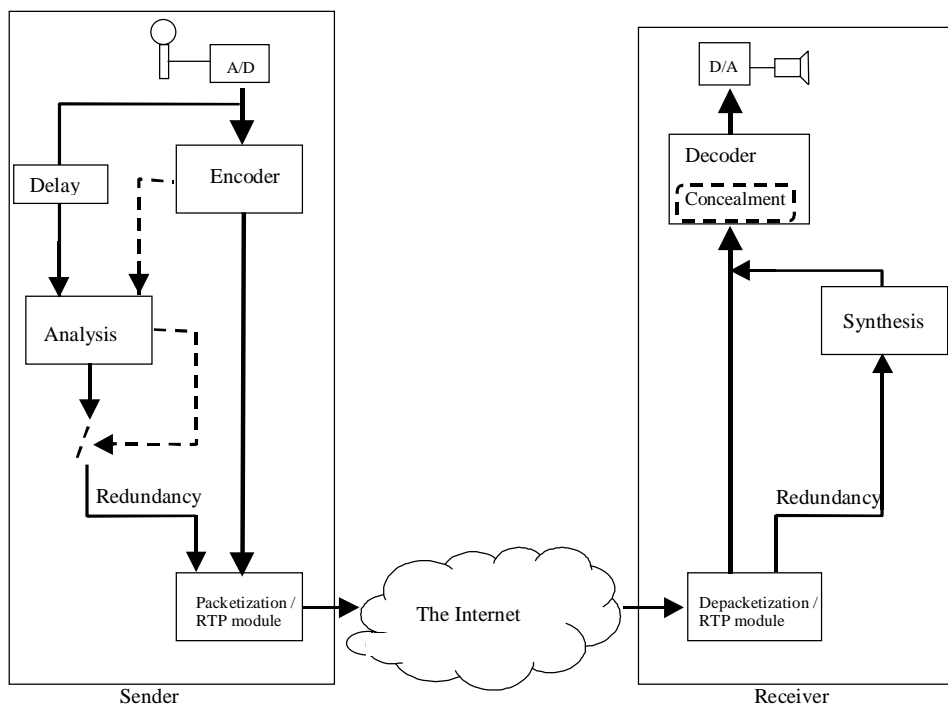


Figure 1 Generic structure of an audio tool with loss recovery.

In this paper, we adopt the approach of using a speech coder (G.729) as the analysis module ([15]), with the following changes (new components are printed with dashed lines in Figure 1):

- Information available at the encoder which can be used for the redundancy / loss recovery is exploited.
- No generic concealment ([19], [20]) schemes are employed, as codec-specific concealment is already implemented in the decoder.
- The amount of redundancy can be adjusted by the analysis module taking into account the decoder concealment process.
- Only one source coder for both the main and the redundant payload is used.

¹ Note that AP/C also influences the packetization lengths.

We only use one source coder to reduce the overall computational complexity. Additionally (if the primary and the redundant data of a packet are coded with different audio encodings and "piggy-backing" on the following packets is used), when an important frame is lost, all decoders suffer loss of synchronization and deliver decoded speech signals with worse quality. An example for this problem is illustrated in Figure 2 where the sender transmits PCM μ -law voice data as primary data and G.729 voice data as redundant data (we assume the transport of one frame per packet for simplicity). When a data packet arrives at the receiver, the PCM μ -law voice frame is played and the G.729 frame is passed to the G.729 decoder to keep it synchronized with the G.729 encoder at the sender. The output of the G.729 decoder for a frame is discarded if the PCM μ -law data for that frame is also received. If a packet is lost and the following packet is received, the G.729 frame is played to cover the gap in the PCM μ -law audio stream. However, because the G.729 decoder also just lost a frame ($n-1$ in Figure 2), it suffers a loss of synchronization, resulting in a worse quality of the speech signal decoded from the replacement frame (n in Figure 2).

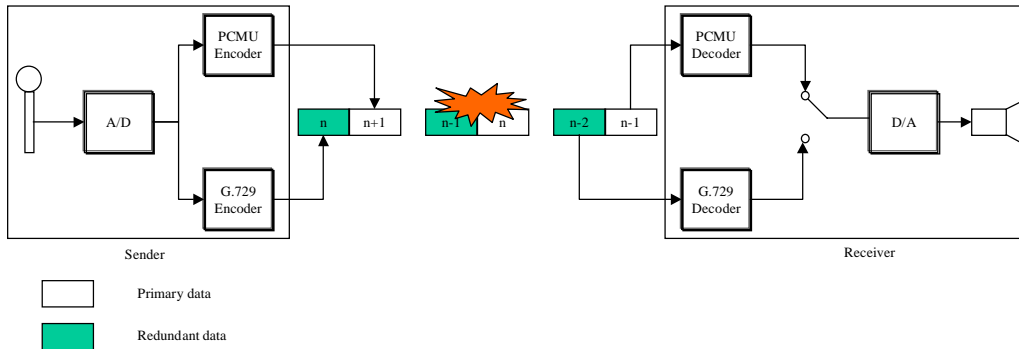


Figure 2 Loss of synchronization of the redundancy decoder during a packet loss.

The remaining sections of the paper are structured as follows: Section 2 presents a brief overview of the G.729 codec and our analysis of the performance of the internal loss concealment scheme with respect to the impact of packet loss at different areas of a speech signal. In section 3, we present the SPB-FEC scheme. Section 4 gives results that evaluate the efficiency of our SPB-FEC scheme using a simple network model and objective speech quality measures. Section 5 concludes the paper.

2. G.729 FRAME LOSS CONCEALMENT

G.729 is also known as Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP) and operates at 8 kBit/s. Input data for the coder are 16-bit linear PCM data sampled at 8 kHz. G.729 is based on a model for human speech production. In this model, the throat and the mouth have the function of a linear filter (synthesis filter) and speech signals are produced by exciting this filter with an excitation vector. In G.729, a speech frame is 10 ms in duration, corresponding to 80 PCM speech samples. For each frame, the G.729 encoder analyzes the input data and extracts the parameters of the Code Excited Linear Prediction (CELP) model such as linear prediction filter coefficients and excitation vectors. The approach for determining the filter coefficients and the excitation is called analysis by synthesis: The encoder searches through its parameter space, carries out the decode operation in each loop of the search, and compares the output signal of the decode operation (the synthesized signal) with the original speech signal. The parameters that produce the closest match are chosen, encoded, and then transmitted to the receivers. At the receivers, these parameters are used to reconstruct the original speech signals.

