LULEÅ
TEKNISKA
UNIVERSITET

1999:309

MASTER'S THESIS

Media-specific Forward Error Correction in a CELP Speech Codec Applied to IP Networks

Magnus Westerlund

Civilingenjörsprogrammet

Institutionen för Systemteknik
Avdelningen för Signalbehandling

1999:309 • ISSN: 1402-1617 • ISRN: LTU-EX--99/309--SE

Abstract

Voice over IP has a problem with the high packet loss present on the Internet, reducing speech quality. For interactive voice applications forward error correction has been suggested as a solution not demanding improved quality of service. The goal was a speech decoder combining GSM enhanced full rate encoded data with redundant vocoder data on the parameter side, with improved speech quality when subject to packet loss. This decoder was designed and implemented and then the quality was measured with SNR, perceptual speech quality measure, and comparative mean opinion score listening tests. The speech quality was primarily improved for single packet losses but also for longer bursts. Distortions which there was no time to correct prevents showing the full potential of the concept. The speech quality improvement potential of this scheme is significant and worth exploiting, despite the increased delay.

Preface

This master’s thesis is the final part of my studies of Computer Science / Signal Processing at Luleå university of technology (LTU). This thesis work was performed at the department of voice processing and radio network research at Ericsson Erisoft AB in Luleå, during the autumn 1999.

Acknowledgments

I would like to thank my supervisor at Ericsson Erisoft AB, Anders Nohlgren for all his guidance. He has always had time for me, answering questions and discussing methods for this thesis. I would also like to thank the other persons working with voice processing research at Ericsson Erisoft AB for kindly enduring all my questions and some listing tests. My examiner, Johan Carlson receives my gratitude for solving some obstacles surrounding this master’s thesis.

Magnus Westerlund
Luleå, December 1999

Contents

1 Introduction

This master’s thesis investigates the use of forward error correction (FEC) for voice over IP (VoIP). On Internet Protocol (IP) networks the error model is different from the wireless radio model normally used in mobile communications. The wireless channel suffers from a high grade of bit errors, while IP has low bit error rate, but packet loss instead. Forward error correction enables an receiver to repair certain losses, depending on extra information that the sender included, i.e redundancy. The redundancy in this work is created using two different speech codecs. The two bitstreams can be used for error correction by delaying one of them one or more frames. This method can handle rather high rates of error but at the cost of overhead and increased delay.

The master’s thesis focuses on designing and implementing an algorithm for combining the Global System for Mobile communications (GSM) enhanced full rate (EFR) speech codec and a LPC-10 like vocoder. The investigation of the combined speech decoder with FEC evaluates if it is sufficiently better than the normal error concealment used in GSM-EFR made for a wireless error model and, if so under which circumstances. How well parameters from a vocoder can be used to do repair in an algebraic code excited linear predictor (ACELP) speech codec is also an issue. The comparison is done using both objective and subjective methods e.g perceptual speech quality measure (PSQM) and listening tests.

1.1 Background

In future communications network for both data and speech, speech coding will be an important part. Speech coding is data compression that is optimized for compressing speech. Speech is often sampled in 8 kHz and then divided into 20-40 ms frames. The speech coder then reduces this to a lower bit rate. Speech codecs of today have after a long evolution become optimized for either circuit connected networks or radio channels. One recent example is Adaptive Multi Rate (AMR) codec developed by Ericsson and accepted as a part of the GSM standard.

The interest to use networks designed for only data transmission also for speech, has grown in recent years. These networks mainly use IP as common protocol have other channel characteristics with new demands and possibilities for efficient speech coding. These new possibilities and challenges create a need for further research in the area of speech coding for IP.

Wired networks usually have large bandwidth, which makes it possible to use FEC. FEC can be used to transport data from a redundant speech encoder, where data have been delayed in time. When a packet is lost the redundant data in the next packet is used instead. This may increase the perceived quality of the communication because single packet losses are more common than double losses. The disadvantages are the overhead information that has to be transported in the network, as well as the increased delay. A framework for FEC together with Real-time Transport Protocol (RTP) is described in RFC 2198 [1]. This framework will probably be part of the coming standard for real-time transport on the Internet.

1.2 Goal

The objective is to design and implement an algorithm for forward error correction with GSM-EFR as primary speech codec and a LPC-10 like vocoder for the redundant data. This design should be optimized with respect to speech quality. The implementation need to be compatible with the format given in RFC 2198 [1]. The algorithm must then be evaluated both objectively and subjectively against frame and bit based error concealment for the GSM-EFR.

1.3 Contents

This report has the following structure: Chapter 2 is an introduction to speech coding and error concealment for speech and audio transport on the Internet. The speech coding part presents the synthesis models used by GSM-EFR and the vocoder. A number of methods for error concealment is introduced and the one used in GSM-EFR are described in more detail. The RTP protocol is presented and the payload format for redundant audio transport is described and its properties discussed.

Chapter 3 describes the design of the used LPC-10 like vocoder, both the encoder and the decoder. The designed error model and its properties and effect on the redundant decoder is explained. The measures taken in different states of the error model are also described. This is followed by comments on the implementation of the design.

Chapter 4 consider how quality of speech is measured with both objective and subjective methods. The measures considered are SNR-SEG and PSQM and a number of different types of listening tests. Error models for the Internet are also shortly considered and the one used in this thesis is presented.

Chapter 5 describes the simulations and tests done. Their purpose and the used disposition is presented. The results are then presented and the implications of them are considered.

Chapter 6 presents the conclusions of this master’s thesis and some suggestions for further studies.

2 Introduction to speech coding and voice over IP

2.1 General

Packet switched networks using IP, e.g. the Internet, has only one service level and that is best effort. This causes many problems for real time network applications like IP-telephony and video conference systems. These applications have requirements, e.g. bounded delay, guaranteed bandwidth, ordered delivery, low jitter, and low or no packet loss.

- Bounded delay: In voice communications high delays disturb the rhythm of a conversation. Therefore the International telecommunication union's sector for telecommunication standardization (ITU-T) recommends a limit of 150 ms end-to-end delay for applications where very little disturbance is acceptable and 400 ms where some disturbance is acceptable [2].
- Guaranteed bandwidth is something many real time applications requires. IP telephones need to send at a constant rate during speech, which requires the network to have this bandwidth available. If not, the network will experience congestion.
- Ordered delivery: If the packets arrive in the same order they were sent, no reordering of packets exist. On the Internet it sometimes happen that packets get reordered by the network.
- Low jitter: Jitter is the variation of the delay between packets. Well provisioned networks with no congestion will have low jitter. But when the data have to go down a certain network link, queues will grow, causing bigger delays.
- Low packet loss: The most common reason for packet loss in the Internet is congestion in network routers. The buffers in the router overflow and the router throws away packets.

There is no single solution to the requirements listed above. To avoid jitter, the network itself has to be well provisioned and this is linked to bounded delay and guaranteed bandwidth. This is a quality of service (QoS) issue and is not easily solved. Today a Resource reSerVation Protocol (RSVP) [3] exist that offers reservations of resource which makes it possible to meet some delay and bandwidth requirements from applications. Due to scalability problems of RSVP there is ongoing work on a new protocol for QoS; differentiated services [4] that scale better to large networks. It also allows several different service levels.

Ordered delivery is easily solved with sequence numbers on the data and by using a buffer where you reorder packets that arrive out of order. Buffers are also used to resolve the jitter problem. By using a buffer and delaying the playback, the application can manage jitter up to the given delay. This method introduces extra delay which makes it more suitable to distribution systems like lecture broadcasting than for interactive real time systems. Packets with too large jitter can also be seen as lost, because packets arriving after the playback point are (almost) useless to a real-time application.

2.2 Speech coding

2.2.1 Speech properties

A speech encoder and its decoder (codec) use the properties of speech to accomplish as good quality as possible for a given bit rate. According to Spanias [5], speech is non-stationary but may for short segments, 5-20 ms be considered quasi-stationary. Speech may be classified in three categories, voiced e.g. “car” or “in”, unvoiced e.g. “she”, or mixed. Voiced speech is also quasi-periodic in the time-domain and its spectral properties are structured harmonically. Unvoiced speech is random and flat in spectrum with less energy than voiced speech.

Speech is created when air from the lungs passes the vocal cords and through the vocal tract [5]. The spectrum envelope for voiced sounds depends on the vocal tract and has a decreasing property with some higher peaks, the formants. The formants are the resonant modes of the vocal tract and are important to the speech synthesis and perception. The voiced sound spectrum also has a fine structure that depends on the vocal cords. When the air bursts pass the vocal cords their vibrations create a periodic signal, seen in Figure 1. The frequency of these periodic vibrations is called the fundamental frequency, or pitch.

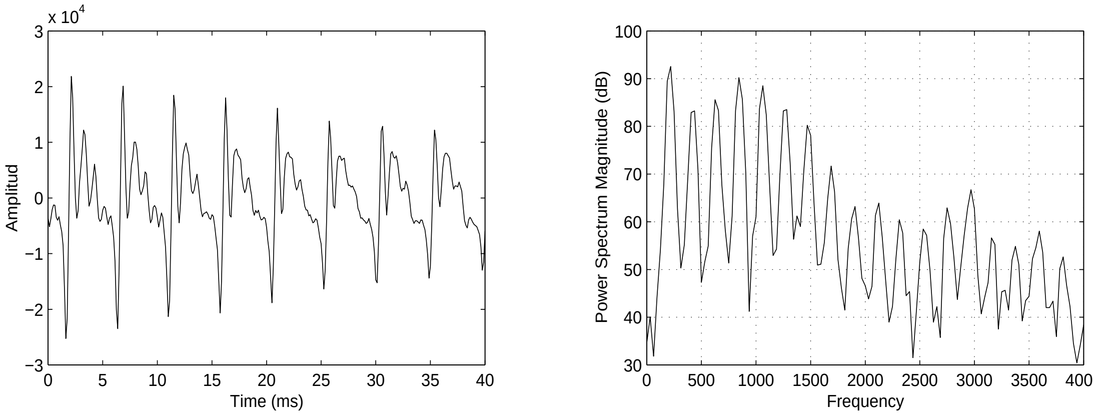

FIGURE 1. Voiced speech segment in both time and frequency domain

Unvoiced sound are formed by a constriction of the vocal tract where air is forced through. This creates a sound which is random-like and has a flat spectrum, see Figure 2. There are a couple of other speech sounds, plosives like “p” which are created from the release of air with a pressure built up by a closure in the vocal tract. Nasal sounds like “n” which use the nasal tract primarily to create the sound.

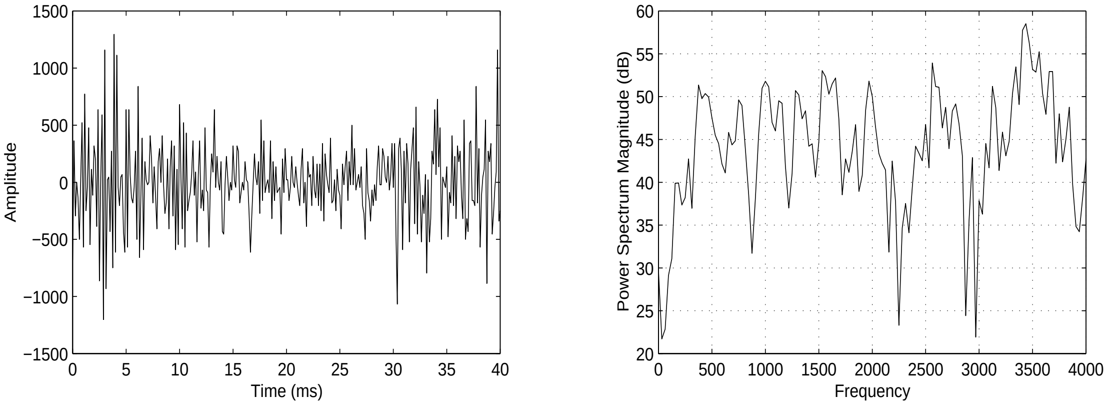

FIGURE 2. Unvoiced speech segment in both time and frequency domain

2.2.2 LPC-10

Linear predictor coding (LPC) with 10 coefficients (LPC-10). Is a USA federal standard voice coder (vocoder) used for military low bitrate digital communication. The model of the speech production system used in the LPC-10 vocoder is the all-pole two state linear source system [5]. This model starts with an excitation that is of two kinds, voiced or unvoiced. Then a linear prediction filter is applied to represent the vocal tract (Figure 3). The voiced excitation is basically an impulse train with a period equal to the pitch convolved with a basic pulse. While the unvoiced is a noise vector. In both cases, amplitude is controlled by the gain factor.

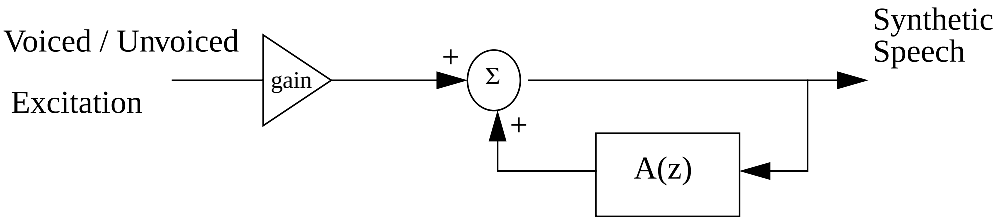

FIGURE 3. Vocoder model for speech synthesis

The linear predictor filter A(z),

$$A(z) = \sum_{k=1}^{n} a_k z^{-k} \qquad \text{(eq. 2-1)}$$

where a_k in eq. 2-1 is determined by minimizing the mean square of the prediction error. The minimizing yields a Toeplitz set of equations that can easily be solved because they are symmetric and Toeplitz. Levinson and Durbin developed a method that allow these equations to be efficiently solved. This method is not used in the LPC-10 encoder. Instead the covariance method is used to solve the Cholesky inversion to determine ten reflection coefficients [6]. These Linear Predictor (LP) parameters are then quantized using a variable number of bits, depending on the order of the parameter [6]. For unvoiced sounds only the first four LP coefficients are encoded with 20 bits, but for voiced speech all ten are encoded with 41 bits.

