Efficient QoS Support for Voice-over-IP Applications Using Selective Packet Marking *

Henning Sanneck[1], Nguyen Tuong Long Le[1] and Adam Wolisz[1 2]
[1]GMD Fokus [2]Technical University Berlin

Abstract

Much research has been devoted recently on maintaining an acceptable Quality-of-Service (QoS) for the transmission of real-time multimedia streams (including voice) over packet-switched networks like the Internet. The work has covered per-flow reservation (Integrated Services), which allows the assurance of tight bounds on loss and delay, however needs the maintenance of state about every flow inside the network. Another research direction (Differentiated Services) focuses on only qualitative QoS assurance on a per-packet basis which has better scaling properties by only maintaining state and enforcing QoS for aggregated traffic.

Besides the advantage of aggregation, per-packet QoS has also the desirable property that an application may control the desired QoS on a per packet (and thus per ADU - Application Data Unit) basis. We explore the potential of this property for packet voice as recent work has shown that some segments of the signal are essential to the speech quality while others, in the event of a packet loss, can be extrapolated at the receiver from data received earlier. This is especially true for modern frame-based codecs like the ITU-T G.729 and G.723.1 which contain an internal loss concealment. Thus, the senders could be less conservative in their network QoS requirements which increases the number of concurrent sessions which can be accepted by the network.

In this paper we first analyze the concealment performance of the G.729 decoder. Using this result, we then develop QoS support schemes which selectively mark packets to a higher priority at the sender dependent on the properties of the speech signal and the expected concealment performance. Objective quality measures (ITU-T P.861A and EMBSD) show that almost the same speech quality as if all packets of the data stream would have been marked can be achieved while approximately halving the amount of actually marked packets. The different markings are then enforced by the network, e.g. using the IETF Differentiated Services architecture.

1 Introduction

In recent years, both the general public and the research community have been showing significant interest in interactive speech transmission over the Internet (Voice over IP, Internet Telephony). Voice over IP has the potential to be integrated with other Internet applications to provide interactive multimedia communication services that are impossible (or at least very difficult) to deploy over the traditional telephone network. Additionally, high complexity speech encoding and decoding can be performed with inexpensive hardware in the end systems at user premises. Examples are the two frame-based codecs G.723.1 and G.729, which are very attractive for Voice over IP because they provide toll quality speech at much lower bit rates (5.3/6.3 kBit/s and 8 kBit/s respectively) than conventional PCM (64 kBit/s). Thus the network resource requirements for a large scale deployment can be reduced significantly.

However, today's packet-switched networks, like the Internet, are based on the ”best effort” principle which does not guarantee a minimum packet loss rate and a minimum delay of packet transmission required for voice communication. Speech packets can be discarded when routers or gateways are congested as well as when they arrive late at the receiver (i.e. their playout time has already passed). Furthermore, considering the backward-adaptive coding schemes of the G.723.1 and G.729 source coders, packet loss results in loss of synchronization between the encoder and the decoder. Thus, degradations of the output speech signal occur not only during the time period represented by the lost packet, but also propagate into following segments of the speech signal until the decoder is resynchronized with the encoder. To alleviate this problem, both G.723.1 and G.729 decoders contain an internal (codec-specific) loss concealment algorithm.

In a related paper we have presented the Speech Property-Based Forward Error Correction (SPB-FEC) scheme. There, only essential parts of the speech signal are protected by Forward Error Correction, while losses within other parts of the signal are treated by the internal G.729 loss concealment. This scheme has been shown to

*to appear in *First International Workshop on Intelligent Multimedia Computing and Networking (IMMCN 2000)*, Atlantic City, February 2000

significantly reduce the amount of needed redundancy while maintaining a good speech quality. However, SPB-FEC still faces the general problem of FEC schemes: transmitting redundant data also adds more load to the network and thus worsens congestion in the Internet. Besides, FEC schemes only reduce but cannot come close to eliminating the possibility of losing important frames. Moreover, if over a time interval no packets are lost on the transmission path, all redundant data transmitted during that interval waste network resources.

Therefore in this work we want to adopt the concept of selective protection of certain packets of a flow according to applications' preferences, however we map it to giving selected packets a higher priority.