2.1. Encoder and Decoder Operation

For each 10-ms frame, the encoder performs a linear predictive analysis to compute the linear prediction filter coefficients. The linear predictive analysis is based on the idea that a speech sample can be approximated as a linear combination of past speech samples. By minimizing the sum of the squared differences between the actual speech samples and the approximated ones (over a certain number of speech samples), a set of filter coefficients can be found². A linear filter that has that set of filter coefficients is called the analysis filter, i.e. when a speech signal is passed through it, we get the excitation for that

² The filter coefficients are the weighting coefficients in the linear combination.

speech signal. The synthesis filter is obtained by inverting the analysis filter. When we filter the excitation through the synthesis filter, the result is an approximation to the original speech signal.

For the sake of stability³ and efficiency, the linear-prediction filter coefficients are not directly quantized but are transformed into line spectral pairs and quantized using a predictive two-stage vector quantization process⁴. The excitation for the speech signal is computed per 5-ms subframe (corresponding to 40 PCM speech samples) and has two components: fixed and adaptive-codebook. First, an open loop pitch delay is estimated once per 10-ms frame. This estimation is based on the auto-correlation of the weighted speech signal that is derived from filtering the speech signal through a perceptual weighting filter⁵. The adaptive-codebook contribution models the long-term correlation of speech signals and is expressed in a closed-loop pitch delay and a gain. The closed-loop pitch delay is searched for around the open loop pitch delay by minimizing the error between the perceptually weighted input signal and the previous excitation filtered by a weighted linear-prediction synthesis filter. The difference of the found excitation filtered by the synthesis filter and the original signal is then used to find the fixed-codebook contribution. The fixed-codebook vector and the fixed-codebook gain are searched by minimizing the mean-squared error between the weighted input signal and the weighted reconstructed speech signal, using a pulse train as excitation. The adaptive-codebook gain and the fixed-codebook gain are then jointly vector quantized using a two stage vector quantization process.

The G.729 decoder at the receivers extracts the following parameters from the arriving bit stream: the line spectral pair coefficients, the two pitch delays, two codewords representing the fixed-codebook vector, and the adaptive- and fixed-codebook gains. The line spectral pair coefficients are interpolated and transformed back to the linear prediction filter coefficients for each subframe. Then, for each subframe the following operations are performed:

- The excitation is the sum of the adaptive- and fixed-codebook vectors multiplied by their respective gains.
- The speech signal is obtained by passing the excitation through the linear prediction synthesis filter.
- The reconstructed speech signal is filtered through a post-processing filter that incorporates an adaptive postfilter based on the long-term and short-term synthesis filters, followed by a high-pass filter and scaling operation. These operations reduce the perceived distortion and enhance the speech quality of the synthesized speech signals by emphasizing the spectral peaks (formants) and attenuating the spectral valleys ([12]).

When a frame is lost or corrupted, the G.729 decoder uses the parameters of the previous frame to interpolate those of the lost frame and performs loss concealment to reduce the degradation of speech quality of the reconstructed speech signal. In particular, the following steps are taken:

- The line spectral pair coefficients of the last good frame are repeated.
- The adaptive- and fixed-codebook gain are taken from the previous frame but they are damped to gradually reduce their impact.
- If the last reconstructed frame was classified as voiced, the fixed-codebook contribution is set to zero. The pitch delay is taken from the previous frame and is repeated for each following frame. If the last reconstructed frame was classified as unvoiced, the adaptive-codebook contribution is set to zero and the fixed-codebook vector is randomly chosen.

When a frame loss occurs, the decoder cannot update its state, resulting in a divergence of encoder and decoder state. Thus, errors are not only introduced in the current frame but also in the following ones. In addition to the impact of the missing codewords, distortion is increased by the missing update of the following internal state parameters:

- The predictor filter memories for the line spectral pairs.
- The linear prediction synthesis filter memories.

2.2. Impact of frame loss at different positions

In [17], Rosenberg investigated the issues of error resiliency and recovery and measured the resynchronization time of the G.729 decoder after a frame loss. He pointed out that the energy of the error signal⁶ increases considerably and the Mean

³ A direct quantization may move some of the poles of the synthesis filter outside of the unit circle, resulting in an unstable synthesis filter.

⁴ In order to save bandwidth, the encoder and decoder predict the value of the line spectral pairs via a 4th order moving average. After prediction, the difference is computed and then vector quantized.

⁵ The perceptual weighting filter is based on the linear prediction filter coefficients and reflects the perceptual distortion of the reconstructed/synthesized speech signal.

⁶ The difference between the decoded signals with and without frame loss.

Opinion Score (MOS) of subjective tests decreases significantly when the number of consecutive lost frames increases from one to two, and gradually from there. He drew the conclusion that a single lost frame can be concealed well by the G.729 decoder but not more. In this section, we take a further step by attempting to answer the question: how does the speech quality degrade and how does the error propagate when a number of consecutive voiced/unvoiced frames are lost?

The first experiment⁷ we carry out is to measure the resynchronization time of the decoder after a number of consecutive frames are lost. We vary the position of the frame loss to cause a number of consecutive voiced/unvoiced frames to be lost and then count the number of the following frames until the signal-to-noise ratio (SNR) exceeds a certain threshold. The SNR is computed on a frame basis and is defined as:

$$SNR_{Frame} = 10 \cdot \log_{10} \left(\frac{\sum_n x(n)^2}{\sum_n (x(n) - x'(n))^2} \right) dB = 10 \cdot \log_{10} \left(\frac{\sum_n x(n)^2}{\sum_n e(n)^2} \right) dB \quad n \in [1, F]$$

where F is the frame length in samples, $x'(n)$ and $x(n)$ are the decoded signal with and without frame loss and $e(n)$ is the error signal (the difference between the decoded signals) with n being the sample index relative to the frame.

We consider 20 dB an appropriate value for the threshold⁸. That is the G.729 decoder is said to have resynchronized with the G.729 encoder after the loss of a number of frames when the energy of the error signal falls below one percent of the energy of the decoded signal without frame loss. Figure 3 shows the resynchronization time plotted against the loss position. The speech sample is produced by a male speaker where an unvoiced→voiced (uv) transition occurs in the eighth frame.