The pitch is determined from a low-pass and inverse filtered signal. An average magnitude difference function (AMDF) is computed from the signal and sixty pitch values are computed in the range 50 to 400 Hz (20 - 156 samples). The AMDF reduces the number of computations compared with a complete correlation calculation. To decide if the speech is voiced or not, a low-pass filtered version of the sampled input signal is used. A decision is taken for each half of the frame. The final decision, if the frame should be voiced or not, is based upon this and the next two frames. The decision is based on the energy of the signal, the maximum to minimum ratio, and the number of zero crossings of the AMDF. The gain is determined from the root mean square value (RMS) of the signal and quantized with five bits. The pitch is encoded with 7 bits together with the voicing decision.

2.2.3 GSM Enhanced Full Rate

GSM-EFR is a hybrid encoder, because it uses both the model based system in encoding formants and pitch, and the waveform model for matching with the input signal [5]. This is done with Analysis by Synthesis Predictive encoding [7]. GSM-EFR has a speech quality that is equal to or better than 32 kbit/s Adaptive Delta Pulse Code Modulation (ADPCM) according to Järvinen & Vaino [8].

GSM-EFR is standardized by the European Telecommunications Standards Institute (ETSI) in [9]. The excitation in a code excited linear prediction (CELP) codec is selected from a codebook to minimize the perceptual weighted error. The weighting is done so that the quantization noise is placed in the high energy formants. This masks the quantization noise and permits the use of fewer bits in encoding.

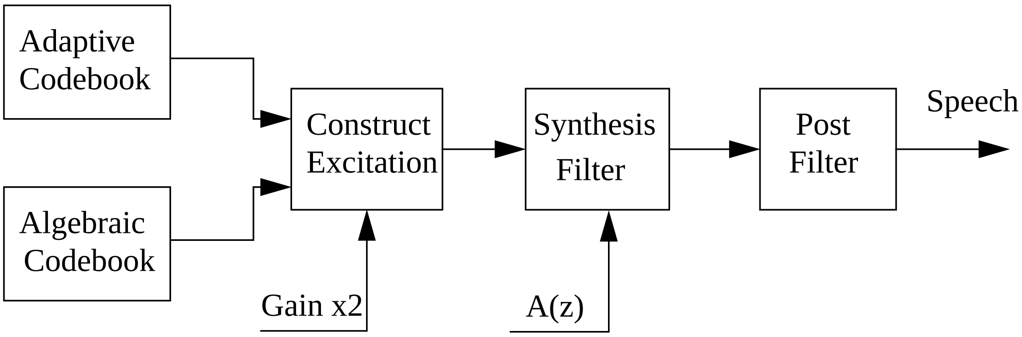

FIGURE 4. GSM-EFR speech decoder model

GSM-EFR encoder uses 20 ms frames which are divided into four subframes of 5 ms each [9]. Initially the input signal is high-pass filtered with the cutoff frequency of 80 Hz and scaled. The LPC is done for two different asymmetrical weighted windows of 240 (30 ms) samples with no lookahead. Lookahead is the use of data from a future frame than the one encoded. The LP coefficients are calculated with a Levinson-Durbin algorithm. Each LP coefficient is then converted to a Linear Spectral Pair (LSP) for quantization and interpolation. LSP maps the filter coefficient on to the unit circle in the range -π to π. Tηε LP coefficients in LSP domain also remain stable in case of bit errors and therefore interpolation is possible. The coefficients are then converted back to LP filter coefficients to be used in synthesis and weighting filters.

Open-loop pitch analysis is also executed for each half of the frame to get an estimate of the pitch. Closed-loop search is then based on the estimate and performed in every subframe. A target signal and the impulse response of the weighted synthesis filter are used to find the best pitch for the adaptive codebook. The target signal is the weighted speech signal minus the zero response from the weighted synthesis filter. The pitch resolution is 1/6 sample in the range 18 to 94 and one sample in the range 95 to 143. In subframe one and three nine bits are used and in subframe two and four they are coded relative to the previous subframe using six bits. The adaptive codebook gain is also computed and non-uniformly quantized with four bits for each subframe.

The algebraic codebook is an interleaved single-pulse permutation design and encodes 40 positions with ten pulses with the values -1 or +1. The 40 positions are divided into five different tracks with two pulses on each track. Each track uses seven bits to encode signs and position. The excitation vector is found by minimizing the mean square error between the weighted input speech and the synthesized speech. The gain of the algebraic codebook is computed and the correction factor for a Moving Average (MA)-predictor is quantized with a five bit codebook.

Decoding is done in the following way: The adaptive codebook parameters pitch and gain is decoded. An excitation vector is then created from the excitation history using interpolation for short and fractional lags. This excitation is also scaled by the gain factor. The algebraic codebook part is recreated and scaled by the gain. These two excitations are added before filtering with LP-coefficients A(z) in the synthesis filter as seen in Figure 4. The synthetic speech is then postfiltered with an adaptive filter consisting of a formant part and a tilt compensation part.

2.2.4 Error Concealment for Speech

Error correction can be divided into sender-based repairs and receiver based concealment [10]. Multiple methods exist for both types.

Receiver based error concealment:

- Silence or noise substitution of synthesis signal. Noise can under short times < 20 ms make the brain mask the effect of loss. Silence is used to prevent changing the rhythm of the speech.
- Repetition of last received frame, works for shorter frames when change is not expected. Primarily usable with waveform codecs e.g. Pulse Code Modulation (PCM).
- Interpolation based repair, interpolates audio before and/or after the loss. Works better then substitution because it can react to some changes. More complex than substitution.
- Regeneration based repair, uses codec parameters and state or model for regeneration of speech. This kind is used in the GSM-EFR error concealment unit where parameters are predicted or averages are used.

All receiver-based methods are incapable of handling longer losses and therefore large frames result in worse errors.

Sender based methods are:

- Forward error correction send extra repair data to cover for possible losses. Two different methods are used, the media independent methods e.g. Reed-Solomon and convolution codes. The other method is media-specific, like the case in this thesis with a primary GSM coder and secondary LPC coder. A method like Reed-Solomon result in bigger delays when whole blocks of packets must be received before repairs are possible.
- Congestion control can be used to avoid or minimize losses due to network congestion. This congestion control is most often done by controlling the bitrate from the application. This is very hard to manage in real-time speech, except through use of multirate coding. Also at lower bitrates lower quality must be excepted.
- Interleaving of smaller speech frames into larger packets. Audio streams become more robust and noise substitution, repetition or interpolation works better. Causes increased delay.
- Retransmission, almost never possible due to the delay bound in real time traffic.

2.2.5 Error Concealment in GSM-EFR

The GSM-EFR codec also has algorithms for error concealment which are proposed in [11]. The channel decoder detects errors through an 8 bit cyclic redundancy check (CRC) on the 65 most important bits [12]. When bit errors are detected the decoder receives a flag which makes the decoder change state in the error concealment state machine. This state machine has seven states representing the amount of errors received. For each successive error the state numbers increase. When a good frame is received the machine returns to state zero in all case except when it was in state 6 representing the worst case. Then it just returns to state 5.

When errors are detected, the gain levels for both the algebraic and the adaptive codebooks are calculated from the median and then attenuated depending on the state. The LSP values used are from the previous frame but shifted towards mean values. The pitch values are taken from the previous good frame’s integer lag. The algebraic codebook is decoded from the received data, ignoring possible bit errors.

The error concealment also contains a flag set in the frame after frame loss. This flag is used when decoding the algebraic and adaptive codebook gains. They are limited and can not become larger then the previous gain value.

2.3 Audio Transport

2.3.1 Real time Transport Protocol

Some of the network problems real-time applications experience can be solved by the Real-time Transport Protocol (RTP) [13]. This protocol consists of two parts. One part carrying data with real-time properties. Part two is the Real-Time Control Protocol (RTCP) that conveys information between users and monitors the quality of service.

The protocol is designed to suit many different applications and therefore has a number of payload formats. Some are static and other dynamically bounded. The protocol offers the following services:

- Sequence numbers to resolve order of packets
- Timestamp for data, e.g. to know when data was sampled. Can be used for audio video synchronization
- Identity of the source that created the payload.
- Identity of contributors to the payload, for such senders that mix a number of sources.
- Payload type information.
- Extension headers

The payload data can be of a number of different types. Some standardized audio and video carrying formats have received static numbers [14]. Payloads can also be dynamically negotiated to a range of numbers. The payload type for redundant audio data is dynamically mapped.

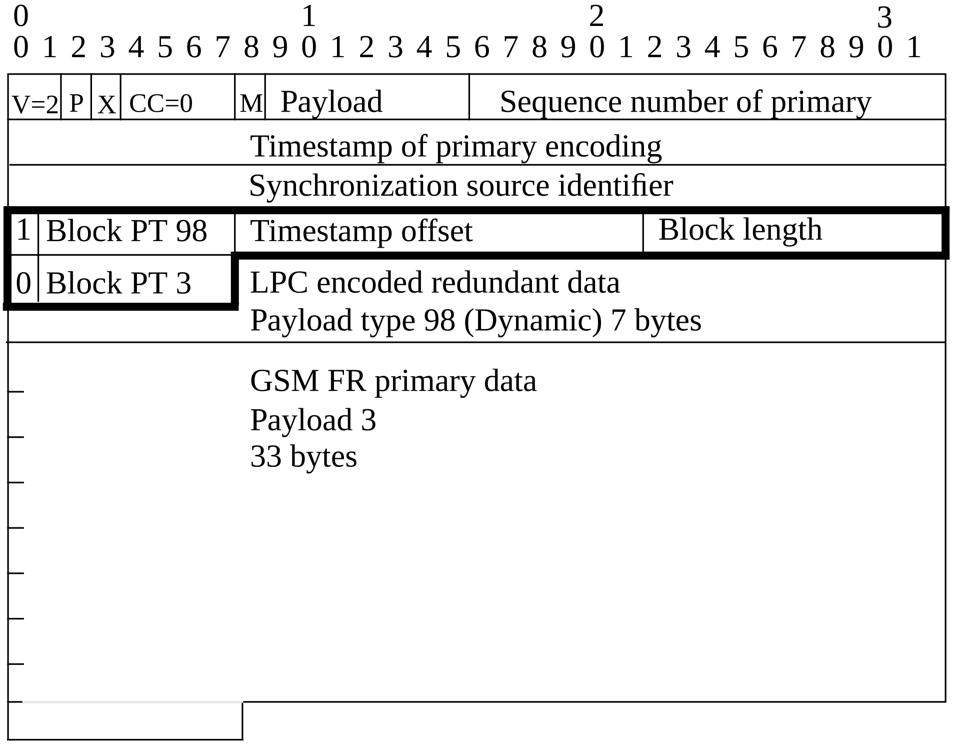

FIGURE 5. Example of RTP header for redundant audio data with GSM Full Rate and LPC encoded redundant payload

The format proposed in RFC 2198 [1] makes RTP able to carry two or more audio payloads in the same packet with less overhead than an RTP extension header would require. After the RTP header there is a four byte (32 bit) header for the redundant data, the header contains, a flag, payload type, timestamp offset, and block length (see Figure 5). The flag specify if there is a another header for redundant data, allowing an unspecified number of redundant payloads. After the redundant encoding’s header, one byte is used to define the payload type for the primary encoding. Both payload types are derived from the same set of RTP payload types and can also be dynamically mapped. In Figure 5 the redundant audio data are mapped to the dynamically assigned payload type 98. The timestamp offset gives the redundant payload a timestamp relative the RTP header timestamp and has the same unit. This forces the data to the redundant encoding to be sampled in the same frequency as the primary. The timestamp offset is also unsigned which prevents applications to send the redundant encoding before the primary encoding. The block length field is used to control where one block of data ends and the next begins.

The redundant encoding can be of the same size or less than the primary, but should be considerably less in order to minimize overhead. Use of multiple redundancy will seldom be needed and if used, each new layer should be considerably smaller than the previous one. This is important since extensive use of redundancy will result in higher network load and worsen the problems with packet loss. The use of the payload for redundant audio data creates an overhead of four bytes (header) plus size of payload data for each redundant payload, plus an extra byte for the primary payload type.

2.3.2 Effects of FEC

When transporting audio on IP networks the properties of the transmission should depend on the given network. This is not easily solved in multicast applications due to scalability problems. But in unicast this will not be a big problem. The RTP control protocol conveys the quality of the transmission and the application can modify the used settings [13].

For low-bandwidth links like modem connections the slow serialization of the data is a problem [15]. The header overhead can be quite large for RTP data sent by User Datagram Protocol (UDP) and IP, which has resulted in the use of larger frames around 80 ms in size. The use of large frames puts higher demand on the error concealment because a frame can contain whole phonemes [16]. The delay also increases.

FEC has been showed to improve speech quality in VoIP by Hardman et al. [16] and Podolsky et al. [17]. However, it is important to consider the effect of adding redundancy to VoIP traffic. If this is done without a control mechanism it might result in increased network load and even more congestion. Podolsky et al. [17] have simulated aggregated VoIP traffic and found that as long as there is other traffic that can reduce its bitrate, the voice traffic with redundancy will achieve better quality. Traffic capable of this is for example Transport Control Protocol (TCP) traffic. If too much traffic lacks congestion control, adding redundancy will increase the problem with packet loss.

Bolot and Towsley [18] have made an adaptive error control for FEC that modifies the amount of redundancy sent based on the RTCP reports sent by the receiver. They also take into account the subjective speech quality. They have tested the concept in a conferencing application, with redundancy, over the Internet. The result measured in packet loss after reconstruction is very promising, but no result on the subjective quality achieved is presented.

