A Real-Time System for Voice over IP using the Adaptive Multi-Rate Speech Coder

Frank Mertz*, Rainer Martin*, Benoit Bossu*,
Tim Fingscheidt†, and Tobias Färber†

*Institute of Communication Systems and Data Processing (ind), Aachen University of Technology
Muffeter Weg 3, 52056 Aachen, Germany / E-mail: {mertz|martin}@ind.rwth-aachen.de

†Siemens AG, Information and Communication Mobile (ICM)
Grillparzerstraße 10a, 81675 München, Germany / E-mail: Tim.Fingscheidt@mch.siemens.de

Abstract

The transmission of coded speech over packet-switched networks, such as the Internet, has to deal with packet loss and packet delays, not occurring in the traditional circuit-switched speech transmission. In this contribution we present packet-switched transmission methods for speech frames encoded by the Adaptive Multi-Rate (AMR) speech coder, a state-of-the-art coding scheme that provides different switchable bit rates. By explicitly adding redundancy to the transmitted data packets we achieve a high degree of robustness with respect to frame erasures. The results we present were obtained by our real-time Voice over IP transmission system which are also described in this paper.

1 Introduction

Future telecommunication systems will probably utilize a single network for all kind of telecommunication services and they will integrate traditionally separated services into powerful unified applications. These systems will be based on an all-IP infrastructure using packet-switched technology. This will also affect the circuit-switched *plain old* telephone system, which will most likely migrate to a packet-switched system as well.

To achieve toll quality, the speech processing technology, mostly designed for circuit-switched networks, has to be adapted to deal with network congestion, packet delay and packet loss, occurring in packet-switched networks. Furthermore, network protocols have to be adapted and/or enhanced to support real-time transmission of, e.g., speech or video data.

An increasing use of Voice over IP systems will lead to higher network congestion, and therefore a reasonable data rate allocation has to be applied. This suggests the use of a sophisticated speech compression algorithm like the modern coders from the field of mobile communication systems. In our studies on Voice over IP transmission we use the AMR speech coder [1] for compressing the speech signal. This coder is based on CELP technology (Code Excited Linear Prediction) using algebraic codebooks (ACELP). The AMR has been standardized by ETSI[1] / 3GPP[2] and will be the mandatory speech codec in UMTS systems. It provides eight different bit rates, ranging from 4.75 kbit/s up to 12.2 kbit/s. It is possible to switch between adjacent modes back and forth even from one frame to the next. The codec was designed to allow a flexible allocation of the overall bit rate to the source and channel coders. To enhance the robustness against errors in bad channel conditions the bit rate used by the channel coder may be increased at the expense of the source coder bit rate. Adapting the coder mode to the channel quality provides highest possible speech quality in good channel conditions and a slightly lower base quality but increased error protection in noisy channel conditions. This flexibility of the AMR codec can also be utilized in packet-switched speech transmission systems, as we will show in section 4 of this paper.

Packet-switched transmission of real-time data like speech will not be restricted to fixed networks. As the upcoming third generation of wireless communication systems will be capable of packet data transmission at reasonable data rates, the application of Voice over IP technology in mobile communications is thinkable. This can be streaming audio as part of a video-stream or real-time telephony, e.g. in a video-conference or

[1] European Telecommunications Standards Institute

[2] Third Generation Partnership Project

as call initiated from a web-site.

In the following sections we will describe which protocols from the Internet protocol stack are already available for real-time data transmission, how our Voice over IP transmission system is structured, and what can be done to enhance the robustness of speech transmission with respect to frame erasures caused by packet losses.

2 Internet Protocols for Real-Time Applications

The standard protocols from the Internet protocol stack were not developed to provide transmission of real-time data streams, e.g. for audio or video applications. There is no means of requesting a special Quality of Service (QoS) within the currently used version of the Internet Protocol (IPv4). However, the shortage of available Internet addresses has been one of the reasons to initiate the development of a new version (IPv6), that among various modifications additionally provides QoS features.

It will still take some time until the Internet infrastructure will be upgraded to use the new protocol features in a large scale and until QoS guaranteeing routing strategies will be available. Until then the real-time dependent data packets have to compete with all other transmitted packets.

2.1 IP - Internet Protocol & UDP - User Datagram Protocol

The Internet Protocol (IP) is the network layer protocol of the Internet, among others providing the function for addressing the target host. One layer above, the transport layer provides two different protocols, TCP (Transport Control Protocol) and UDP (User Datagram Protocol).