2 G.729 Frame Loss Concealment

G.729 is also known as Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP) and operates at 8 kBit/s. Input data for the coder are 16-bit linear PCM data sampled at 8 kHz. G.729 is based on a model for human speech production. In this model, the throat and the mouth are represented by a linear filter (synthesis filter) and speech signals are produced by exciting this filter with an excitation vector. In G.729, a speech *frame*[1] is 10 ms in duration, corresponding to 80 PCM speech samples. For each frame, the G.729 encoder analyzes the input data and extracts the parameters of the Code Excited Linear Prediction (CELP) model such as linear prediction filter coefficients and excitation vectors. The approach for determining the filter coefficients and the excitation is called analysis by synthesis: The encoder searches through its parameter space, carries out the decode operation in each loop of the search, and compares the output signal of the decode operation (the synthesized signal) with the original speech signal. The parameters that produce the closest match are chosen, encoded, and then transmitted to the receivers. At the receivers, these parameters are used to reconstruct the original speech signal.

The experiment we carry out is to measure the resynchronization time of the decoder after k consecutive frames are lost. The G.729 decoder is said to have resynchronized with the G.729 encoder when the energy of the error signal falls below one percent of the energy of the decoded signal without frame loss (this is equivalent to a signal-to-noise ratio (SNR) threshold of $20dB$). The error signal energy (and thus the SNR) is computed on a per-frame basis. Figure 1 shows the resynchronization time (expressed in the number of frames needed until the threshold is exceeded) plotted against the position of the loss for different values of k. The speech sample is produced by a male speaker where an unvoiced/voiced (uv) transition occurs in the eighth frame.

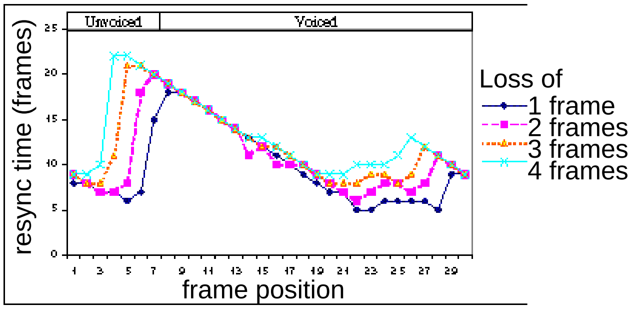

Figure 1. Resynchronization time (in frames) of the G.729 decoder after the loss of k consecutive frames ($k \in [1,4]$) as a function of frame position.

We can see from Figure 1 that the position of a frame loss has a significant influence on the resulting signal degradation[2], while the degradation is not that sensitive to the length of the frame loss burst k. The loss of unvoiced frames seems to have a rather small impact on the signal degradation and the decoder recovers the state information fast thereafter. The loss of voiced frames causes a larger degradation of the speech signal and the decoder needs more time to resynchronize with the sender. However, the loss of voiced frames at an unvoiced/voiced transition leads to a significant degradation of the signal. We have repeated the experiment for different male and female speakers and obtained similar results. Taking into account the used coding scheme, the above phenomenon could be explained as follows: Because voiced sounds have a higher energy than unvoiced sounds, the loss of voiced frames causes a larger signal degradation than the loss of unvoiced frames. However, due to the periodic property of voiced sounds, the decoder can conceal the loss of voiced frames well once it has obtained sufficient information on them. The decoder fails to conceal the loss of voiced frames at an unvoiced/voiced transition because it attempts to conceal the loss of voiced frames using the filter coefficients and the excitation for an unvoiced sound. Moreover, because the G.729 encoder uses a moving average filter to predict the values of the line spectral pairs and only transmits the difference between the real and predicted values, it takes a lot of time for the decoder to resynchronize with the encoder once it has failed to build the appropriate linear prediction filter.

[1] We use the term *frame* for the unit of the encoding/decoding operation and *packet* for the unit of transmission. One packet carries typically several frames.