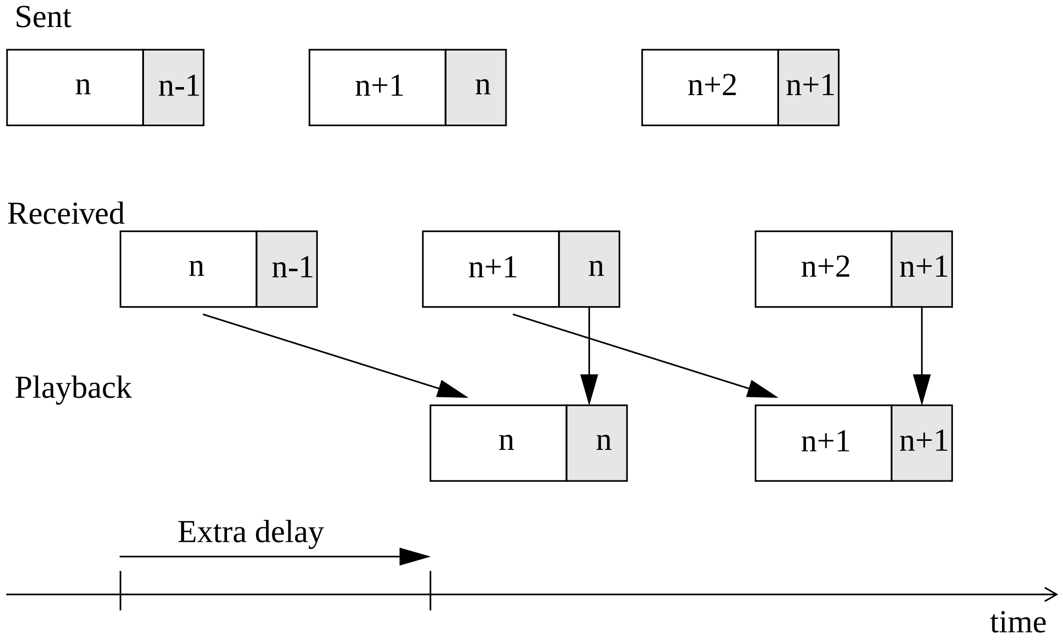

FIGURE 6. Distribution of primary and redundant data and the delay of playback point for one layer of redundant data.

Figure 6 visualizes how data are packed together for one level of redundancy. In the first packet, GSM-EFR data for frame n are packed together with redundant data for frame n-1. It will be possible to decode frame n when the frame containing the n+1 primary and n redundant data arrives. The delay is added on the receiver side as it has to await the next packet.

The effect on receiver buffering when adding redundancy is that the playback point moves further into the future. This will also have the good effect of reducing the number of packets that arrive too late, at the cost of delay. In some applications high levels of error recovery could be accomplished when the playback point only was moved half a frame length [15]. This was investigated using frame sizes of 80, 130, and 206 ms and the best result was achieved with the 80 ms frame.

3 Design and Implementation

The goal is a speech decoder that uses redundant data on the parameter side of the decoder. The design of such a decoder is dependent on which data that are used. The primary speech codec used is the standardized GSM-EFR and as redundant codec a LPC-10 implementation especially made for this thesis. This conforms with the RTP payload for redundant audio data, which permits payload formats to be selected with almost total freedom. Each of the two data streams can be decoded to intelligible speech.

The GSM-EFR speech codec is an Algebraic Code Exited Linear Prediction Coder (ACELP) which codes a 20 ms frame of 160 samples to 244 bits/frame which is equal to an encoded bitstream of 12.2 kbit/s [8]. The LPC-10 vocoder has 22.5 ms frames and 54 bits/frame equal to 2.4 kbit/s [6]. LPC-10 also has low demands on hardware compared to the GSM-EFR, much depending on what years they where designed. GSM-EFR was designed in the mid 1990's and LPC-10 is from the early 1980's.

Due to the odd frame length (22.5 ms) of the LPC-10 vocoder, it is not suitable to use as a redundant encoder with GSM-EFR. Therefore an implementation of an LPC-10 was done using the structure from GSM-EFR. This also has the advantage that the algorithm for handling lost frames will be easier to implement because both coders use the same parameter format.

The combined decoder will use both sets of parameters to decode the speech when subjected to packet loss. By designing a decoder that uses the parameters instead of combing the speech streams from two separate decoders, a significantly better quality than the normal error concealment performed in GSM-EFR will be achieved. The two encoders use different synthesis models which create difficulties when they are combined. However, they share the most important parameters, namely the linear prediction coefficients and the pitch.

A vocoder solution uses voicing decision, pitch and energy to create an excitation. Vocoders use little bandwidth, and work well for either voiced or unvoiced segments of speech. For segments that are neither, for example plosives, they perform much poorer. This solution is also less state dependent which causes problems when combining with a codec like the GSM-EFR. Although this is also positive in the sense that the vocoder can run with less history and is more stable in an environment with losses. The largest problem is the phase in the pitch period that must be detected in the excitation history state, if distortion is to be avoided.

Another type of encoding that has been considered is multipulse coding, where a number of the most important pulses from the residual is encoded. This solution will react better to changes and transitions from unvoiced to voiced. No phase problem will arise when combining it with GSM-EFR. One disadvantage is the high bandwidth demand for each pulse. To achieve better results the number of pulses must be increased. If too few pulses are used some pitch periods with short lag will be impossible to represent. As an example, a four pulse system can only describe pitches down to 40 samples in lag equal to pitch frequencies below 200 Hz.

3.1 Vocoder

The implementation of a vocoder based on the GSM-EFR codec (GSM-VOC) was made for two reasons. First, the frame length for the LPC-10 speech coder is 22.5 ms compared to GSM-EFR's 20 ms, which would have complicated the combination considerably. Secondly, the parameters would not have been as well matched if they came from two different encoders. This also resulted in a vocoder that uses more advanced and resource demanding methods in determining the parameters and also with slightly better speech quality. GSM-VOC combines methods from both the LPC-10 and the GSM-EFR.

3.1.1 Encoder

From the incoming speech that is HP filtered with a cut-off frequency of 80 Hz, the RMS energy value is calculated. The LP coefficients are then calculated and quantified with the method from GSM-EFR. Where GSM-EFR calculates two sets of LP coefficients, only one, derived from the window with more weight on the last data is used in GSM-VOC. After the LP coefficients are found the residual is calculated.

GSM-EFR does one open loop pitch search for each half of the frame. This search is done by calculating the auto-correlation over 80 samples for lags 18 to 143. The calculated correlations are then weighted to favour small lags. This weighting is done by dividing the span of 18 to 143 into three sectors, 18 - 35, 36 - 71 and 72 - 143. The maximum value from each sector is then weighted and the largest one is selected. Then the result from the two halves are compared, and the LTP lag of half frame with the largest correlation is used in GSM-VOC.

The voicing is calculated based on the unweighted maximum correlation from the open loop searches. The correlations from the two previous, current and next two half frames are used in the voicing decision seen in Figure 7. To calculate the correlations for the next frame a 20 ms lookahead is required. The lookahead is available at no cost for the redundant encoder, since in the FEC scheme used, the redundant data represents a frame earlier than the one that is primarily encoded (see Figure 6). The delay can be used to achieve a lookahead by encoding the redundant frame at the same time as the primary.

To determine if the speech is voiced or not, the five correlations are compared to three different thresholds. Firstly, a median calculated from three correlations, the present and the next two half frames, is compared against a threshold. This threshold is used to quickly react to the start of a voiced segment. Secondly, the median of all the five correlations is compared to a second threshold. This threshold is lower than the first one, and is used to detect voicing during a voiced segment. The third comparison, also involves the median of the five correlations, but with hangover. The hangover is a condition; the previous half frame must have been voiced. If that is not the case this threshold is not used. The hangover threshold value is the lowest of the three. The purpose of the third threshold is to hold out voiced segments to or past the true point of transition. The third threshold will make sure that the half frame where the transition from voiced to unvoiced speech occurs, will be marked as voiced. The voicing for both half frames are sent to the decoder.

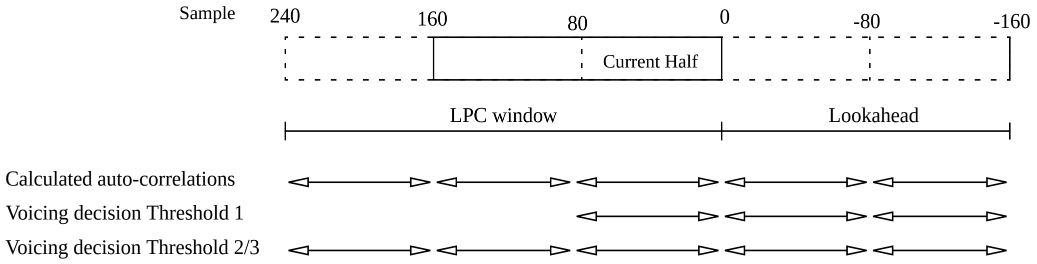
Sample
240
160
80
0
-80
-160
Current Half
LPC window
Lookahead
Calculated auto-correlations
Voicing decision Threshold 1
Voicing decision Threshold 2/3

FIGURE 7. Division of subframes for GSM-VOC and their use in different methods.

The LP coefficients are quantized using a modified method from the speech coder IS-641, which uses prediction of the linear spectral frequencies (LSF). The modification is in the predictor, which uses mean LSF values instead of a prediction factor based on the previous frame's LSF's. This eliminates dependencies on the previous frame for the LPC's. The ten residuals from the prediction are grouped into three vectors. These vectors are then matched against a statistically produced table for the best match and the index in the table is returned. These three indices use 26 bits.

The RMS value is converted into dB and then linearly quantized using seven bits. This is unnecessary many, five or six should be sufficient. The voicing state uses two bits to represent the voicing in each half frame. The pitch has a range of {18..143} samples, 18 is subtracted so the valid numbers fit into seven bits {0..125}.

Parameter	Nr of Bits
LPC	26
Pitch lag	7
RMS value	7
Voicing state	2
Pitch pulse position	8
Pitch pulse sign	1
Total (Bandwidth)	51 (2550 b/s)

TABLE 1. Bit allocation for GSM-VOC

The encoder generates two parameters that are unnecessary when used as a stand-alone vocoder, namely the pitch pulse position and its sign. This parameter tells, with a resolution of one sample, where in a frame the pitch pulse starts to keep the excitation and its synthesis in phase with the original speech, which is important when used for FEC. In a stand-alone vocoder no parameter specifies anything to a certain position and the phase is irrelevant as long as pitch epochs has the given pitch lag distance. This parameter is found by correlating the residual and a fixed pulse form. The position and sign is then located in the correlation curve with help of the voicing decision to point to the correct frame half.

3.1.2 Decoder

First an excitation vector is created from the voicing decision and pitch. The voicing has 6 different states, two steady states, and four transition states. The steady states are, voiced and unvoiced. The transition states are from unvoiced to voiced and from voiced to unvoiced and they occur in either half of the frame. For voiced parts of the frame the given pitch is used to determine the epochs that are calculated. Unvoiced frame's is dived into four epochs of 40 samples each for interpolation purposes.

For each pitch epoch, the value of RMS and pitch are interpolated between the new and old values for softer transitions. Furthermore an excitation is created, for voiced speech a 25 sample long pulse and low intensity noise are used. For unvoiced parts the excitation consists of noise only. In a voiced pitch epoch the pulse is low-pass filtered and the noise high-pass filtered. The created excitation is then filtered with $1 + 0.7\alpha A(z)$, where α is the gain of A(z). This is to reduce the peaked nature of the synthetic speech according to Tremain [6]. For unvoiced frames where the RMS value is increased more than eight times the previous frame's value, a plosive is added. The position of the plosive is random in the first unvoiced pitch epoch and consists of two consecutive pulses, one added and the other subtracted. Then the RMS value of the epoch is adjusted to match the interpolated value. This is done by calculating the present RMS value of a synthesis filtered excitation.

The LPC's are interpolated in the LSF domain for each 40 sample subframe and then applied to the excitation. The pulse used for voiced excitation is biased and to remove this bias a high-pass filter with cut-off frequency of 80 Hz is used.

3.2 GSM-EFR decoder with redundancy

This decoder designated GSM-RED is designed with aim to produce the highest possible quality of speech with both low and high levels of packet loss. It is restricted to one level of redundancy even if more levels could be added as an extension of the current design. The design also works under the assumption that primary and redundant data are sent in the same packet. This assumption is important in the fact that the decoder when receiving a packet always has the previous redundant frame and the current primary frame. The decoder is designed to have a receiver buffer with increased delay to await the next frame's packet with the current frame's redundant data.

Which data that are available and which are not, are presented in Figure 8. The machine starts in state *EFR Norm* which represents primary decoding and having received the packet containing the next frame's data. The transitions in the machine are based on the arrival or not of the packet containing the next frame's data. The transition from *EFR Norm* to *EFR Nxt E* is done if the next frame's packet is lost. The current frame’s primary data arrived in the previous packet, therefore primary decoding can be done.

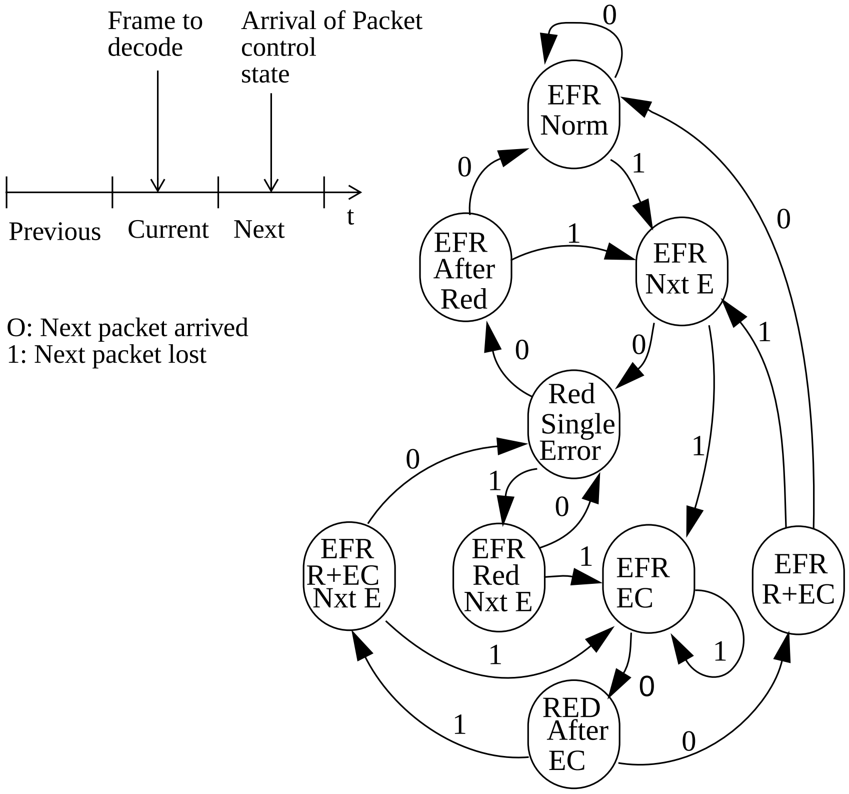

FIGURE 8. GSM-RED Loss state-machine

The states in this machine need further presenting:

- *EFR Norm*: Primary decoding and next packet has arrived.
- *EFR Nxt E*: Primary decoding and next packet is lost.
- *Red Single Error*: The current frame’s primary data are lost, but the next frame’s packet arrived carrying redundant data for the current frame. Decoding is done of the redundant data with the knowledge that a single packet was lost.
- *EFR After Red*: Next frame’s packet has arrived and there are primary data for the current frame although the previous frame was decoded with only redundant data.
- *EFR Red Nxt E*: Next frame’s packet was lost. The current frame’s primary data have been received and the previous frame was decoded with redundancy.
- *EFR EC*: Multiple packets were lost in sequence, resulting in that no data are available for this frame. Error concealment (EC) is applied as it is done in GSM-EFR.
- *Red after EC*: Next frame’s packet has arrived containing the current frame’s redundant data. Decoding of redundant data after one or more frames of EC.
- *EFR R+EC Nxt E*: Next frame’s packet is lost. The current frame is decoded with primary data but the previous frame was decoded with only redundant data which was preceded by EC.
- *EFR R+EC*: Next frame’s packet arrived. Current frame is decoded with primary and redundant data but previous frame was decoded with only redundant data and the frame preceding that frame was created with EC.