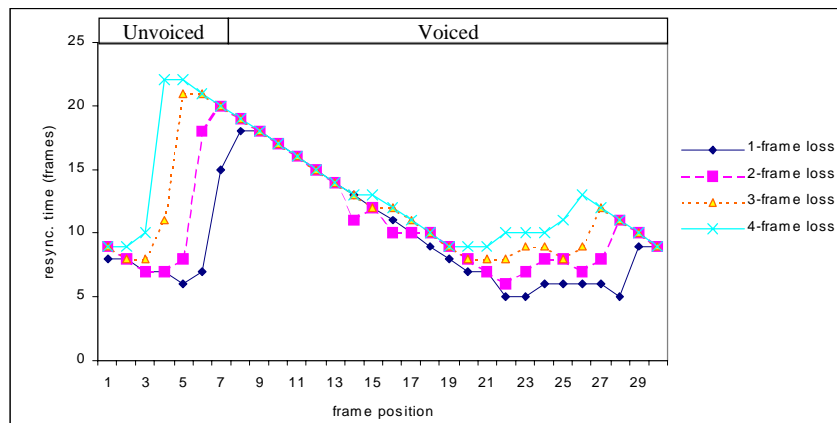


Figure 3 Resynchronization time (in frames) of the G.729 decoder after the loss of k consecutive frames ($k \in [1,4]$) as a function of frame position.

Our second experiment is to measure the energy of the error signal over a number N of frames after k consecutive frames are lost. We vary the position where the frame loss (burst) occurs and then compute the mean SNR over the N following frames. In our experiment, we measure the mean SNR over $N=15$ consecutive frames after the frame loss which we consider an appropriate mean value for the resynchronization time. We have also measured the mean SNR over 10 and 20 consecutive frames after the frame loss and obtained similar results. (Our first experiment has shown that the resynchronization time ranges from 5 to 22 frames depending on the position of the frame loss and the burst size. Previous experiments in [17] came to comparable results). Figure 4 shows the mean SNR plotted against the frame loss position for the same speech sample.

We can see from Figure 3 and Figure 4 that a loss of a consecutive number of frames at different positions has significantly different levels of impact on the error introduced into the speech signal and thus on speech quality. The loss of unvoiced

⁷ For all experiments in this paper we used the ITU reference implementation of the G.729 codec available at <http://www.itu.int/itudoc/itu-t/rec/g/g700-799/software/g729/>.

⁸ This threshold is also used in [17].

frames seems to have a rather small impact on the speech quality and the decoder recovers the state information fast thereafter.

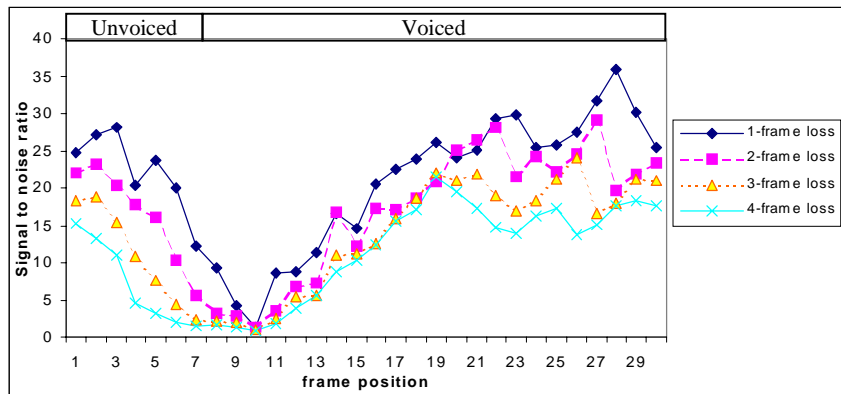


Figure 4 Mean SNR (dB) of the G.729-decoded speech signal after the loss of k consecutive frames ($k \in [1, 4]$).

However, the loss of voiced frames causes a larger degradation of the speech quality and the decoder needs more time to resynchronize with the sender. Moreover, the loss of voiced frames at an unvoiced/voiced transition leads to a very significant degradation of speech quality. We have repeated our two above experiments for different male and female speakers and obtained similar results. The above phenomenon could be explained as follows:

- Because voiced sounds have a higher energy and are also more important to the speech quality than unvoiced sounds, the loss of voiced frames causes a larger degradation of speech quality than the loss of unvoiced frames.
- Due to the periodic property of voiced sounds, the decoder can conceal the loss of voiced frames well once it has obtained sufficient information on them.
- The decoder fails to conceal the loss of voiced frames at an unvoiced/voiced transition because it attempts to conceal the loss of voiced frames using the filter coefficients and the excitation for an unvoiced sound. Moreover, because the G.729 encoder uses a moving average filter to predict the values of the line spectral pairs and only transmits the difference between the real and predicted values, it takes a lot of time for the decoder to resynchronize with the encoder once it has failed to build the appropriate linear prediction filter.

3. SPEECH PROPERTY-BASED FORWARD ERROR CORRECTION (SPB-FEC)

The experiments we have carried out in the previous section have shown that the loss of frames at the beginning of a voiced signal causes a significant degradation in speech quality and a frame-based decoder like the G.729 decoder can conceal the loss of other voiced segments well once it has obtained sufficient information on the voiced signal. The loss of unvoiced frames is also concealed well by the decoder. This knowledge is exploited to develop a new FEC scheme called Speech Property-Based FEC (SPB-FEC). In contrast to other FEC schemes that equally distribute the amount of redundant data on all data packets, the SPB-FEC scheme concentrates the amount of redundant data on the frames essential to the speech quality and relies on the decoder's concealment for other frames.

Figure 5 shows the simple algorithm written in a pseudo-code that is used to detect a uv transition and protect the voiced frames at the beginning of a voiced signal. In the algorithm, the procedure *analysis()* is used to classify a block of k frames as voiced, unvoiced, or uv transition⁹. Senders can either run a parallel algorithm for voiced/unvoiced decision or couple this algorithm with the encoder's operation. The first method is a generic approach (useful when coder-internal state cannot be accessed) but may duplicate functionality already available in the encoder and thus unnecessarily consume CPU resources. In our experiments we have chosen the second method. The voiced/unvoiced decision in G.729 is made in the decoder only however, so that the sender also has to run a decoder to decode its own frames and detect voiced/unvoiced transitions.