The connection-oriented TCP numbers the packets and requests repeated transmission of lost packets. In real-time applications, however, there is no time for re-transmission in case of a packet loss. Therefore the smaller UDP is used in such applications, basically providing the addressing of the target application via port numbers. UDP works connection-less and operates a best-effort transmission of data packets. No guarantee is given whether the packets arrive in correct order or that they arrive at all. It does not number the packets and does not request repeated transmission of lost packets as TCP does. The ability to reassemble the speech data in correct order at the receiver is provided by the Real-Time Transport Protocol (RTP), described in the next section.

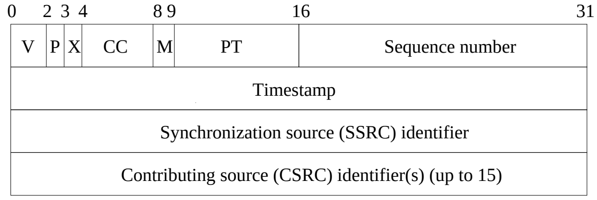
0
2 3 4
8 9
16
31
V
P
X
CC
M
PT
Sequence number
Timestamp
Synchronization source (SSRC) identifier
Contributing source (CSRC) identifier(s) (up to 15)

Figure 1: RTP Header
V: version, P: padding octets ,
X: header extension, CC: CSRC count,
M: marker bit, PT: payload type

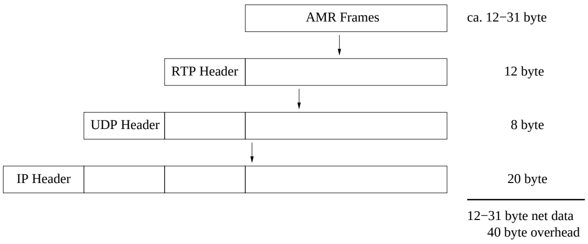

Figure 2: Assembling Internet packets for Voice over IP transmission

2.2 RTP - Real-Time Transport Protocol

The Real-Time Transport Protocol (RTP) [2] has been developed for the transmission of real-time data streams over the Internet. It provides packet numbering and timestamps to insure correct reordering of packets at the receiver.

The header that is added to the data packets by RTP is shown in Fig. 1. The most significant fields of the RTP header are the following. The payload type (PT) field defines the type of data contained in the RTP packet (e.g. MPEG-4 video or AMR speech coder frames). The SSRC (synchronization source) identifier is a unique number within an RTP session, identifying the sending source of the packet. Timestamp and sequence number allow correct reordering at the receiver and the synchronization of parallel video/audio packet streams. The sequence number is incremented with each packet being sent, and the timestamp reflects the sampling time of the first data sample from the packet's payload.

The payload of RTP packets itself gets a special RTP payload header, dependent on the type of payload which is sent via RTP. There is currently a standardization in progress to specify the payload format for AMR encoded speech frames [3]. The board responsible for standardization of Internet protocols is the Internet Engineering Task Force (IETF) [4], and responsible for audio and video transmission is the Audio Video Working Group within the IETF.

According to the current draft, the RTP payload header for AMR encoded speech frames provides fields for optionally specifying the use of CRC bits (cyclic