To simulate a destination buffering, GSM-RED has a buffer mechanism that sorts the data to its correct time and data slots. This buffer lacks the capability to handle the case when a packet arrives after the redundant data could be decoded but before the primary data is needed. So this is not considered in the design. Data can be fetched from the buffer prior to its normal decoding time.

State EFR Norm: Speech is decoded according to GSM-EFR standard [9].

State EFR Nxt E: Same as *EFR Norm*, the state is only used to represent that next frame’s packet is lost. Because the redundant data for this frame are missing, the RMS value is calculated and entered into history. The voicing of this frame is also calculated by taking the maximum of the autocorrelation and feeding it to the voicing decision module used in the encoder. This design is done without the lookahead which results in a less accurate decision.

State Red Single Error: In this state decoding is done with redundant data for the current frame and primary data from next frame. The LPC’s for subframe four are decoded from the redundant frame. The decoded values are used to update the predictor of the primary LPC decoder. The predictor factor is calculated from the previous frame’s LSF residual and the decoded LSF values in this frame. The difference from the real predictor value is the quantization error that has occurred depending on different encoding of the data. The other subframes LPC values are interpolated between decoded value and the previous frame’s LPC in the LSF domain.

The LTP lag, RMS value, and pitch pulse position and sign, are extracted and decoded into parametric values. The voicing decisions are also extracted from the frame and used to create a voicing state. The voicing state depends on the previous half frame’s decision and the two current half frame values. This state is used to control which actions are taken in constructing the excitation.

The possibility to prefetch primary data are used in the decoding in this state. On LTP gain and algebraic codebook (Alg CB) gain for the current frame EC is applied. Then when predictor and histories has reacted to the current frame next frame’s parameters are decoded. These values are used for predicting the RMS of the next frame. The prediction is done by calculating a mean LTP gain and the energy of the Alg CB vector with gain applied i.e. eq. 3-1.

$$R\hat{M}S = \sqrt{\overline{LTPgain} \cdot prevRMS^2 + RMS(AlgCB \cdot Alggain)^2} \qquad \text{(eq. 3-1)}$$

In frames with voicing state representing steady state voiced the excitation is created in a different way from the other states. The excitation is created in the same way as GSM-EFR normally does it. The LTP vector is created, by copying in the excitation history, with LTP lags that are interpolated between the values from the redundant data and the previous frame. This is done only if the difference is small enough i.e. less then eight, otherwise the new lag is used in all subframes. The check is done to avoid interpolating a gap that are a result of the encoder choosing a LTP lag that are two periods long. The Alg CB is randomized to avoid ringing, and the gain is calculated so the Alg CB vector will have one tenth of the LTP vector.

The excitation is the sum of the LTP vector and the Alg CB vector. The excitation vector’s amplitude is then adjusted with a RMS value for each subframe. Adjusting on subframe basis is not the best option, because the pitch pulses distribution of energy are not even. So if two high energy parts of the pitch pulses are in a subframe they will receive a smaller amplitude than if only one high energy part was in the subframe. The adjustment should be done on pitch pulse basis instead. The RMS value is interpolated in the three first subframes between the RMS value in the last subframe in previous frame and the current frame's RMS value. In the last subframe the value is interpolated between the current frames value and the predicted value of next frame. This results in a softer transition into the next frame.

In frames with other voicing state than steady state voiced, the excitation is created more similar to the GSM-VOC. In steady state unvoiced the excitation is noise which amplitude is adjusted so that the subframes receive the correct RMS. In transitions to unvoiced the position of the last pitch pulse is located. This is done by correlating the previous frame's synthesis with a pulse form. From the correlation maximum, the next local pulse maximum is located with steps of LTP lag size until the last possible maximum is found. The vocoder excitation module is updated to start at the end of the last pulse, i.e. somewhere in the current frame. The missing samples are copied from the positions just before the start of the last pulse. If this position is not beyond the position where the unvoiced segment starts, one or more vocoder pulse will be added, RMS values are interpolated towards the frame's value. From the end of the last voiced pulse, noise is instead generated to the frame boundary. The noise RMS is also interpolated so that a soft transition to unvoiced is achieved.

If the voicing state represents transition to voiced, the pulse position and sign are crucial. The excitation consist of noise until the given pitch pulse position. This noise's RMS is interpolated towards the received value. At the pitch pulse position, the first vocoder pulse is placed, with an interpolated RMS value. All pulses use the received lag. The RMS interpolation is between the value of previous frame's last subframe and the received value in the first half of the frame and between the received value and the predicted value in the second half.

When calculating the RMS value for the excitation, the excitation are actually synthesis filtered with the correct filter state so that the filter gain is taken into account. After the adjustment of the energy the excitation is HP-filtered so that the biased part of the vocoder pulse is removed. To give the LTP something to work with in following frame the created excitation is entered into the excitation history. A synthesis filter is then applied a final time to create the synthesis. The synthesis from a steady state voiced is also postfiltered.

State EFR After Red: In this state the decoding is done in the GSM-EFR way except that already decoded gain parameters are used. The synthesis that is created has its amplitude adjusted so that the RMS value of the whole frame corresponds to the received value from the redundant data. To avoid discontinuities in the synthesis that can produce high frequency noise, the adjustment is done on the excitation. The excitation is then fed into the excitation history for consistency with the next frame. The synthesis filter is reset to the state it had initially in this frame, and then used on the excitation again.

State EFR nxt E: In this state there exists no redundant data to use when correcting the energy level of the synthesis. Instead a prediction is calculated as in eq. 3-1.

State EFR EC: In this state, when no data was received the error concealment used in GSM-EFR is used. This include taking the mean of the gain histories (LTP and Alg CB) and attenuating that value and feeding it back into the history. Because the data are lost instead of distorted by bit errors, the algebraic codebook vector can not be used as received, a new one is randomized. This method is used in GSM-EFR adapted for packet based networks. If the vector was instead copied from the last frame, ringing in the speech might occur. The RMS value and voicing state are calculated from the synthesised speech as in state *EFR nxt E*. The use of the last good frame's pitch can result in a large phase drift of pulse positions in the excitation history.

State Red after EC: The big difference between this state and Red single error is that there was one or more frames with EC before it. Therefore the excitation history is very uncertain and should not be used. The excitation in steady state voiced is created from vocoder pitch pulse and the energy is interpolated from: previous frame, the current value, and the prediction for the next frame. The position and sign of the pulses are taken from the received data so that the phase of the excitation history is as good as possible. The points before the given position is copied from excitation history like in steady state voiced of the Red Single Error state.

State EFR R+EC nxt E: This state is the worst case of the states that have primary data to decode. The LSF predictor is very probably out of line and can not be corrected with the data available. Therefore the GSM-EFR LPC's are decoded normally and then slightly bandwidth expanded. This is done in the same way as the in the GSM-EFR's EC but in lesser amount to avoid creating another type of instability. The energy adjustment of the excitation and synthesis are done against a predicted value, i.e. eq. 3-1. Afterwards the RMS and voicing for the current frame are calculated from the synthesis.

State EFR R+EC: After EC has been applied to the LP coefficients the predictor loses its precision. In this state this can be corrected with the redundant data. The redundant LPC coefficients are decoded, and they represent the same value as the second series of LPC coefficients in GSM-EFR. Both are used to calculate an estimate of the predictor value for the current frame.

$$LSF = LSFres + meanLSF + predFactor \cdot prevLSFres \qquad \text{(eq. 3-2)}$$

$$prevLSFres = \frac{(redLSF - meanLSF - LSFres)}{predFactor} \qquad \text{(eq. 3-3)}$$

A LSF is predicted as eq. 3-2 where *LSFres* is the value that is decoded from the data, *meanLSF* and *predFactor* are constants. This makes it possible to calculate eq. 3-3 to produce the previous frame's LSF residual when the LSF for this frame and the LSF residual are available. This estimation gives the advantage that LP coefficients for the current frame have an error equal to the redundant LPC quantization error. The predictor would otherwise have been correct in the next frame when it had been updated with the current frames LSF residuals.

There exist another predictor in the GSM-EFR and that is for the algebraic codebook gain. The values of the codebook gain is rather stochastic and no available redundant parameter matches that. This prevented designing a method that estimate the Alg CB gain. The predictor takes approximately one frame before it has become stable after a frame loss. The predictor could be updated with help of the energy changes seen between frames. The distribution between the LTP gain and algebraic gain could be measured in the encoder and sent with very few bits, two or three. The updating of the predictor should also consider the voicing state. In transition to voiced the algebraic gain is often too large, to build up a history for the LTP to use in later frames. In steady state the gain is more moderate and for unvoiced it produces most of the randomness found in unvoiced.

3.3 Implementation

This design was implemented in C++ by adding and modifying an existing floating point implementation of the GSM-EFR. This implementation was done in Baseline Codec Library (BC-lib) which is an Ericsson research speech coder development environment. In BC-lib a large number of different speech coders are implemented which speeded up my work considerably since I was able to use algorithms from BC-lib.

The concentration of the implementation work has been on the methods specially developed for this thesis. The current implementation has been modified numerous times in trying to reduce the decoding distortion and trying different algorithms. The implementation still suffers from distortion, especially when multiple sequential packet losses occur. This is possible to correct, but not in the time available for this thesis project.

Things that could be improved are:

- The pulse position search in the encoder should be moved so that it uses the voicing decision based on lookahead. The search algorithm should also be better at deciding if a found local maximum in the correlation actually are a pulse or merely noise.
- The RMS measure of the last subframe should be changed to measure the last complete pitch epoch so that only one pitch pulse can be measured. With the current measure over the last subframe, zero, one, or two high energy parts can be present depending on the pulse’s position and the pitch lag.
- Same type of problem as above arises in the energy adjustment that is done on subframe basis in state *Red Single Error* and *steady state voiced*. The energy interpolation should be adjusted based on the amount of pitch pulses.
- When in the error state *Red after EC* the placing of the first pitch pulse should be adjusted. This adjustment should consider both the received pulse position and the phase information in the previous frame’s synthesis. To minimize phase discontinuities the whole frame should be used to correct the phase error. This under the assumption that the previous frame’s synthesis consist of voiced speech.
- Improved interpolation for the pitch pulses. Instead of the linear interpolation used today, interpolation with a polynomial should be used. The polynomial should be matched to the following values: previous frame’s total RMS, RMS for previous frame’s last pulse, current frames RMS and next frame’s predicted RMS.
- The prediction of the energy should be more advanced. There exist enough data to determine the energy envelope for the next frame. From the envelope, the energy and its derivative at the start of the next frame could be predicted. This information

could be used to improve the energy interpolation so a even softer frame boundary could be accomplished. When the energy of the next frame is depending on the excitation, some iterative method must be designed.

- If the above prediction was slightly wrong, the energy level needs adjustment in the next frame. To avoid discontinuities, some kind of uneven adjustment could be used, e.g. the gain adjustment is almost zero in the beginning of a frame and then grow to the needed value by the middle of the frame.

3.4 Parameter usage

To reduce the amount of redundant data (overhead) transmitted over the network, some parameters could be discarded. The prime selector of which parameters that can be removed from the frame is the voicing state. This depends on the characteristics of speech and is also used in LPC-10 to make room for extra channel coding in unvoiced segments [6].

Parameter	Nr of Bits
LPC	26
RMS value	7
Voicing state	2
Total (bit rate)	35 (1750 b/s)

TABLE 2. Parameters in unvoiced speech

In unvoiced segments the parameters in Table 2 are needed. The LPC’s are needed to shape the spectral properties of the noise. RMS to know the energy of the noise. The voicing state could also be removed and instead use the data size as an indicator of unvoiced speech. That would give a frame size of 33 bits and bit rate of 1650 b/s. The spectral shaping of the noise may not need as precise values as voiced segments and because of that you could use a other type of quantization and save some bits. That would, however reduce the effectiveness of updating the predictor for the primary LPC decoder.

Parameter	Nr of Bits
LPC	26
Pitch lag	7
RMS value	7
Voicing state	2
Pitch pulse position	8
Pitch pulse sign	1
Total (bit rate)	51 (2550 b/s)

TABLE 3. Parameters in voiced speech

In transitions from unvoiced to voiced speech all parameters in Table 3 are needed. The LPC parameters normally drastically change, the voiced speech has a pitch and a new level of energy exists in the frame. The pitch pulse and sign are needed to generate a correct phase for the excitation. All these factors are needed because of the transition.

In *steady state voiced* and transitions to *unvoiced* the pitch pulse position and sign could be removed, reducing the total bit amount to 42 bits (2100 b/s). That will, however have the negative effect that the decoder will not receive any phase information in these frames. This will force the decoder to search the phase in the previous frame which can result in larger phase errors since the algorithm can not detect the phase due to loss of a burst of packets. It also makes it impossible to correct any phase drift that has occurred under a period of error concealment.

4 Evaluation methods

Speech quality is measured with both objective and subjective methods. The objective measurements try to put a number to the speech quality. They must have two properties to be useful. Firstly, low and high objective quality must correspond with low and high subjective quality. Secondly, it must be possible to mathematically analyse and implement it in some algorithm [19]. Subjective quality is done with listening tests and grading on different scales.