The procedure *send()* and *sendFEC()* are used to send a block of k frames and redundant data to protect these frames respectively. N is a pre-defined value and defines how many frames at the beginning of a voiced signal are to be protected. Our simulations have shown that the range from 10 to 20 are appropriate values for N (depending on the network loss

⁹ The voiced→unvoiced (vu) transition is unimportant in our algorithm and is classified as unvoiced.

condition). In the simulation presented in section 4, we choose $k=2$, a typical value for interactive speech transmissions over the Internet (20 ms of voice data per packet). A larger number of k would help to reduce the relative overhead of the protocol header but also increases the buffer delay and makes sender classification and receiver concealment in case of packet loss (due to a large loss gap) more difficult.

```

protect = 0
foreach (k frames)
    send(k frames)
    classify = analysis(k frames)
    if (protect > 0)
        if (classify == unvoiced)
            protect = 0
        else
            sendFEC(k frames)
            protect = protect-k
    endif
else
    if (classify == uv_transition)
        sendFEC(k frames)
        protect = N-k
    endif
endif
endfor

```

Figure 5 **SPB-FEC Pseudo Code**

4. EVALUATION OF THE SPEECH PROPERTY-BASED FEC SCHEME

4.1. Speech Quality Measurement

In general, there are two ways to measure the speech quality: subjective and objective tests. In subjective tests, listeners listen to a set of speech signals without being told about their nature. The listeners evaluate the quality of the speech signals by giving a score that typically ranges from 1 for bad to 5 for excellent quality. The listeners' scores are averaged, resulting in a Mean Opinion Score (MOS) that represents the speech quality. While subjective tests should be considered to be a very important tool to evaluate the performance of any speech-related system, they are time-consuming, expensive, error-prone and only difficult to reproduce.

In objective tests, speech quality is evaluated by measuring the distortion of the decoded speech signals compared to the original speech signals. The simplest objective test method is to use the Signal-to-Noise Ratio (SNR) to assess the speech quality. The most common ways to compute the SNR are: the overall (or "classical") SNR method that computes the SNR over the whole signal duration and the frame-based SNR method that computes the SNR on a frame basis and then averages the results. The frame-based SNR is known to provide a much better estimate of the subjective quality than the overall SNR ([6]). It has been feasible to employ the frame-based SNR for the experiments in section 2.2 because there we have examined only *one* system (G.729 without any FEC protection) under different error conditions. Now, however, we will compare two systems (G.729 with permanent or SPB-FEC) under the same error condition. The first system employs mathematically exact reconstruction where possible, whereas the second system relies much more on the internal concealment of the G.729 decoder, which is able to maintain a good output quality under the conditions described in section 2.2. However this effect cannot at all be captured by an SNR (e.g. the gradual dampening of the gain coefficients of the previously received frame during the loss concealment (as described in section 2.1) improves the speech quality, but lets the recovered signal largely deviate from the original signal in the mathematical sense).

Unlike the SNR methods, novel objective quality measures attempt to estimate the subjective quality as closely as possible by modeling the human auditory system. In our evaluation we use two objective quality measures: the Enhanced Modified Bark Spectral Distortion (EMBSD) ([24]) and the Measuring Normalizing Blocks (MNB) described in the Appendix II of the ITU-T Recommendation P.861 ([11]). These two objective quality measures are reported to have a very high correlation

with subjective tests and are suitable for the evaluation of speech degraded by transmission errors in real network environments such as bit errors and frame erasures ([24]).

4.2. Network Model

We use a simple network model (Gilbert model) to drop voice packets, which is well accepted for the modelling the Internet end-to-end packet loss process ([3], [4]). The model has two states reflecting whether the previous packet is received (state 0) or lost (state 1).

Let p be the probability for the network model to drop a packet given that the previous packet is delivered, i.e. the probability for the network model to go from state 0 to state 1. Let q denote the probability for the network model to drop a packet given that the previous packet is dropped, i.e. the probability for the network model to stay in state 1. This probability is also known as the *conditional loss probability (clp)*. Let p_0 and p_1 denote the probability of the network model to be in state 0 and state 1, we have:

$$\begin{aligned}
 p_1 &= p_0 \cdot p + p_1 \cdot q \\
 p_0 + p_1 &= 1 \\
 \Rightarrow p_0 &= \frac{1-q}{p+1-q} \quad , \quad p_1 = \frac{p}{p+1-q}
 \end{aligned}$$

The probability for a packet to be dropped regardless whether the previous packet is delivered or dropped, i.e. the *unconditional loss probability (ulp)*, is exactly the probability for the network model to be in state 1 (p_1). Figure 6 shows the Gilbert model with its transition probabilities.

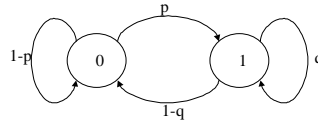


Figure 6 Gilbert model.

4.3. Reference FEC schemes

In general, there are two methods to send redundant data: in a separate flow ([18]) or “piggy-backed” on the following packets ([15]) containing the main payload. While the first method has the advantage of backwards compatibility, we choose the second method for our simulation because of the lower protocol header and router processing overhead. We use two other FEC schemes as reference to evaluate the SPB-FEC: In the first FEC scheme, the two frames of the packet (n) are piggy-backed on the packet ($n+2$) (we do not piggy-back the two frames of the packet (n) on the packet ($n+1$) to further mitigate the effect of packet burst loss, [3]). This FEC scheme has a redundancy overhead of 100%. In the second FEC scheme, the four frames of the packet (n) and ($n+1$) are XORed and the result is piggy-backed on the packet ($n+2$). If the packet ($n+2$) and one of the packets (n) or ($n+1$) arrive at the receiver, a lost packet can be recovered. This FEC scheme has a redundancy overhead of 50%.

Our speech property-based FEC scheme is similar to the reference FEC scheme 1. However, in our scheme, only when an unvoiced/voiced transition is detected, the FEC mechanism is turned on to protect the voiced frames at the beginning of a voiced signal, resulting in a redundancy overhead of 41.9% (for the speech material used in the experiments below). Figure 7 illustrates the two reference FEC schemes.