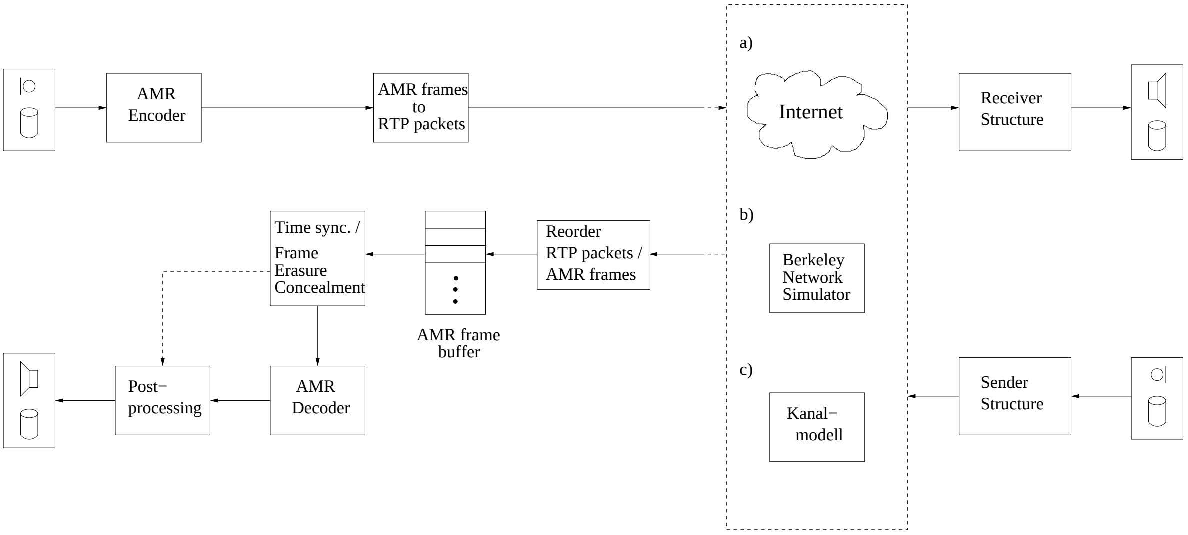
a)
AMR Encoder
AMR frames to RTP packets
Internet
Receiver Structure
b)
Time sync. / Frame Erasure Concealment
Reorder RTP packets / AMR frames
AMR frame buffer
Berkeley Network Simulator
c)
Post−processing
AMR Decoder
Kanal−modell
Sender Structure

Figure 3: Real-time system for Voice over IP transmission using the Adaptive Multi-Rate (AMR) speech coder and frame erasure concealment.

redundancy check), a field for requesting the usage of a special coder bit rate (Codec Mode Request - CMR), etc. It is possible to send an arbitrary number of speech frames within one RTP packet, the timestamp field describing the first speech frame within the packet. Following frames have to be interpreted accordingly.

2.3 RTCP - RTP Control Protocol

The RTP standard [2] additionally specifies a RTP Control Protocol (RTCP) which provides the transmission of feedback information on the quality of the current transmission. RTCP periodically transmits so-called Sender and/or Receiver Reports containing information on the amount of packets sent/received, the number of packets lost and an estimate of the inter-arrival jitter. The inter-arrival jitter is defined as the mean deviation of the time difference in packet spacing at the receiver compared to the sender. From timestamp fields an estimate of the current round-trip time can be calculated.

This feedback information can be utilized by the sender to choose an appropriate transmission scheme that is suitable for the current channel conditions. The receiver can use the information to adapt its jitter buffer length.

2.4 Assembling the Internet Packets

Fig. 2 depicts how the Internet packets are assembled. The speech data, together with the payload header, is first embedded into RTP packets that are subsequently attached to UDP and IP headers. The assembled RTP/UDP/IP packets are transmitted over the network.

From the given numbers it becomes clear that the overhead in data rate introduced by the various headers is enormous compared to the fairly small speech data rate. For instance, when sending one speech frame per RTP packet, encoded by the highest AMR bit rate of 12.2 kbit/s, a 40 byte header is necessary to transport about 31 byte of speech data. This becomes particularly relevant when there is a wireless link within the transmission route, like the link between a base station and an UMTS hand-held. In this case header compression algorithms [5] are absolutely necessary to reduce the protocol overhead.

3 Voice over IP System

We developed a system that provides a framework to study, develop and test transmission schemes and methods of frame erasure concealment for Voice over IP applications. Its structure is shown in Fig. 3. The speech data can be in-/exported by files or a microphone/speaker. The program works in real-time, providing studies with real-life networks and the capability of live telephone conferences. The speech data is encoded in frames of 20 ms by the AMR coder and packed into RTP packets using a specified packing scheme (e.g. one or more frames per packet, optional redundancy). The RTP packets are attached to UDP/IP protocol headers and submitted into a emulated or real-life network.

The receiver block, realized as separate program thread, evaluates the received RTP packets and retrieves the enclosed speech frames. The timestamp field from the RTP header describes where the first frame from the packets payload belongs in the current speech stream. Other fields from the payload header define if there are further frames enclosed in this packet and where they are placed in time relative to the first frame. The frames are retrieved from the packet and sorted into a frame buffer, the so-called jitter buffer. This buffer is needed to deal with the deviation in packet inter-arrival times, but also introduces a delay which depends on its length. The jitter buffer can be made adaptive to adjust its length to the current transmission delay [6].