[2] While SNR measures often do not correlate well with subjective speech quality, the large differences in the SNR-threshold-based resynchronization time clearly point to a significant impact on subjective speech quality.

```
protect $= 0$
foreach ($k$ frames)
    classify = analysis($k$ frames)
    if (protect $> 0$)
        if (classify $==$ unvoiced)
            protect $= 0$
            send($k$ frames, ”0”)
        else
            send($k$ frames, ”$+1$”)
            protect = protect$-k$
        endif
    else
        if (classify $==$ uv_transition)
            send($k$ frames, ”$+1$”)
            protect $= N - k$
        else
            send($k$ frames, ”0”)
        endif
    endif
endfor
```

Figure 2. SPB-MARK Pseudo Code

3 Speech Property-Based Selective Packet Marking

The result on the ability of the G.729 decoder to conceal packet loss is exploited to develop a new packet marking scheme called Speech Property-Based Selective Packet Marking (SPB-MARK). The SPB-MARK scheme concentrates the higher priority packets on the frames essential to the speech signal and relies on the decoder’s concealment for other frames.

Figure 2 shows the simple algorithm written in a pseudocode that is used to detect a uv transition and protect the voiced frames at the beginning of a voiced signal. In the algorithm, the procedure *analysis()* is used to classify a block of k frames as voiced, unvoiced, or uv transition. The procedure *send()* is used to send a block of k frames as a single packet with the appropriate priority (either ”+1” or ”0”). N is a pre-defined value and defines how many frames at the beginning of a voiced signal are to be protected. Our simulations have shown that the range from 10 to 20 are appropriate values for N (depending on the network loss condition). In the simulation presented in section 4, we choose $k = 2$, a typical value for interactive speech transmissions over the Internet ($20ms$ of audio data per packet). A larger number of k would help to reduce the relative overhead of the protocol header but also increases the packetization delay and makes sender classification and receiver concealment in case of packet loss (due to a large loss gap) more difficult.

4 Evaluation of the Speech Property-Based Marking Scheme

It has been feasible to employ the frame-based SNR for the experiments in section 2 because there we have examined only one system (G.729 without any per-packet protection) under different error conditions. Now, however, we will compare several systems (G.729 with permanent and different partial protection modes) under similar error conditions. The system with permanent protection will deliver the signal without any degradation (assuming no loss occurs when a packet marker is set) whereas the other systems partially rely on the internal concealment of the G.729 decoder, which is able to maintain a low signal degradation under the conditions described in section 2. However the relation of the resulting speech qualities cannot adequately be captured by an SNR (e.g. the gradual dampening of the gain coefficients of the previously received frame during the loss concealment improves the speech quality, but lets the recovered signal largely deviate from the original signal in the mathematical sense). Unlike the SNR methods, novel objective quality measures attempt to estimate the subjective quality as closely as possible by modeling the human auditory system. In our evaluation we use two objective quality measures: the Enhanced Modified Bark Spectral Distortion (EMBSD) and the Measuring Normalizing Blocks (MNB) described in the Appendix II of the ITU-T Recommendation P.861. These two objective quality measures are reported to have a very high correlation with subjective tests, their relation to the range of subjective test result values (MOS) is close to being linear and they are recommended as being suitable for the evaluation of speech degraded by transmission errors in real network environments such as bit errors and frame erasures.

We use a simple one-state Markov model (Bernouilli model) to describe the network behaviour as seen by each class of packets. ”Best effort” packets (designated by ”0” in Fig. 3) are dropped with the probability p (NO MARK case in Fig. 3). When full protection (FULL MARK) of the flow is assumed (packets are marked as ”+1”) the drop probability is 0. Packets of flows using the SPB-MARK scheme will either see no drops (”+1”, Fig. 2) or the drop probability p (”0”, Fig. 2). For comparison we use a scheme (ALT-MARK) where packets are alternatingly marked as being either ”0” or ”+1”. The system as a whole could then be described by a Markov model itself (e.g. a two-state Gilbert model), however as different marking schemes change the parameters of this model[3] we use only the internal system parameter p to be able to compare different approaches. Finally, we also will present results for schemes we classify as *”differential”* marking. Differential means here that any packet which is sent using a higher priority (”+1”) has to

[3] The ALT-MARK scheme e.g. sets the conditional loss prob. to 0.