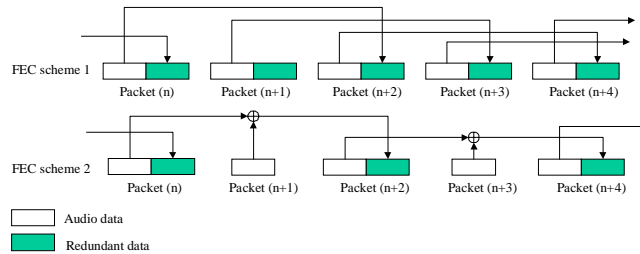


Figure 7 Two reference FEC schemes.

4.4. Simulation Description

We first simulate a network where voice data flows using packets containing two frames (i.e. 20 ms speech segments) without any redundant data are transmitted. We vary the network loss parameters p and q in constant steps to obtain an impression on the sensitivity and expected range of the objective quality measurements' result values (Figure 8 shows the network loss rate (unconditional loss probability) associated with the pairs of p and q in the first simulation step). The voice data flow with frame loss is decoded. The results are then compared to the decoded speech signal without frame loss using the objective quality measures.

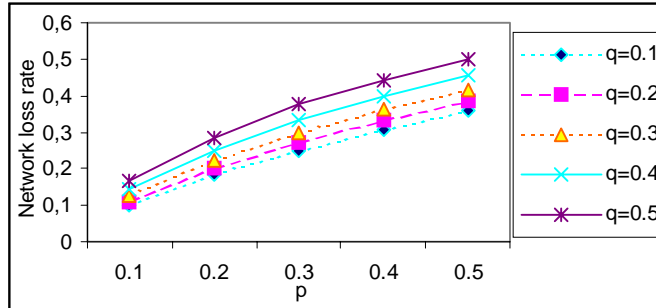


Figure 8 Network loss rate (unconditional loss probability) in simulation step 1.

In the second step, the simulated network is applied to voice data flows using our SPB-FEC scheme, the two reference FEC schemes described in section 4.3, and a scheme without redundant data respectively. Every speech data packet contains two frames and possibly some redundant data depending on the respective FEC scheme. We use five (p, q) value pairs reflecting real network loss conditions (Table 1) measured in the Internet ([4]). The FEC schemes are then used to recover the information contained in the lost packets to the largest extent possible.

Network loss condition 1	Network loss condition 2	Network loss condition 3	Network loss condition 4	Network loss condition 5
$p=0.05, q=0.2$	$p=0.1, q=0.3$	$p=0.15, q=0.4$	$p=0.2, q=0.5$	$p=0.25, q=0.6$

Table 1 Network loss condition parameters used in simulation step 2.

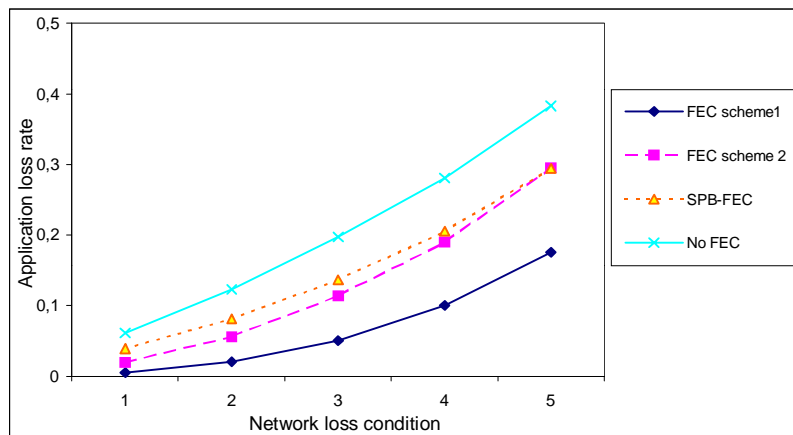


Figure 9 Application loss rate of different FEC schemes and network loss conditions.

Figure 9 shows the application loss rate of the schemes with and without FEC, i.e. the loss rate seen by the G.729 decoder after FEC decode (if any) has been performed for the five network loss conditions. Obviously, the more redundant data is

transmitted, the lower is the application loss rate¹⁰. Then, the voice data streams (possibly still with some frame losses) are decoded. These decoded speech signals and the decoded speech signal without frame loss are then evaluated by the objective quality measures to demonstrate the efficiency of the FEC schemes. The two simulation steps for the evaluation of the FEC schemes are illustrated in Figure 10.

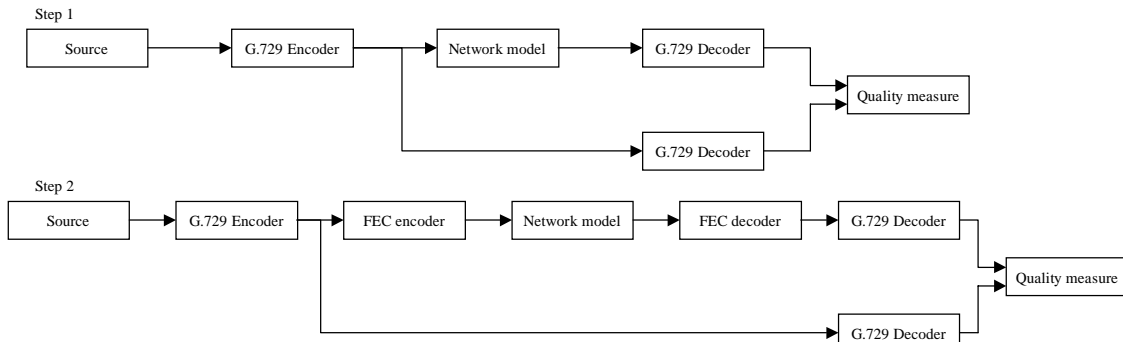


Figure 10 Simulation steps for the evaluation of the FEC schemes.

For each pair of p and q , we use the same speech sample¹¹ containing different male and female voices as input to our simulation but use different seeds for the pseudo-random number generator to generate different loss patterns. This is important because, as we have seen in section 2.2, different loss patterns can have largely different levels of impact on the speech quality, e.g. a loss pattern dropping only voiced frames would result in a worse speech quality than a loss pattern dropping only unvoiced frames. By averaging the result of the objective quality measures for several loss patterns, we have a reliable indication for the performance of the G.729 codec and the FEC schemes under a certain network loss condition.