Some of the packets might get lost during the transmission. This can be caused by overflowing queues at network nodes, or the packets do not reach the receiver in time and therefore have to be discarded. In a real-time application there is only a limited time to wait for packets to arrive. When a frame is still missing at the time it has to be decoded by the AMR decoder and played out, a frame erasure concealment technique has to be activated to replace the missing packet by an approximation in order to maintain an uninterrupted output signal. One possible method is to set the so-called BFI Flag (Bad Frame Indication) which is part of the AMR decoder. In wireless communication systems this flag is set in case a frame contains too many bit errors. When set the decoder uses information of previous frames to approximate the missing speech parameters. In the case of single missing packets this method is very effective, only slightly degrading the speech quality. In case several consecutive frames are lost, the signal is muted.

In addition to the RTP packet stream a RTCP packet stream is generated on another port to provide feedback information as described in section 2.3.

Besides the possibility to use the real Internet for transmission, we implemented a simple channel model that introduces determined packet loss rates. The packets may also be sent through a network emulation software. We use the emulation features of the Berkeley Network Simulator *ns* [7, 8] to emulate a complex packet-switched network, consisting of various nodes and links with specified attributes like queue length, transmission bandwidth, delay, etc. The emulation feature of this simulator allows to send a real-life packet stream (generated by our Voice over IP program) through the emulated network, running in real-time. The real-life stream is affected by emulated concurrent TCP and UDP streams and delayed or even lost in case of overflowing queues at certain network nodes.

4 Transmission Methods

The amount of speech frames effectively lost during a Voice over IP transmission can be reduced by explicitly introducing redundancy into the transmitted packets. To maintain the overall transmission data rate the base quality of the speech has to be slightly reduced by using a lower encoder rate. Several strategies are possible and the choice which to use might be adaptively controlled by the information on the current transmission quality supplied by the RTCP sender and receiver reports.

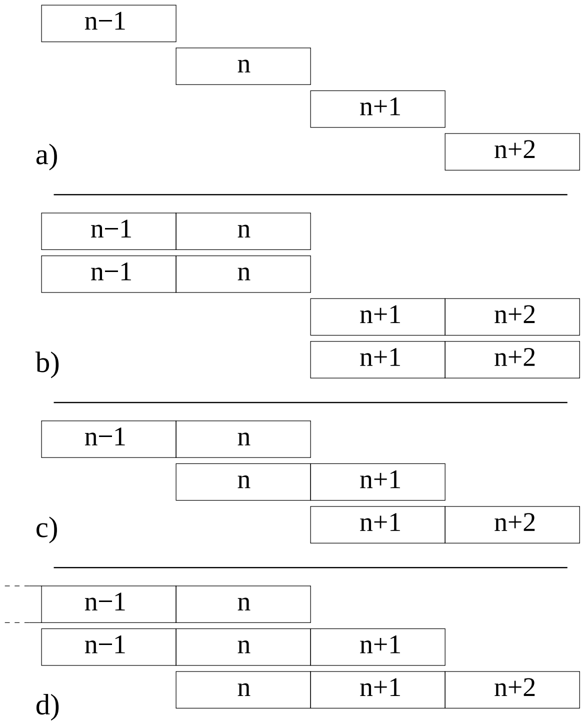

Figure 4: Schemes for packing AMR frames in RTP packets
a) single frame / packet
b) 2 frames / packet, packet sent twice
c) 2 frames / packet, overlap of 1 frame
d) 3 frames / packet, overlap of 2 frames

Fig. 4 shows some possible transmission schemes that have an overall payload data rate of about 12-14 kbit/s. When the packet transmission is reliable and only very few packets get lost or arrive too late, method a) is chosen, using the highest possible AMR coder rate (12,2 kbit/s) for the best possible base quality. When more packet losses occur, a switch to method b) or c) would be reasonable. Method b) packs two successive encoded speech frames into one RTP packet and transmits it twice. By reducing the coder rate to 5.9 kbit/s the overall bit rate including headers would be the same as with transmitting 12.2 kbit/s with method a). Sending the packets twice results in a reduction of the effective loss rate, as only one of them needs to be received in time. Alternatively method c) sends each packet only once but packs two successive speech frames in one packet in an overlapping fashion (each at 5.9 kbit/s). When a packet gets lost its backup sent with the next packet can be used instead. Thereby, single packet losses have no effect on the speech quality. Both methods

Packet Loss Rate	0 %	2 %	5 %
Method a)	91 %	55 %	20 %
Method c)	9 %	45 %	80 %

Table 1: Subjective preference of transmission methods at different packet loss rates

reduce the base quality of the speech signal by reducing the coder rate and utilize this to add redundancy to the transmission and thereby increase the robustness against packet losses.