4.1 Objective methods

A time-domain method is the signal to noise ratio (SNR) which is defined as,

$$SNR(z) = 10 \cdot \log\left\{ \sum_{n=z}^{N+z} x(n)^2 \Big/ \sum_{n=z}^{N+z} [x(n) - y(n)]^2 \right\}, \qquad \text{(eq. 4-1)}$$

where *x(n)* is a reference signal and *y(n)* is a measured signal, *z* is the starting point and *N* the number of samples to process. SNR does not correspond too well with subjective assessments, as an example, in a pause without speech activity, a small amount of noise will result in large negative SNR values. This can be solved with segmental SNR, which is defined as

$$SNR_{seg} = 10 \cdot \log\left(exp\left(\frac{1}{N} \sum_{m=1}^{N} \log\left\{ \sum_{n=1}^{M} x(n)^2 \Big/ \sum_{n=1}^{M} [x(n) - y(n)]^2 \right\} \right) - 1 \right), \qquad \text{(eq. 4-2)}$$

where *N* is the number of *M* sized blocks to calculate SNR-SEG over, x(n) a reference signal and y(n) the measured signal. SNR-SEG works better than eq. 4-1 as a speech quality predictor for 32 kbit/s ADPCM and 64 kbit/s PCM [19]. But SNR-SEG is not a usable measure for distortion of the synthesised speech from the vocoder. When the vocoder encoding does not preserve the timing of speech at sample level, the phase shifts are measured as distortion, resulting in much lower SNR-SEG values than actual speech quality.

Perceptual Speech-Quality Measure (PSQM) is an objective measurement based on a psychoacoustic representation [20]. The frequency spectrum is calculated with FFT over 40 ms of samples. This spectrum is then perceptually weighted and filtered to transform it into the psychoacoustic domain. Perceptual weighting assigns different weights to different parts of the spectrum depending on how sensitive the human ear is to these frequencies. This process is done on both the measured signal and the reference. The measured signal is also scaled to remove gain differences. The difference between the signals are then weighted with a function for how critical different frequency bands are.

PSQM also needs time synchronization between the reference and measured signal. Because of the frequency approach, exact sample synchronization like SNR is not needed but the measured quality will drop if the phase shift is too large. Important is also that the reference and test signal are normalized to the same energy level. The measure is on the scale 0-6.5 where 0 is equal to reference and 6.5 is worst. The measure depends on the sentences that are measured and therefore has some variance.

4.2 Subjective methods

There is a number of different methods for subjective testing. Not only the method can vary between tests, but also the scope of the test. From small laboratory tests up to actual field tests. The method and scope of a test exert a large influence on how the results can be interpreted [21].

4.2.1 Absolute category rating

Absolute category rating (ACR) assessments [21] of speech quality is performed by letting a group of people listen to 6-10 seconds of speech and then rate that sentence. The rating is done with a five point scale {excellent, good, fair, poor, bad} which is mapped to a numeric scale of 5 to 1. The person making the test rates a number of sentences including a number of references. The order of the sentences in the test are important, a fair sample played after a good can receive a different rating then if played after a poor one. Therefore randomization of the order for different groups of listener must be applied. The final result is a mean opinion score (MOS), calculated from the results of the listeners.

The test material can also effect the MOS scale by stretching or compressing it. Therefore reference sentences using the Modulated Noise Reference Unit (MNRU) [22] scale are often added so that this effect can be considered. Even though ACR test uses an absolute scale, results from different tests can not be compared. The factors effecting the test are many and the subjective nature of the test also prevents comparisons. ACR test tend to saturate in both ends of the quality scale, since subjective rating is harder to do when differences are small.

4.2.2 Degradation category rating

Degradation category rating (DCR) is performed by giving a degradation rating compared to a high quality original [21], this results in a relative assessment. The test subject first hears the high quality original and than the test signal, and rates the degradation. The degradation is graded in a five point scale {inaudible; audible, but not annoying; slightly annoying; annoying; very annoying}. The method makes DCR more sensitive and works better with high quality. The disadvantage is that only material measured in the same test is comparable. The result is presented as a degradation mean opinion score (DMOS).

4.2.3 Comparative rating

Comparative MOS (CMOS) is a listening test where the test subject hears two samples. One is a reference to compare against, the other is processed by the methods to be tested. They are played in random order and then rated with the question: Is the first sample compared with the second. 1: Definitely better, 2: slightly better, 3: equal, 4: slightly worse, 5: definitely worse. This gives a relative measure of how much better or worse the tested process is compared to the reference [23].

4.2.4 Speech intelligibility

It is important to measure how intelligible the decoded speech is, especially for low quality codecs (below telephone standards). This can be measured by articulation tests measuring the percentage of speech sounds that are correctly perceived [21]. One articulation test is the Diagnostic Rhyme Test (DRT) where the listeners hear one-syllable words that are one of a rhyming pair that are predefined e.g. meat and beat. Then the listeners judge which word in the pair it is. The intelligibility score normally is in the range 70-92 percent, and a high quality telephone system scores around 92-96%.

4.3 Error Model for Voice over IP channels

It is practically impossible to create a model for traffic on the Internet [24]. This depends on the basics in IP networks. IP hides the detail of the underlaying link technology. Because of the enormous amount of links with a wide variety of properties, the Internet is very heterogeneous. Another problem with modelling and simulating the Internet is that it is constantly changing. Therefore it is important to combine methods like simulation, experiments, measurements and analysis when testing things for the Internet.

Investigations have shown that audio streams will not have a large amount of consecutive packet losses unless the network is heavily loaded [18]. This can be modelled by a two-state Gilbert model according to Bolot [25]. This model has two state, state 0 meaning no packet loss and state 1 which represents a packet loss, see Figure 9. There is one probability *p* that you leave state 0 and go to state 1. Another probability *q* that you return to state 0. The *q* parameter enables the Gilbert model to have a larger or smaller probability for consecutive errors than a single error.

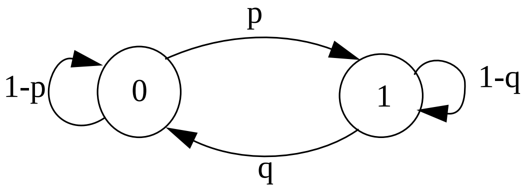

FIGURE 9. The Gilbert model

The distribution for consecutive losses for a very large measurements of TCP traffic was fitted to a Pareto distribution with $\alpha = 1.06$ by Paxon [26]. This distribution has an infinite variance which results in very large variability. The failing of the Gilbert model compared to real traces from the Internet is the lack of extremely large consecutive losses that occurs because of network pathologies. These pathologies can depend on a number of reasons, one common is synchronization of routing updates. When routing updates arrive from a number of different routers simultaneously, they overload the router, which results in packet loss. These losses have been observed in measurements made by Bolot et al. [18]. Loss bursts have also been observed in lengths up to 1206 packets [27] but more commonly in lengths of 30 packets. If so many consecutive packets are lost it will result in a half second, or longer, loss of speech.

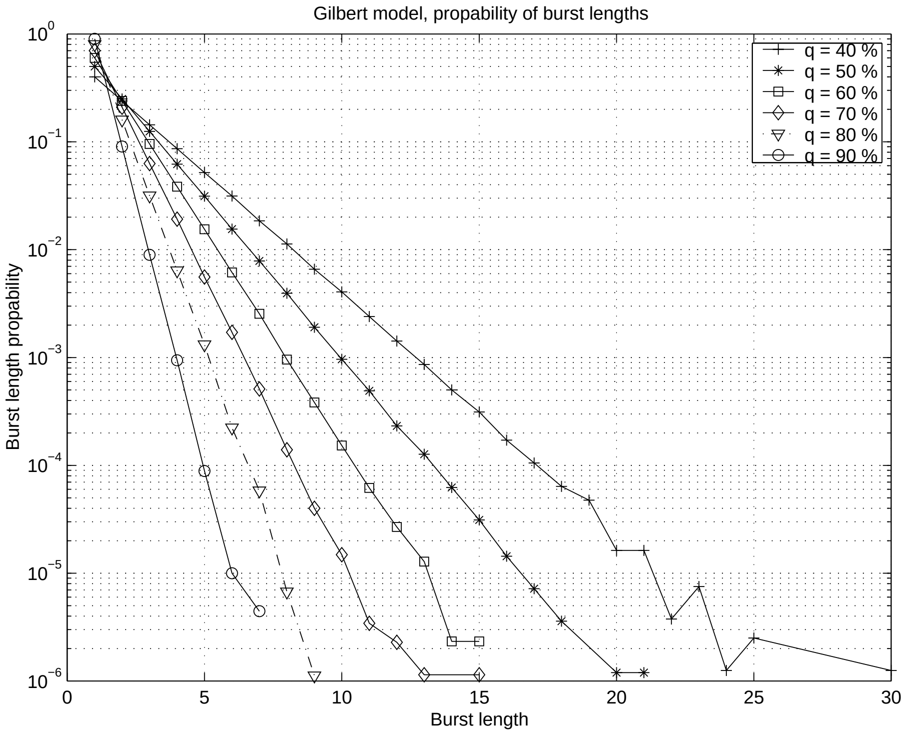

FIGURE 10. Burst length probabilities for different values of q

The Gilbert model burst length can be controlled by selecting the probability q. In Figure 10, the probabilities for a couple of values of q are presented. The value of q found in Internet measurements is around 0.7 according to Bolot et al. [25].

Packet loss patterns created by the Gilbert model will be used in simulations and analysis. Most important is, however, investigations to determine the effect of only single, double and triple packet losses. These methods seem satisfactory for the analysis of how forward error correction will improve performance in a VoIP application. A big problem is the lack of good objective measurements that would enable a larger amount of material to be tested.

5 Simulations and Results

The implementation of GSM-RED has been successful in most aspects, resulting in a decoder that can be used for analyses. The analyses done are objective measurements with SNR-SEG and PSQM, and a subjective listening test of CMOS type to compare GSM-RED with GSM-EFR. The results from the above tests have been analysed and commented with the purpose of explaining the reasons behind the measured results.

5.1 Simulation Chain

The simulation of the system for different channels is done in the following way.

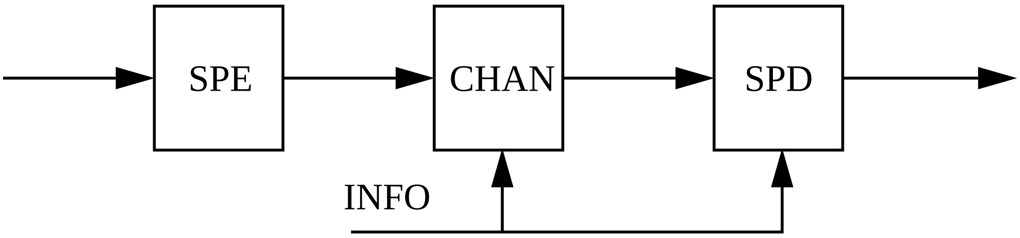

FIGURE 11. Simulation chain

The original speech is first run through the speech encoder (SPE), in Figure 11, then through a channel (CHAN) that modifies the encoded data in the frames that are marked lost by the INFO flow. These data are then decoded by the speech decoder (SPD). To simulate the mechanism detecting packet loss, the speech decoder also receives the INFO flow.

The modification done in the CHAN is that all data in frames marked as lost are randomized. The speech decoder does not actually need the CHAN to remove the packets that are lost, since this is signalled in the INFO flow. The randomization helps the GSM-EFR to avoid ringing, created by the algebraic codebook, that can occur if the data are copied from a previous good frame.

The INFO flows was created with a MATLAB program and four kinds of error patterns were used in the tests.

- Single packet loss only.
- Double packet loss, two consecutive packets are the only possible pattern and two double losses is never placed back to back creating longer bursts.
- Triple packet loss, three consecutive packets are lost in each loss occurrence. After each loss burst at least one packet arrives.
- Gilbert model with $q = 0.7$.

5.2 Results of Objective measures

Two objective measures are used, PSQM and SNR-SEG. They have both been used on the same material, which is a 100 seconds long audio file named wd99. This audio file consist of 58 seconds of speech with and without background noise. Speaker and processing are listed in Appendix B. The remaining 40 seconds consists of noise, music and tones. The last 40 seconds is not used in the measurements because the goal is to improve the quality of speech. GSM-RED is not as good at music and tones as on speech due to the known shortcomings of the vocoder model to reproduce this kind of sounds.

For each type of loss pattern 51 different probabilities of loss in the range 0 to 45-50% have been simulated. The same pattern has been used for both GSM-EFR and GSM-RED in each measuring point. The PSQM and SNR-SEG results have been calculated on the same simulated material.

5.2.1 SNR-SEG

SNR-SEG is measured with a lossless, clean decoding with GSM-EFR as reference. Therefore the point with no packet loss in all four measures have infinite SNR-SEG.

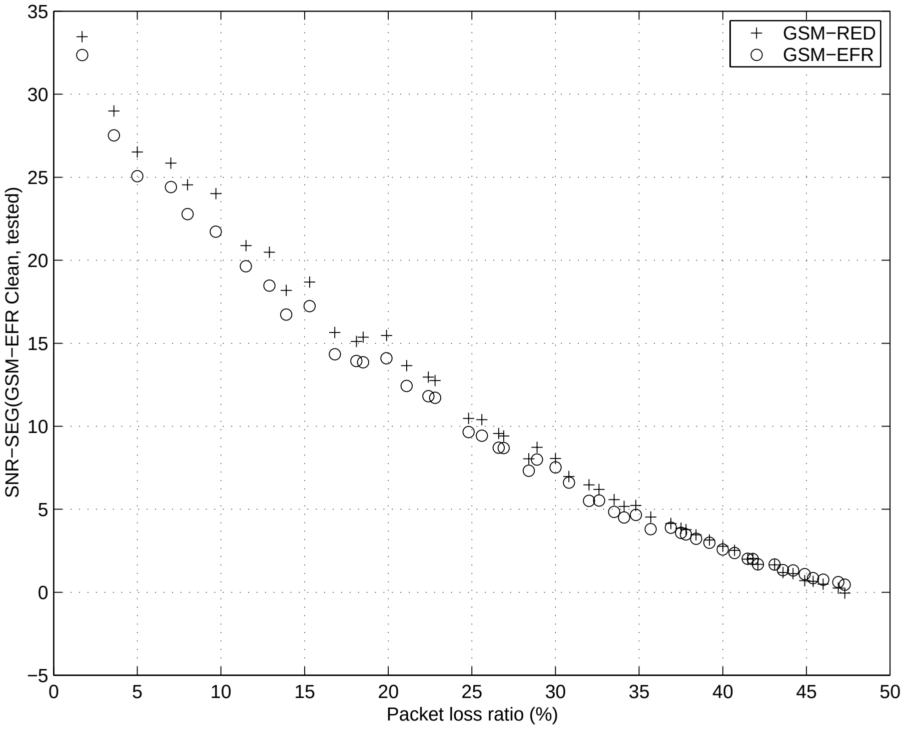

FIGURE 12. SNR-SEG for single packet loss in audio file wd99 compared with a lossless GSM-EFR decoding.