4.5. Results

In MNB, the perceptual difference between the test signal and the reference signal is measured at different time and frequency scales. The perceptual difference, also known as Auditory Distance (AD), between the two signals is a linear combination of the measurements where the weighting factors represent the auditory attributes. The higher AD is, the more the two signals are perceptually different and thus the worse the speech quality of the test signal is. Figure 11 and Figure 12 show the auditory distance evaluated by MNB resulting from the two simulation steps.

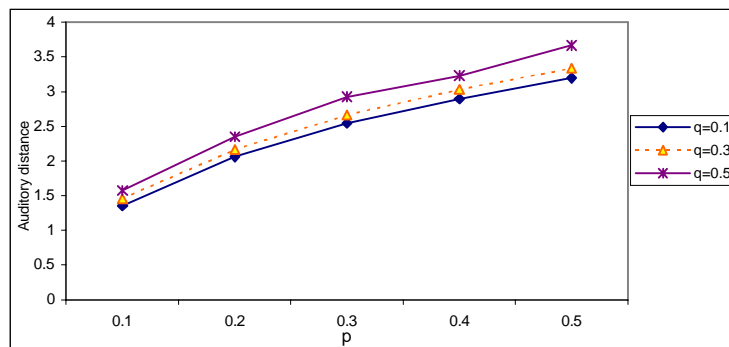


Figure 11 Auditory distance (simulation step 1) evaluated by MNB.

The Bark Spectral Distortion (BSD) measure ([22]) assumes that speech quality is directly related to the speech loudness which is defined as the perceived feeling for a given frequency and sound pressure level ([13], [23]). The BSD measure is the perceptual distortion computed as average squared Euclidean difference between the estimated loudness of the test and

¹⁰ We do not account for the increase in network congestion caused by the increase in bandwidth of the single flow under consideration.

¹¹ The length of the speech sample is 11.25 s. The sample, as well as samples used for the simulation step 2 can be obtained at <http://www.fokus.gmd.de/glone/products/voice/spb-fec>.

the reference signal. The Modified BSD measure (MBSD) defines the perceptual distortion as the estimated loudnesses' average difference and introduces a noise masking threshold below which perceptual distortion is not taken into account ([23]). The difference between the MBSD and the enhanced MBSD (EMBSD) is that a new cognition model based on postmasking effects and 15 loudness components are used, loudness vectors are normalized, and the spreading functions in the noise masking threshold calculation are removed in the EMBSD ([24]).

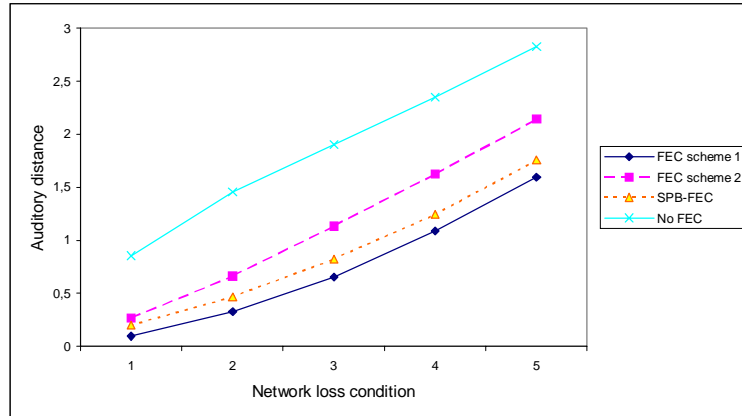


Figure 12 Auditory distance of FEC schemes evaluated by MNB.

Figure 13 and Figure 14 show the perceptual distortions evaluated by the EMBSD resulting from the two simulation steps.

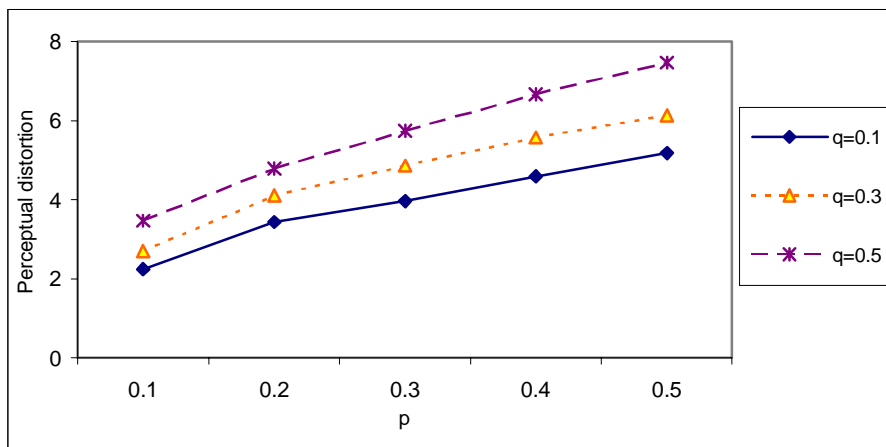


Figure 13 Perceptual distortion (simulation step 1) evaluated by EMBSD.

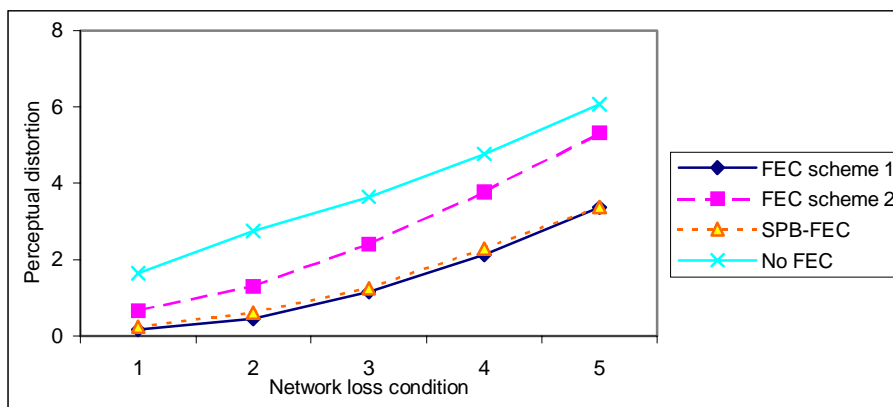


Figure 14 Perceptual distortion of FEC schemes evaluated by EMBSD.