When there occur even more packet losses resulting in an increasing amount of losing two successive packets, the coder rate might be further reduced and three successive speech frames with an overlap of two frames could be packed in each RTP packet.

5 Simulation Results

Using our Voice over IP system described in section 3 we ran some first simulations to study the different frame packing methods presented in the previous section. In case all redundant data of a speech frame is lost, the BFI flag is set, activating the concealment algorithm of the AMR decoder as explained in section 3.

Tab. 1 shows the results of a listening test at our laboratory. In an A/B comparison test the listeners had to judge between speech files transmitted by the methods a) - 12.2 kbit/s encoded frames without additional redundancy - and c) - 5.9 kbit/s encoded frames with redundant transmission - at different packet loss rates. Ten listeners participated in this test. Eight phonetically balanced speech files were used for each test condition. With zero packet losses the quality judgment is reduced to a decision between the AMR coder modes 12.2 kbit/s and 5.9 kbit/s. For this condition the listeners mostly preferred the quality of the higher bit rate mode. At 2 % packet loss rate the quality of the 12.2 kbit/s mode is slightly degraded by the use of BFI frame erasure concealment in method a). In method b), however, single missing packets do not result in missing speech frames because of the redundant copy in the following packet. The listeners judged the quality of both methods as about the same in this case, slightly preferring the higher bit rate mode because of its higher base quality. When the packet loss rate is increased to 5 %, the concealment efforts in method a) become more audible, resulting in a speech quality inferior to method c). The latter only needs to use the BFI concealment in case two successive packets are lost.

The quality of method b) proved to be about the same as method c). Because of its poorer base quality, method d) will not perform better until the loss rate increases much further and/or a higher burstiness of packet losses occur.

6 Conclusions

In this paper we have given an introduction into the transmission of speech data via packet-switched networks and presented our Voice over IP transmission system. As we have shown, a fairly high robustness with respect to frame erasures can be reached by explicitly adding redundancy to the transmitted packets. However, some speech frames might still be lost. This requires an effective frame erasure concealment technique at the receiver to make these losses subjectively less audible. The BFI algorithm of the AMR speech codec has proved to be effective to conceal the remaining frame losses. Possible improvements of the frame erasure concealment technique, e.g. by exploiting information from already received packets, both before and behind the missing packet, will be part of our further studies. In this context MMSE parameter estimation techniques are useful [9].

References

[1] ETSI, *Spec. GSM 06.90: Adaptive Multi-Rate (AMR) Speech Transcoding.* European Telecommunications Standards Institute, 2000.

[2] IETF, *RFC 1889: RTP: A Transport Protocol for Real-Time Applications.* Internet Engineering Task Force, Jan. 1996.

[3] INTERNET-DRAFT, *RTP payload format and file storage format for AMR and AMR-WB audio.* IETF, http://www.ietf.org/ietf/lid-abstracts.txt, 5 ed., June 2001.

[4] "IETF (Internet Engineering Task Force)." http://www.ietf.org.

[5] IETF, *RFC 2508: Compressing IP/UDP/RTP Headers for Low-Speed Serial Links.* Internet Engineering Task Force, Feb. 1999.

[6] Y. J. Liang, N. Färber, and B. Girod, "Adaptive Playout Scheduling Using Time-scale Modification in Packet Voice Communications," in *Proc. of the Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2001.

[7] S. McCanne and S. Floyd, "ns Network Simulator." http://www.isi.edu/nsnam/ns/.

[8] K. Fall, K. Varadhan, and the VINT project, *The* ns *Manual.* VINT Project, Feb. 2001. www.isi.edu/nsnam/ns/ns-documentation.html.

[9] R. Martin, C. Hoelper, and I. Wittke, "Estimation of Missing LSF Parameters Using Gaussian Mixture Models," in *Proc. of the Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, IEEE, 2001.