SNR-SEG for single packet loss (Figure 12) shows that GSM-RED has a few dB better result. This improvement gets less and less as the loss ratio increases and finally disappears when the packet loss ratio exceeds 35%. The improvement in SNR is likely due to the improved energy envelope in frames with packet loss. This advantage remains for the whole range but gets diminished by phase errors as the packet loss increases. Since the SNR-SEG measure is very sensitive to phase differences the measure drops rapidly for higher packet loss ratios. Despite all efforts in GSM-RED to keep the phase, it drifts a couple of samples over a frame and this causes the low SNR values. GSM-EFR error concealment has the same problem since the LTP vector is created from a previous good integer lag.

For small packet loss ratios (5-10%), the double packet loss (Figure 13) SNR values are not as high as for the single packet losses. This depends on the phase drift problem that gets worse since error concealment is applied on two consecutive frames. For GSM-EFR, error concealment is used on both frames while GSM-RED uses GSM-EFR EC for the first frame. The second frame is created from the redundant data and this increases the phase error(s). Also worth noticing is the SNR-SEG value for large packet loss ratio(>35%), it is 3-4 dB higher than for single packet loss (Figure 12). This depends on the length of the arrived packet bursts. For single errors at 40% packet loss ratio, almost every other packet is lost. The result of this is that the decoder never recovers after an error. When the loss bursts get longer the received packets also come in bursts that are longer, resulting in better recovery after loss bursts.

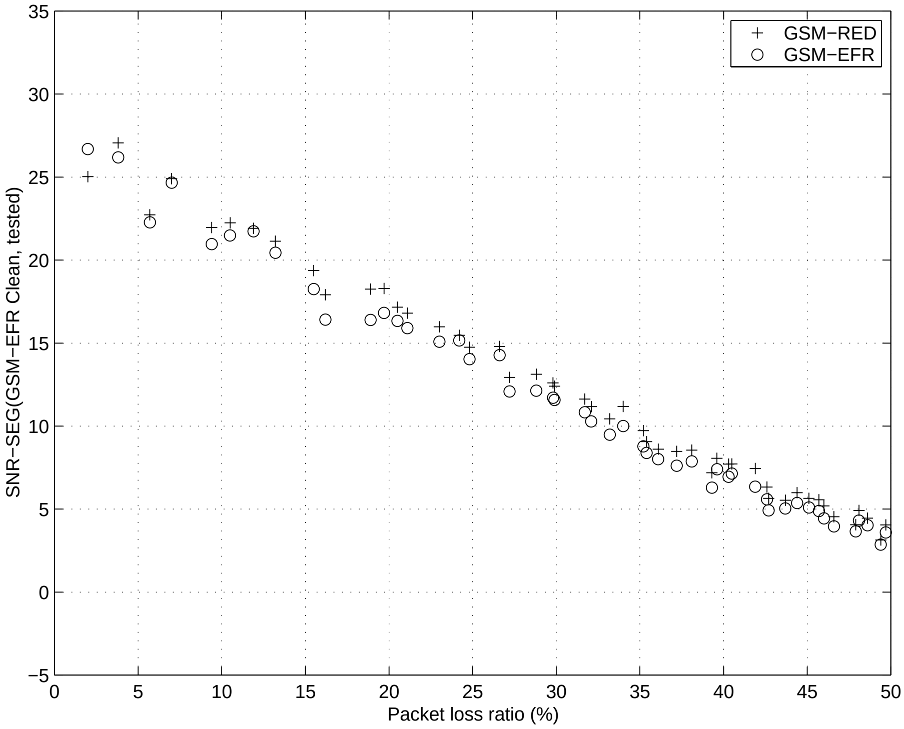

FIGURE 13. SNR-SEG for double packet loss in audio file wd99 compared with lossless GSM-EFR decoding.

Figure 14, for triple packet losses, shows that the previously mentioned effect of burst lengths is even more noticeable for triple losses. The distribution is more stochastic for low loss ratios than for double losses and that is probably caused by the positions of the loss bursts.

The Gilbert model (Figure 15) produces a result which is a combination of single packet losses and the longer bursts. The points distribution is almost as smooth as for single packet losses. The SNR-SEG for low packet loss ratios starts almost as high as for single packet losses and does not drop any lower in the high range than for double packet losses. A small but consistent improvement of SNR-SEG for GSM-RED compared to GSM-EFR can also be seen.

The SNR-SEG measurements are not especially conclusive and that has much to do with the properties of SNR-SEG. Primarily the sensitivity for phase errors makes this measure unreliable for measuring error concealment methods.

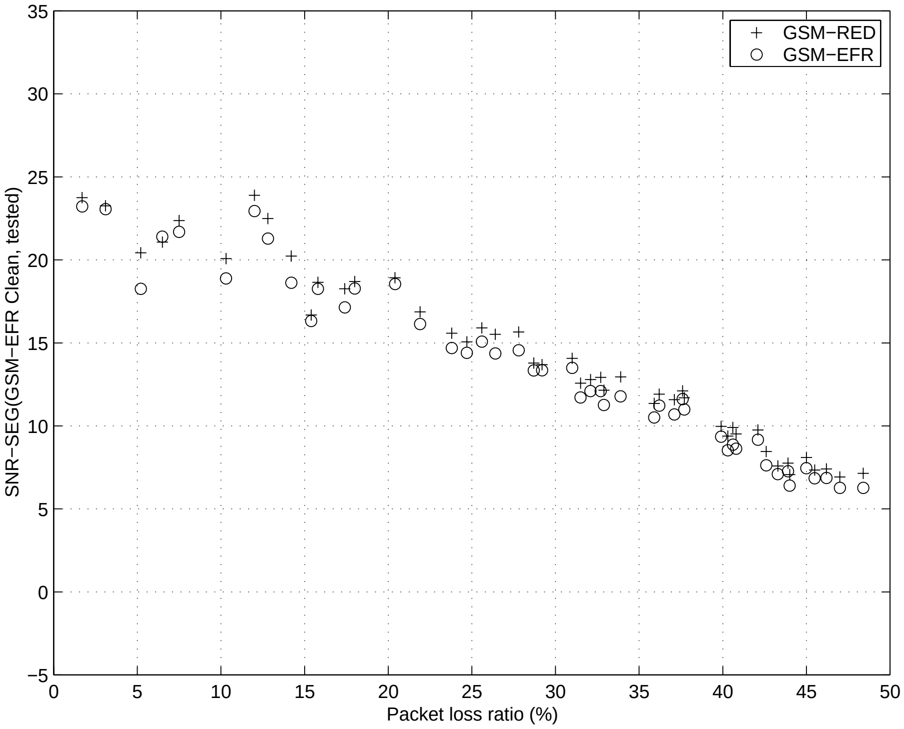

FIGURE 14. SNR-SEG for triple packet loss in audio file wd99 compared with lossless GSM-EFR decoding.

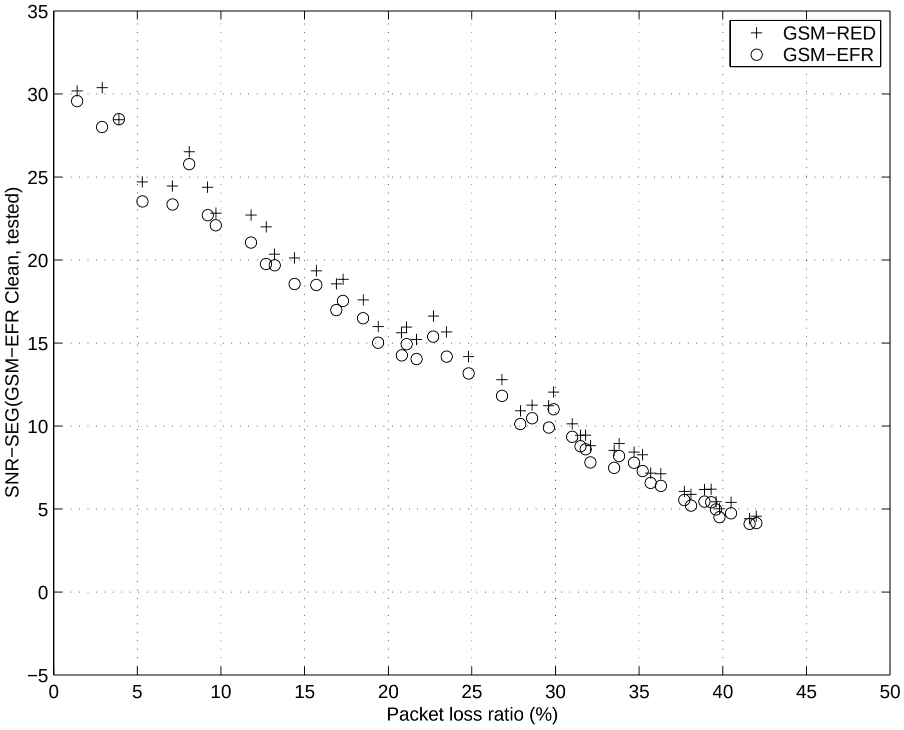

FIGURE 15. SNR-SEG for Gilbert(p,0.7) modelled packet loss in audio file wd99 compared with lossless GSM-EFR decoding.

5.2.2 PSQM

The PSQM measure is much less sensitive to small phase errors, since it measures the energy and frequency match. PSQM needs a reference file with energy adjusted to -26 dBov. dBov is decibel relative to the overload of a digital system. The tested files were also scaled, which resulted in some measuring points experienced amplitude clipping, i.e. when the value exceeds the available range of sixteen bits. The effect of amplitude clipping on PSQM measurements is not known [28]. The number of occurrences of clipping are probably very few, resulting in a minimum effect. A PSQM value of 0 equals the reference signal and 6.5 represents the worst.

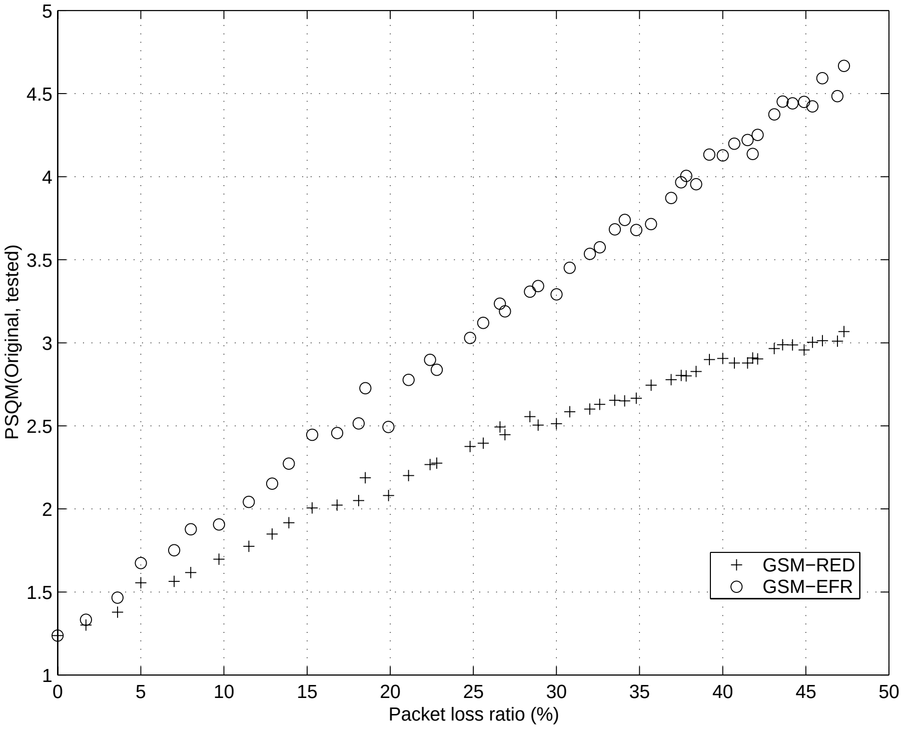

FIGURE 16. PSQM for single packet loss only in audio file wd99.

As seen in Figure 16 the improvement for GSM-RED over GSM-EFR is observable at as small packet loss ratios as 4%. The degradation for GSM-RED compared to GSM-EFR is significantly less for the whole range from 4-5% and upwards, especially in the high range (40-50%) of packet loss.

For double packet loss, presented in Figure 17, GSM-RED is an improvement compared to GSM-EFR in all measured points except for the case when no errors occur. The degradation for GSM-RED is slightly higher than for single packet loss only. This trend is even more noticeable for triple packet loss (Figure 18) where the gap between GSM-RED and GSM-EFR is the smallest of the four tests. This probably depends on the use of GSM-EFR’s EC for the first frame of a double loss, in GSM-RED. In a length three loss burst, two of the three frames are error concealed with the same method, creating more similar results. The results would probably converge even more if GSM-RED was not more responsive after a burst of lost packets.