The results of MNB and EMBSD for the first simulation step (Figure 11 and Figure 13) show that with increasing p and q in the network model (and thus increasing packet loss rate and loss correlation), the auditory distance (MNB) and the perceptual distortion (EMBSD) are increasing, i.e. the speech quality of the decoded speech signals is decreasing. These results together with informal subjective testing indicate that the two objective quality measures are reasonably related to the network model parameters and can be used for the speech quality evaluation of the FEC schemes influencing these parameters.

The results of MNB and EMBSD for the second simulation step (Figure 12 and Figure 14) show the quality of the decoded speech signals for the different FEC schemes. We can see that the decoded speech signal without FEC has the highest auditory distance and the highest perceptual distortion and thus the worst speech quality. This is very obvious because the scheme without FEC transmits no redundant data and has the highest application loss rate. However, the auditory distance and the perceptual distortion of our SPB-FEC is significantly lower than those of the reference FEC scheme 2 even though SPB-FEC has a higher application loss rate. The auditory distance and the perceptual distortion of our SPB-FEC come even very close to those of the reference FEC scheme 1 although the application loss rate of this scheme is much lower. These results validate the strategy of our SPB-FEC scheme that does not distribute the amount of redundant data equally on all packets but rather protects a subset of frames that are essential for the speech quality.

5. CONCLUSIONS

We have investigated the impact of frame loss at different positions within a speech signal on the quality and gained the knowledge that the loss of voiced frames at the beginning of a voiced signal segment leads to a significant degradation in speech quality while the loss of other frames are concealed rather well by the G.729 decoder's concealment algorithm. We have then exploited this knowledge to develop a speech property-based FEC scheme (SPB-FEC) that protects the voiced frames that are essential to the speech quality while relying on the decoder's concealment in case other frames are lost. Simulations using a simple but well understood network model and subsequent evaluation using objective quality measures show that our FEC scheme performs almost as good as other FEC schemes at a significant lower redundancy overhead.

Although we only investigated the inter-operation of the G.729 codec and our speech property-based FEC scheme, we believe that a similar gain in speech quality can be expected when our scheme is applied to support other frame-based codecs (e.g., the G.723.1 codec) that operate in a similar way (in particular, G.723.1 incorporates an algorithm similar to that of G.729 to conceal frame loss using the codewords of the previous frames).

Despite its promising results, SPB-FEC faces the general problem of FEC schemes: transmitting redundant data also adds more load to the network and thus worsens congestion in the Internet. Thus, a *network*-adaptive SPB-FEC scheme like the one presented in [3] is highly desirable. In such a scheme, the sender receives feedback information on the network loss conditions from the receivers and uses this information to determine the optimal amount of redundant data.

Besides, SPB-FEC, as any other FEC scheme, only reduces but cannot eliminate the possibility of losing important frames. Moreover, if over a time interval no packets are lost on the transmission path, all redundant data transmitted during that interval wastes network resources. An alternative solution could be to develop mechanisms that allow an application to make the relative importance of packets of its stream known to the network. Then, e.g. by using simple queue management, such packets can receive a lower drop priority than others ([21]). Such drop preference on a *per-packet* basis is supported e.g. by the Differentiated Services (DiffServ) architecture ([1], [2]).

6. ACKNOWLEDGEMENTS

We are thankful to the Speech Processing Lab of the Electrical and Computer Engineering Department at Temple University, especially Dr. Wonho Yang and Prof. Robert Yantorno, for providing us with the Enhanced Modified Bark Spectral Distortion (EMBSD) software to evaluate the efficiency of the SPB-FEC scheme and compare it with other FEC approaches.

7. REFERENCES

1. S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An Architecture for Differentiated Service. Request for Comments RFC 2475, IETF, December 1998.
2. Y. Bernet, J. Binder, S. Blake, M. Carlson, B. E. Carpenter, S. Keshav, E. Davies, B. Ohlman, D. Verma, Z. Wang, and W. Weiss. A Framework for Differentiated Service. Internet Draft, draft-ietf-diffserv-framework-02.txt, February 1999.
3. J. C. Bolot, S. Fosse-Parisis, and D. Towsley. Adaptive FEC-Based Error Control for Interactive Audio in the Internet. Proceedings IEEE Infocom 1999, New York, NY, March 1999.
4. J. C. Bolot. Characterizing End-to-End Packet Delay and Loss in the Internet. Journal of High-Speed Networks, vol. 2, no. 3, pp. 305-323.
5. J. C. Bolot and A. Vega-Garcia. The Case for FEC-Based Error Control for Packet Audio in the Internet. ACM Multimedia Systems, 1997.
6. J. R. Deller, J. G. Proakis, and J. H. L. Hansen. Discrete-Time Processing of Speech Signals. Maxwell Publishing Company, 1993.
7. N. Erdöl, C. Castelluccia, and A. Zilouchian. Recovery of Missing Speech Packets Using the Short-Time Energy and Zero-Crossing Measurements. IEEE Transactions on Speech and Audio Processing, vol. 1, no. 3, July 1993.
8. V. Hardman, M. A. Sasse, M. Handley, and A. Watson. Reliable Audio for Use over the Internet. Proceedings INET 95.
9. International Telecommunications Union. Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s. ITU-T Recommendation G.723.1, March 1996.
10. International Telecommunications Union. Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP). ITU-T Recommendation G.729, March 1996.
11. International Telecommunications Union. Objective Quality Measurement of Telephone-Band (300-3400 Hz) Speech Coders. ITU-T Recommendation P.861, February 1998.
12. D. Minoli and E. Minoli. Delivering Voice over IP Networks. John Wiley & Sons, Inc., 1998.
13. R. J. Novorita. Improved Mean Opinion Score Objective Prediction of Voice Coded Speech Signals. Master's thesis, Department of Electrical Engineering and Computer Science, University of Illinois at Chicago, 1996.
14. C. Perkins, O. Hodson, and V. Hardman. A Survey of Packet Loss Recovery Techniques for Streaming Audio. IEEE Network September/October 1998.
15. C. Perkins, I. Kouvelas, O. Hodson, V. Hardman, M. Handley, J. C. Bolot, A. Vega-Garcia, and S. Fosse-Parisis. RTP Payload for Redundant Audio Data. Request for Comments RFC 2198, IETF, September 1997.
16. M. Podolsky, C. Romer, and S. McCanne. Simulation of FEC-Based Error Control for Packet Audio on the Internet. Proceedings IEEE Infocom 1998.
17. J. Rosenberg. G. 729 Error Recovery for Internet Telephony. Project Report, Columbia University, May 1997.
18. J. Rosenberg and H. Schulzrinne. An RTP Payload Format for Generic Forward Error Correction. Internet Draft, draft-ietf-avt-fec-08.txt, August 1999.
19. H. Sanneck, A. Stenger, K. Ben Younes, and B. Girod. A New Technique for Audio Packet Loss Concealment. Proceedings IEEE Global Internet 1996, London, England, 1996.
20. H. Sanneck. Concealment of Lost Speech Packets Using Adaptive Packetization. Proceedings IEEE Multimedia Systems, Austin, TX, June 1998.
21. H. Sanneck and G. Carle. A Queue Management Algorithm for Intra-Flow Service Differentiation in the "Best Effort" Internet. Proceedings ICCCN '99, Natick, MA, October 1999.
22. S. Wang, A. Sekey, and A. Gersho. An Objective Measure for Predicting Subjective Quality of Speech Coders. IEEE Journal of Selected Areas in Communications, vol. SAC-10, 1992.
23. W. Yang, M. Benbouchta, and R. Yantorno. Performance of the Modified Bark Spectral Distortion as an Objective Speech Quality Measure. Proceedings ICASSP, vol. 1, Seattle 1998.
24. W. Yang, K. R. Krishnamachari, and R. Yantorno. Improvement of the MBSD Objective Speech Quality Measure Using TDMA Data, submitted to IEEE Speech Coding Workshop, 1999.