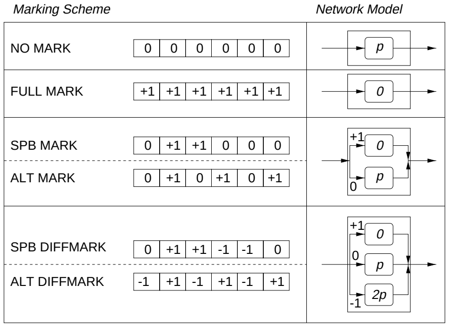

Figure 3. Marking schemes / network models.

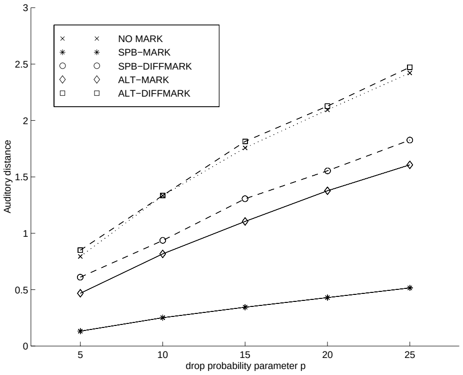

Figure 4. Auditory distance of marking schemes evaluated by MNB.

be compensated for by a less-than-”best effort” packet (”-1”). We compare again two flavours of this scheme: in the first (ALT-DIFFMARK) packets are alternatingly marked as being either ”-1” or ”+1”. The second scheme (SPB-DIFFMARK) is triggered by the SPB marking algorithm, however after a burst of ”+1” has been sent, a corresponding number of ”-1” packets is sent immediately[4]. The drop probability for ”-1” is $2p$. Thus seen over time intervals which are long as compared to the burst times the loss probability for the flow will be equal to the ”best effort” case.

In MNB, the perceptual difference between the test signal and the reference signal is measured at different time and frequency scales. The perceptual difference, also known as Auditory Distance (AD), between the two signals is a linear combination of the measurements where the weighting factors represent the auditory attributes. The higher AD is, the more the two signals are perceptually different and thus the worse the speech quality of the test signal is. Figure 4 shows the auditory distance evaluated by MNB[5].

The results for the unprotected flows (”NO MARK”) show that with increasing p in the network model (and thus increasing packet loss rate and loss correlation), the auditory distance is monotonically increasing, i.e. the speech quality of the decoded speech signals is decreasing. When comparing the ”NO MARK” results to the curves when marking is enabled, we can see that the decoded speech signal without marking has the highest auditory distance and thus the worst speech quality. The ALT-MARK scheme (50% of the packets are marked) enhances the perceptual quality. However, the auditory distance of the SPB-MARK scheme (with 40.4% of all packets marked[6]) is significantly lower and even close to the quality of the decoded signal without losses ($AD = 0$). This also shows that by protecting the entire flow only a minor improvement in the perceptual quality is obtained. The differential marking scheme (SPB-DIFFMARK) offers a better speech quality even when only using a network service which amounts to ”best effort” in the long term (while the ALT-DIFFMARK marking strategy does not differ from the ”best effort” case), needing only very limited network support. These results validate the strategy of our SPB marking schemes that do not equally mark all packets with a higher priority but rather protect a subset of frames that are essential to the speech quality.

5 Conclusions

We have investigated the impact of frame loss at different positions within a speech signal on the quality and gained the knowledge that the loss of voiced frames at the beginning of a voiced signal segment leads to a significant degradation in speech quality while the loss of other frames are concealed rather well by the G.729 decoder’s concealment algorithm. We have then exploited this knowledge to develop speech property-based marking schemes that protect the voiced frames that are essential to the speech quality by marking them with a higher priority while relying on the decoder’s concealment in case other non-marked frames are lost. Simulations using a simple network model and subsequent evaluation using objective quality measures show that the SPB-MARK scheme performs almost as good as the protection of the entire flow at a significantly lower number of necessary high-priority packets. The “differential” packet marking scheme SPB-DIFFMARK performs much better than the conventional best effort service, requiring only per-hop control over the loss patterns rather than the loss rates. All proposed marking schemes can be realized within the IETF Differentiated Services architecture.

[4] State about the necessary number of to-be-sent ”-1” packets is kept in the event that the SPB algorithm triggers the next ”+1” burst before all ”-1” packets necessary for compensation are sent.

[5] We have obtained similar results using the EMBSD measure (results are given in the full version of this paper).

[6] We obtained similar marking percentages using other speech material.