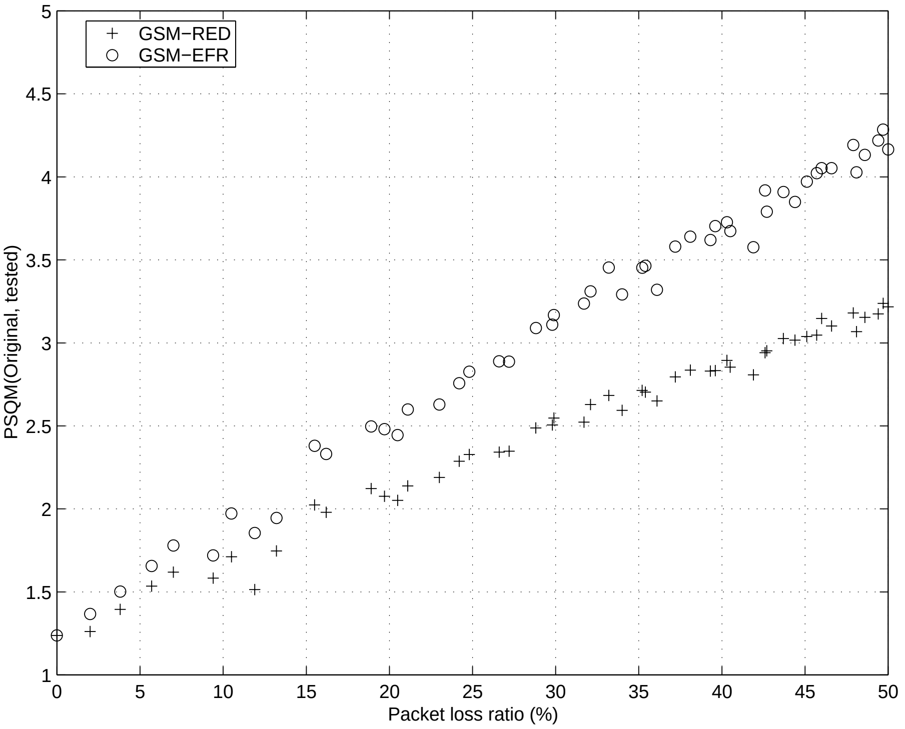

FIGURE 17. PSQM for double packet loss only in audio file wd99.

The random tendencies of the measured values also increase with increased length of the bursts. This is probably due to the placement of the errors. For longer bursts the number of occasions where errors occur, are fewer than for single packet losses, for the same packet loss ratio. This stresses the importance of where the losses actually occur in the sound file.

For the Gilbert model (Figure 19) the GSM-RED's curve is somewhere between the one of single losses and double losses. The random tendencies are also in the same magnitude as double losses which seems reasonable considering the distribution of burst lengths. GSM-EFR seems to have a performance envelope comparable with double packet loss.

Notice that the degradation for GSM-EFR is worst for single errors (Figure 16), all the other three tests have lower PSQM values for the high range. This depends on the error concealment algorithm in GSM-EFR. When there are only single packet losses and the packet loss ratio is around 40% almost every second packet is missing. The EC applies damping of gain factors in the frames after frame loss. This will result in almost no frames with unattenuated gain values. For longer burst errors, the error occasions are fewer and multiple packets in a row are received more frequently. This results in that the decoder has time to return to unattenuated decoding.

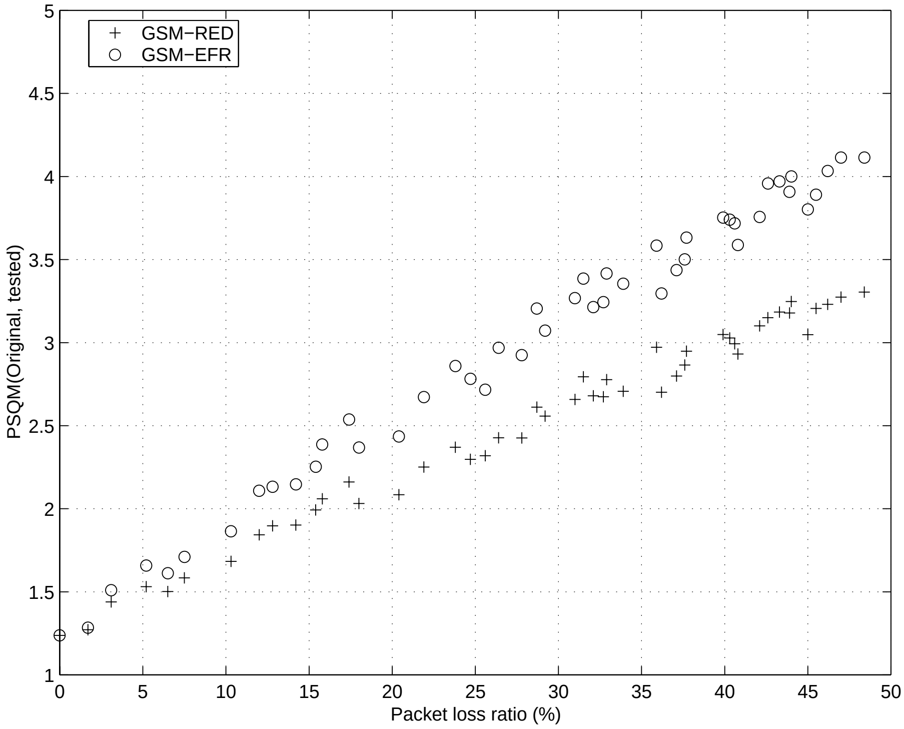

FIGURE 18. PSQM for triple packet loss only in audio file wd99.

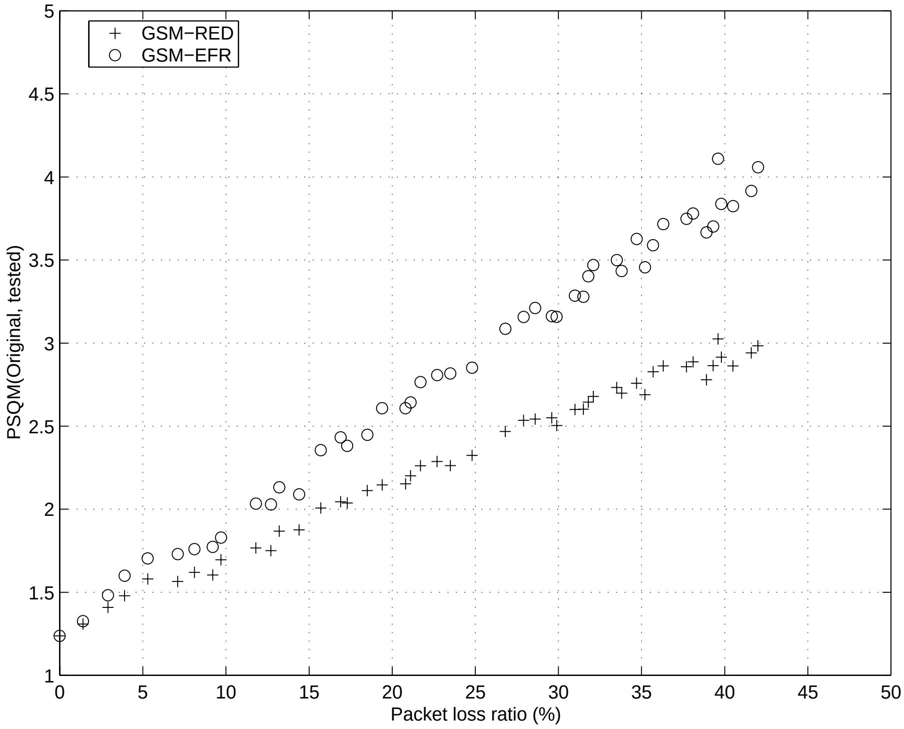

FIGURE 19. PSQM for Gilbert(p,0.7) distributed packet losses in audio file wd99

5.3 Subjective measures

To measure the subjective quality of GMS-RED, listening test were performed. The first listening test tries to identify the minimum packet loss ratio needed to detect the performance improvement for the GSM-RED compared with GSM-EFR. A second test measures the improvement of speech quality for GSM-RED compared to GSM-EFR when the packet loss ratio is higher than in the first test.

5.3.1 Test 1

This test tries to identify the minimum packet loss needed to detect an improvement in speech quality. The test is a CMOS test so a higher resolution in difference can be achieved than with an absolute measure, like MOS. In the same time the difference is measured, unlike the AB-tests, where only preference is measured.

The test consists of six swedish sentences from the audio file wd90, listed in Appendix C. These sentences were disturbed with three different packet loss patterns. The error patterns had a mean error rate of 3.2, 5.1, 7.9% and were generated according to the Gilbert model. The packet loss was not controlled in the sentences, resulting in widely fluctuating error rates in different sentences, as seen in Table 4. The burst lengths existing in each sentence are also presented. A score of 1 represents that GSM-RED was definitely better, 2 slightly better and 3 equal. The GSM-EFR is slightly better on 4 and definitely better on 5. The test was made by 11 persons all experienced listeners using binaural output to high quality headphones. The sentence number represent which sentence {1,2,3,4,5,6} it was and the letters represent which of the error patterns that was used {A= 3.2%,B=5.1%,C=7.9%}.

Sentence	Error rate (%)	Burst length distribution	CMOS	Variance
1A	0	0/0/0/0/0	2.95	0.24
2A	3.7	1/0/1/0/0	2.68	0.61
3A	3.6	4/0/0/0/0	2.27	0.68
4A	1	1/0/0/0/0	3.00	0.48
5A	1.8	2/0/0/0/0	2.09	1.13
6A	5.76	1/1/1/0/0	1.73	0.59
1B	5	1/0/0/1/0	2.36	0.43
2B	7.33	3/1/1/0/0	3.68	1.47
3B	6.4	3/2/0/0/0	3.00	1.33
4B	3	3/0/0/0/0	2.00	0.48
5B	7.2	5/0/1/0/0	2.95	1.00
6B	2.88	1/1/0/0/0	3.14	0.69
1C	5	4/1/0/0/0	3.14	1.27
2C	4.6	3/1/0/0/0	2.00	0.67
3C	6.4	7/0/0/0/0	2.95	0.81
4C	8.0	4/2/0/0/0	2.09	0.56
5C	10	5/3/0/0/0	1.73	0.87
6C	6.7	2/0/0/0/1	2.64	0.43

TABLE 4. CMOS test result with low MOS values are preference of GSM-RED and high for GSM-EFR.

Since the error rates and burst lengths are so different between different sentences, a meaningful mean is not possible to compute. A effort possible is to calculate mean with error rates in the same range, which is done in Table 5. The variance increases as the error ratio increases. The reason for this can be seen in Table 4. For the category with an error rate over 7%, sentence 2B has a CMOS value of 3.68 i.e. preference for GSM-EFR. Sentence 5B has been ranked as equally good and the last two has strong preference for GSM-RED. This shows inconclusive results for the overall performance of GSM-RED in this test.

There are few sentences 2B, 6B and 1C where the GSM-EFR is preferred and another couple, 4A, 3B, 5B and 3C where GSM-EFR and GSM-RED are rated as equal. For the sentences with preference for GSM-EFR, the reason is probably distortions in the speech that make GSM-RED less pleasant to listen to than GSM-EFR. In the case when they are rated equal, the reason is probably that the errors occur in positions that have no or small effect on speech quality. For the remaining sentences, except for 1A which has no errors, the GSM-RED is preferred more or less. Studying the sentences that have strong preferences for GSM-RED 5A, 4B, 2C, 4C and 5C, all except 5A have rather small variances. The error rate and error patterns in this test are widely dispersed. This shows that significant improvements in speech quality can be achieved in error rates from 2-3% and for both single errors and longer bursts in these sentences.

Error rate range	Sentences	CMOS	Variance
1.0-4.0	2A, 3A, 4A, 5A, 4B, 6B	2.53	0.85
4.0-7.0	6A, 1B, 3B, 1C, 2C, 3C, 6C	2.55	1.01
> 7.0	2B, 5B, 4C, 5C	2.62	1.53

TABLE 5. Mean over all sentences in Table 4 with a given error rate.

This listening test failed in some aspects. It could not conclusively show at which error rate GSM-RED gives improvements. The reason is that the distortions that the decoder creates in some cases lowers speech quality. This test also fails to bring out the most important improvement over GSM-EFR, the improved speech intelligibility. It was also badly planned in respect to error rates and burst distribution. The conclusions that can be made are; improvement of speech quality in some cases and at very low loss ratios.

5.3.2 Test 2

When planning the second listening test the problems from the first was considered. To show the improved speech intelligibility another type of test should have been done. The problem is that there was no material available for this kind of test and there was no time to construct one. Therefore another CMOS test was planned, this time with higher error rates, where more conclusive results can be obtained despite the audible clicks and bangs.

The error rates were controlled so that each sentence will have the desired error ratio. These sentences were also from the audio file wd90 but a different selection, listed in Appendix D. The error patterns used were single and double packet loss. The number of sentences were four; two male and two female speakers. The test consisted of error rates of 6, 13 and 20% packet loss for either single or double packet loss only. The demand of having as correct packet loss ratio as possible in each sentence was not possible to fulfil for 13% double losses. This depends on the length of the sentences that are around 100 frames and for which 13% losses is an odd number of frames. The loss ratio nearest 13% that was possible to create was 13.7%.

The listening test was performed by 11 persons using binaural output to high quality headphones. No special instructions on how to grade were given. Low grades represents preference for GSM-RED and high for GSM-EFR i.e GSM-RED is: 1 definitely better, 2: slightly better, 3: equal, 4: slightly worse, 5: definitely worse, than GSM-EFR.

Error model	CMOS	Var	95% confidence interval (Low/High)
Single losses 6.0%	2.84	1.01	2.16/3.52
Single losses 13.0%	1.98	0.64	1.44/2.52
Single losses 20.0%	1.67	0.73	1.10/2.24
Double losses 6.0%	3.13	1.08	2.43/3.84
Double losses 13.7%	2.80	1.04	2.11/3.45
Double losses 20.0%	2.75	1.29	1.99/3.51

TABLE 6. Result of the CMOS test with higher error rates

Table 6 contains the results from the test. The 95% confidence interval assumes a normal distribution of the votes. For 13.0% single losses only there is a significant preference for GSM-RED. At 6.0% it is hard to make any conclusions, the variance is too big compared to the difference from 3.0 for the CMOS value. This is because the GSM-RED speech distortions are more unpleasant for a listener than the fade out produced by GSM-EFR. The result for each sentence for 6% single losses is in Table E-1 and can be used to draw the same conclusions as in listening test 1. At 20.0% single losses the preference is even more resolute for GSM-RED, but the variance is slightly bigger than for 13.0%.

For 6.0% double losses there exist a small preference for GSM-EFR. The reason for this is that the audible speech distortions, in the form of bangs and clicks, are more common. This is the case for all three error rates, making GSM-RED harder to prefer. The variance for all three ratios are also larger than for the single error case with corresponding error rate. For 13.7% and 20.0% there are a small preference for GSM-RED. However the large variance shows that the preference is not significant. There exists one sentence with 20.0% double losses that was rated with a large preference for GSM-RED. This indicates that improvements are also possible for double losses, however they are smaller. The algorithm for repairs following error concealment also needs more work.