APPENDIX

Figure 15 demonstrates the impact of frame loss at different position on the decoded speech signal (in this case a male voice is used) in the time domain. We can clearly see that a frame loss at the beginning of the voiced signal causes a significant distortion of the decoded speech signal while the loss of other voiced and unvoiced frames are concealed rather well by the G.729 decoder. Using several different male and female voices, we also obtained similar results.

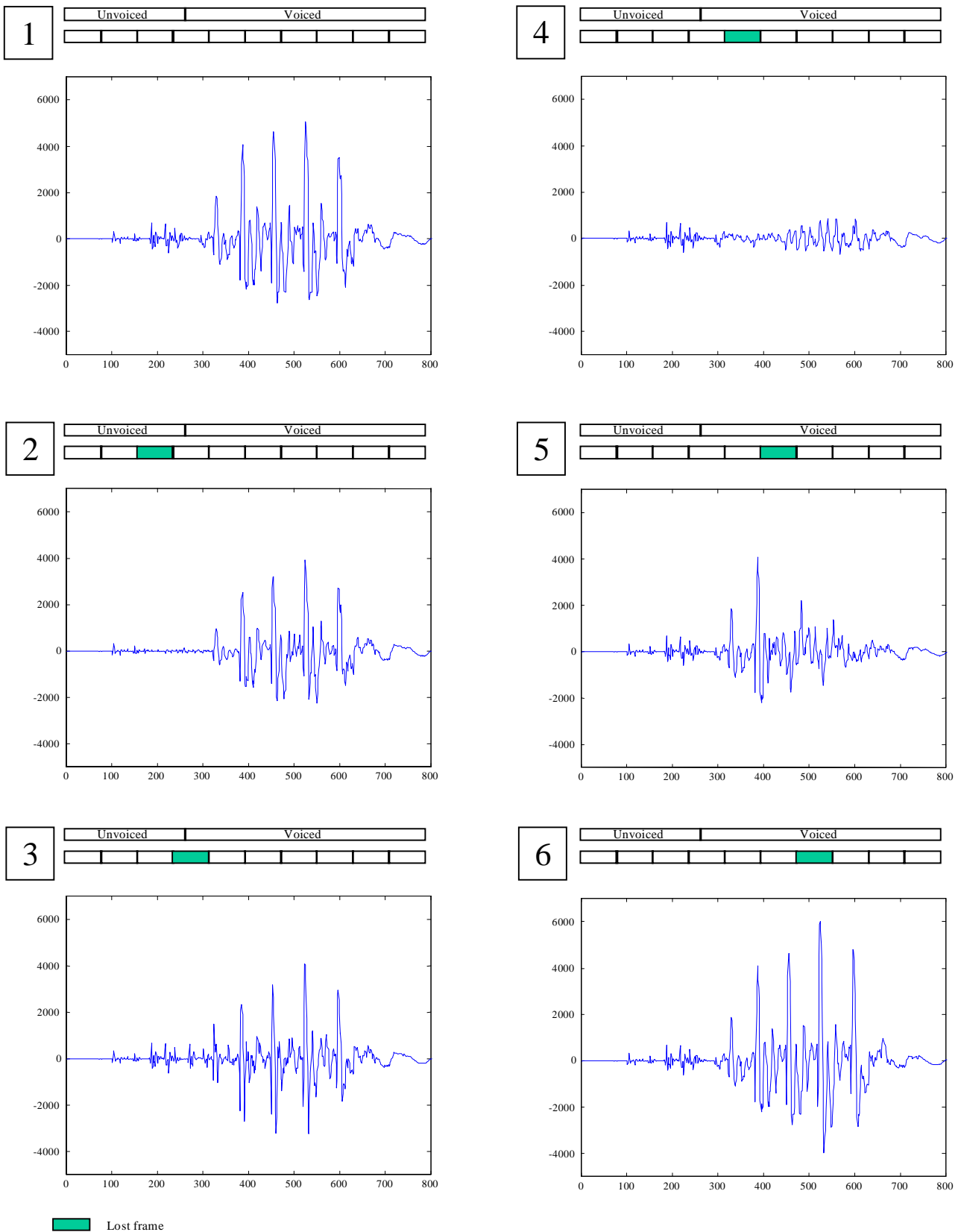


Figure 15 Decoded speech signal without and with frame loss at different position.