The relative speech quality improvement for GSM-RED increases as the packet loss ratio increases. The reason for this is that the number of frames that receives any information compared to being lost in a non redundant scheme, increases. This makes it possible to have intelligible speech at 50% single packet losses, when some information is received for all frames. This trend can be seen for single packet losses as the loss ratio increases. The CMOS test also changes nature from a speech quality measure to an intelligibility measure when the packet loss rates increase. This happens when the speech quality becomes so low that intelligibility becomes the most important aspect.

6 Conclusions

One goal of this work was to design and implement a speech decoder that combines GSM-EFR and redundant LPC-10 type data. This has been done successfully. The other goal was to achieve better speech quality than GSM-EFR for both low and high packet loss ratios. This goal is almost achieved, but some speech distortions remain that are believed to be possible to remove with further work.

The improved quality has been shown with PSQM as objective measure for the whole measured range and all types of loss patterns. The subjective listening tests show a definitive quality improvement at 13% single packet loss. For double packet losses GSM-RED and GSM-EFR produce more similar results. If the remaining speech distortions are removed or reduced, a much more conclusive result would be achieved. There are indications in the tests that even for a few percent of packet loss, an improvement can be detected.

When lacking the possibilities to do an intelligibility test, only the opinion on this matter can be presented. The writer and others who have listened agree that intelligibility is improved, especially for loss rates over 10%.

The amount of redundant data can vary depending on the speech. One controlling factor that can be used is the voicing state. This makes it possible to change the amount of data per frame between 35, 42 and 51 bits in this design.

When the achieved speech quality did not reach the expected quality in all aspects a number of measures are proposed. These are methods which are outside the scope of this thesis project to implement and test.

6.1 Suggestions for further studies

A speech coder using both redundant and multi-layered techniques. An encoder that sorts parameters in quality order and divides them between multiple layers, where each layer could be used to receive basic quality and for each extra layer received improve the speech quality. The number of layers used can then be adapted to the network transporting the data. Can such a scheme perform better than the scheme in this thesis?

A study of how Discontinuous Transmission (DTX) affects the performance of FEC transmission and decoding. Can FEC information be sent in another way when DTX is used?

Make an IP speech coder, that uses the following methods: No predictors in quantizations, less dependencies on histories, encoding of important states. However it is important that the error free speech quality is not less than for other speech codecs with the given bitrate.

Appendix A, Abbreviations

ACELP	Algebraic Code Excited Linear Predictor
ACR	Absolute Category rating
ADPCM	Adaptive Delta Pulse Code Modulation
Alg CB	Algebraic CodeBook
AMDF	Average Magnitude Difference Function
AMR	Adaptive Multi-Rate
BC-lib	Baseline Codec Library, Ericsson speech coder development environment
CELP	Code Excited Linear Predictor
CMOS	Comparative Mean Opinion Score
CRC	Cyclic Redundancy Check
dBov	Decibel relative to overload of a digital system
DCR	Degradation Category Rating
DMOS	Degradation Mean Opinion Score
DRT	Diagnostic Rhyme Test
DTX	Discontinuous transmission(TX)
EC	Error Concealment
ETSI	European Telecommunications Standards Institute
FFT	Fast Fourier Transform
FEC	Forward Error Correction
GSM	Global System for Mobile communications
GSM-EFR	GSM Enhanced Full Rate
GSM-RED	GSM-EFR with GSM-VOC redundancy
GSM-VOC	Vocoder like LPC-10 implemented with GSM-EFR methods
ITU-T	International telecommunication union’s sector for telecommunicationstandardization
IP	Internet Protocol
IRS	Intermediate Reference System
LP	Linear Predictor
LPC	Linear Predictive Coding
LPC-10	U.S.A. government standardized vocoder
LSF	Linear Spectral Frequency
LSP	Linear Spectral Pair
LTP	Long Term Predictor
MA	Moving Average
MNRU	Modulated Noise Reference Unit
MOS	Mean Opinion Score
PCM	Pulse Code Modulation
PSQM	Perceptual Speech Quality Measure
RMS	Root Mean Square
RSVP	Resource reSerVation Protocol
RTCP	Real-Time Control Protocol

RTP	Real-Time Protocol
SNR	Signal to Noise Ratio
SNR-SEG	Segmental SNR
TCP	Transport Control Protocol
TX	Transmission
UDP	User Datagram Protocol
VoIP	Voice over IP
QoS	Quality of Service

Appendix B, Sentences in audio file wd99

Table B-1 contains information about the speech sentences in audio file wd99. All sentences are in english spoken by three male and three female speakers. The sentences are processed in different ways. The first four are unchanged since sampling and recording. The next four have been filtered by the Intermediate Reference System (IRS) sender filter [29]. This filter is a band-pass filter simulating the normal frequency response from telecommunication systems with a pass band between 300 and 3400 Hz. Six sentences have been subject to interfering noise from the inside of a car. The last six are subject to background babble from a restaurant.

Speaker	Processing	Start time	Stop time	Length (s)
Female 1		0	3.74	3.74
Female 2		3.74	5.48	1.74
Male 1		5.48	8.52	3.04
Male 2		8.52	11.32	2.8
Female 1	IRS send filter	11.32	15.10	378
Female 2	IRS send filter	15.10	16.80	1.70
Male 1	IRS send filter	16.80	19.80	3.00
Male 2	IRS send filter	19.80	22.64	2.84
Female 1	Car noise	22.64	25.20	2.56
Female 2	Car noise	25.20	29.20	4.00
Female 3	Car noise	29.20	31.58	2.38
Male 1	Car noise	31.58	34.00	2.42
Male 2	Car noise	34.00	37.10	3.10
Male 3	Car noise	37.10	39.87	2.77
Female 1	Background babble	39.87	45.00	5.13
Female 2	Background babble	45.00	48.58	3.58
Male 1	Background babble	48.58	52.50	3.92
Male 2	Background babble	52.50	55.20	2.70
Male 3	Background babble	55.20	57.98	2.78

TABLE B-1 Speech sentences in audio file wd99.

Appendix C, Sentences in listening test 1

The sentences from audio file wd90 used in the first CMOS listening test are listed in Table C-1. The sentence number corresponds with the one used in Table 4. The sentence spoken and by which speaker are presented in the second column. Positions in the wd90 audio file and each sentence length are also presented.

nr	Sentence (in swedish)	Start time (s)	Stop time (s)	length (s)
1	Male 1: Trädet blåste omkull i stormen	0	2.40	2.40
2	Female 1: Far läser högt för sin flicka	2.40	4.58	2.18
3	Male 2: Tyskar och fransmän är ganska lika	4.58	6.78	2.20
4	Female 2: Studenterna hälsar våren med sång	6.78	8.77	1.99
5	Male 1: Fjällen lockar många turister	8.77	10.98	2.21
6	Female 1: Kost och logi ingick i lönen	10.98	13.06	2.08

TABLE C-1 Sentences in audio file wd90 used in CMOS listening test 1

Appendix D, Sentences in listening test 2

The sentences in Table D-1 was used in the second CMOS listening test.

nr	Sentence (in swedish)	Start time (s)	Stop time (s)	length (s)
1	Male 1: Fjällen lockar många turister	8.77	10.98	2.21
2	Female 1: Kost och logi ingick i lönen	10.98	13.06	2.08
3	Male 2: Tupparna slogs så fjädrarna rök	13.06	15.23	2.17
4	Female 2: Träden växer vid den lugna dammen	15.23	17.25	2.02

TABLE D-1 Sentence from audio file wd90 used in CMOS listening test 2

Appendix E, Test results from listening test 2

Error model	Sentence	CMOS	Var	low outlier	high outlier	Most pref. listener	Least pref. listener
Single 6.0%	1	3.23	0.95	1	5		
	2	2.18	0.54	1	3		
	3	3.27	1.16	2	5		
	4	2.68	0.70	1	4		
	All	2.84	1.00			2.25	3.25
Single 13.0%	1	1.86	0.88	1	4		
	2	1.81	0.44	1	3		
	3	2.27	0.87	1	4		
	4	1.95	0.33	1	3		
	All	1.98	0.64		3.5	1.625	2.375
Single 20.0%	1	2.00	1.05	1	4		
	2	1.45	0.55	1	4		
	3	1.5	0.45	1	3		
	4	1.72	0.78	1	4		
	All	1.67	0.73			1.00	2.50
Double 6.0%	1	2.81	1.20	1	5		
	2	3.41	1.02	2	5		
	3	3.68	0.89	2	5		
	4	2.59	0.54	2	4		
	All	3.13	1.08			2.25	4.125
Double 13.7%	1	2.68	0.80	1	4		
	2	3.05	1.09	1	5		
	3	2.91	1.42	1	5		
	4	2.55	0.83	1	4		
	All	2.80	1.04			2.00	3.75
Double 20.0%	1	2.18	1.30	1	4		
	2	3.18	1.11	1	5		
	3	3.45	0.83	2	5		
	4	2.18	0.73	1	5		
	All	2.75	1.29			1.875	3.88

TABLE E-1 Result of second CMOS test for each individual sentence

Table E-1 contains results for each individual sentence in the second CMOS listening test. Tested for 6 different error conditions seen in the error model column. Results are for four different sentences and the joint result for that error model. Lower outlier is the lowest score that sentence received i.e. the largest preference for GSM-RED. High outlier is highest score this sentence received in the test i.e the least preference for GSM-RED. Most pref. listener is the mean for the listener who had the largest preference for GSM-RED in that error model. Least pref. listener is the listener who had most preference for GSM-EFR in that error model.

7 References

[1] C. Perkins et al., “RTP Payload for Redundant Audio Data,” RFC 2198, Sep 1997.

[2] International Telecommunication Union, “One-way transmission Time,” ITU-T Recommendation G.114, Feb. 1996

[3] L. Zhang, S. Deering, D. Estrin, S. Shenker and D. Zappala, “RSVP: A New Resource ReSerVation Protocol,” IEEE Network Magazine, Sep, pp 8-18, 1993

[4] M. Carlson, W. Weiss, S. Blake, Z. Wang, D. Black and E. Davies, ”An Architecture for Differentiated Services,” RFC 2475, Dec 1998.

[5] A. Spanias, ”Speech Coding: A tutorial review,” Proceedings of the IEEE, vol. 82, no. 10, Oct, pp. 1541-1582, 1994.

[6] Tremain T., “The government standard linear predictive coding algorithm: LPC-10,” Speech technology, April, pp. 40-48, 1982.

[7] P. Kroon and F. Deprettere, “A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbits/s,” IEEE Journal on selected areas in communications, vol. 6, no 2, Feb, pp. 353-363, 1988.

[8] K. Järvinen and J. Vainio, ”GSM enhanced full rate speech codec,” IEEE International conference on acoustics, speech and signal processing, vol. 2, pp. 771-774, 1997.

[9] European Telecommunications Standards Institute, ”Digital cellular telecommunications system; Enhanced Full Rate (EFR) speech transcoding (GSM 06.60),” 1996.

[10] C. Perkins, O. Hodson and V. Hardman, “Survey of packet loss recovery techniques for streaming audio,” IEEE Network vol. 12 no. 5 Sep-Oct 1998, IEEE p 40-48 0890-8044.

[11] European Telecommunications Standards Institute, ”Digital cellular telecommunications system; Substitution and muting of lost frames for Enhanced Full Rate (EFR) speech traffic channels (GSM 06.61),” 1996.

[12] European Telecommunications Standards Institute, ”Digital cellular telecommunications system (Phase 2+); Channel coding (GSM 05.03 version 8.0.0 Release 1999)”, 1999

[13] H. Schulzrinne, S. Casner, R. Frederick and V. Jacobson, “RTP: A Transport Protocol for Real-time Applications,” RFC 1889, Jan 1996.

[14] H. Schulzrinne, "RTP Profile for Audio and Video Conferences with Minimal Control," Internet Draft from Audio Video Transport Working Group, Feb 26 1999.

[15] B. Dempsy and Y. Zhang, “Destination Buffering for Low-Bandwidth Audio Transmissions using Redundancy-Based Error Control,” Conference on Local Computer Networks Oct 13-16 1996, Sponsored by: IEEE, IEEE pp. 345-354, 0742-1303

[16] V. Hardman, M.A. Sasse, M. Handley, and A. Watson, “Reliable Audio for Use over the Internet,” Proc. INET’95, 1995

[17] M. Podolsky, C. Romer, and S. McCanne, “Simulation of FEC-Based Error Control for Packet Audio on the Internet,” Proceedings. IEEE INFOCOM ’98, 3 vol. xxviii+1477 pp. p.505-15 vol.2, 1998.

[18] J-C. Bolot,H. Crepin, A. Vega-Garcia, “Analysis of Audio Packet Loss in the Internet,” Proceeding of 5th international workshop on network and operating system support for digital audio and video 1995.

[19] S. Dimolitsas, “Objective speech distortion measures and their relevance to speech quality assessments,” IEE Proceedings, Vol. 136, Pt. 1, No. 5, Oct 1989.

[20] J. Beerends and J. Stemerdink, “A perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation”, J.Audio Eng. Soc. Vol. 42, No. 3, March 1994.

[21] S. Dimolitsas, “Subjective Assessment Methods for the Measurement of Digital Speech Coder Quality,” Kluwer Academic Press, 1993.

[22] International Telecommunication Union, “Modulated Noise Reference Unit (MNRU),” ITU-T Recommendation P.810, Feb 1996.

[23] International Telecommunication Union, “Subjective performance assessment of telephone-band and wideband digital codecs,” ITU-T Recommendation P.830, Feb 1996.

[24] V. Paxson and S. Floyd, “Why We Don’t Know How To Simulate The Internet”, Proc 1997 Winter Simulation Conference, Dec 1997.

[25] C. Bolot, S. Fosse-Parisis, D. Towsley, “Adaptive FEC-Based error control for Internet Telephony,” Proc. Infocom ’99, New York, NY, March 1999.

[26] V. Paxon, “End-to-End Internet Packet Dynamics,” Proc. ACM SIGCOMM ‘97, 1997.

[27] J. Rosenberg, “Slides on Internet Loss Measurements”, presented to The voice on the Net Conference, Boston MA, sep 25, 1997, Avaiable at http://www.cs.columbia.edu/~jdrosen/papers/von-slides.pdf, sep 16 1999.

[28] International Telecommunication Union, “Objective quality measurement of telephoneband (300-3400 Hz) speech codecs,” ITU-T Recommendation P.861, Aug 1996.

[29] International Telecommunication Union, “Specification for an intermediate reference system,” ITU-T Recommendation P.48, 1988.