Available online at www.sciencedirect.com
SCIENCE DIRECT
Computer Communications 28 (2005) 582–588
computer communications
www.elsevier.com/locate/comcom

Assessment of objective voice quality over best-effort networks

Jari Turunen*, Pekka Loula, Tarmo Lipping

Tampere University of Technology, Pori Pohjoisranta 11, P.O. Box 300, FIN-28601 Pori, Finland

Received 14 May 2003; revised 21 January 2004; accepted 17 February 2004
Available online 22 December 2004

Abstract

VoIP calls transferred over dedicated bandwidth or QoS capable networks is a cost-effective alternative for PSTN in large enterprises. However, the calls made over the best effort network, such as the global Internet, suffer packet loss and jitter. In some VoIP-codecs, such as ITU G.723.1 and G.729a, there are built-in recovery mechanisms for concealing packet-based errors in the audio/speech stream. These recovery mechanisms can conceal up to 5% packet losses without significant quality degradation, as shown in this article. The 5% quality degradation is approximately within 0.5 MOS scale when compared to the original signal. Beyond 5%, the speech quality will drop gradually. The overall quality of MOS scale 3 can be maintained even with 14–17% packet loss rates. The influence of delay variation or jitter cannot be eliminated with the concealment algorithms unless the jitter time exceeds packet loss indication delay. The influence of jitter is not critical below 20 ms but beyond 20 ms limit its influence will decrease the speech quality very steeply. This suggests that packet losses can be recovered in normal conditions, but the influence of jitter must be eliminated somehow. The interleaving and piggybacking-based stream manipulation enhances speech quality in packet dropout situations. The guaranteed delay over the whole Internet would enhance the possibilities of VoIP to achieve success.

Keywords: Packet concealment; Jitter; Delay; VoIP

1. Introduction

The Voice over Internet Protocol (VoIP) calls were predicted to replace ordinary phone calls in all IP-based communication scenarios. However, it seems that the full-scale implementation and of the VoIP is still waiting its final success. Several reasons are pointed out in numerous research reports that have partially affected to the VoIP breakthrough as a communication channel. The main reason for the delayed success of the VoIP might be that the Internet was designed to be a fault tolerant data exchange/transmission medium and the traffic re-routing is the primary target in the case of Internet server additions and removals. The delivery time or transmission delay was never the primary goal in the design phase although the Internet Protocol (IP) has a support for real-time transmission. The IP-protocol comprises User Datagram Protocol (UDP) and Real-Time Protocol (RTP) that are primarily designed for stream transmissions. Despite the benefits that UDP and RTP can offer to the stream transmission, VoIP has suffered speech quality problems over the Internet, as reported for example in [1–3]. The voice packets are lost or damaged in transmission, the transmission delay might be too long for interaction and the changes in transmission rate can make the interaction with VoIP very uncomfortable.

The average global packet loss is around 5% in the Internet core networks and peak losses are well beyond 10%. The average round-trip delay is below 200 ms with peaks up to 500 ms, as seen in [4]. The round-trip delay consists of transmission time from sender to destination and the echoed message return time to sender. The end-to-end transmission delay time is approximately half of the round trip delay time. These values vary from time to time due to network traffic, congestion and routing situations [5]. Although the traffic and its changes on the Internet can be

* Corresponding author. Tel.: +358 2 627 2748; fax: +358 2 627 2727.
E-mail addresses: jari.j.turunen@tut.fi (J. Turunen), pekka.loula@tut.fi (P. Loula), tarmo.lipping@tut.fi (T. Lipping).

0140-3664/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2004.02.027

predicted with some accuracy, this information does not help the caller who tries to establish an urgent VoIP call over a congested network.

Jitter, or delay variation, is a harmful characteristic of the Internet. There are several aspects that cause jitter, starting from the traffic re-routing and different packet queuing mechanisms over the Internet [5,6]. The influence of jitter in media streams can be smoothed using buffers, but there are some problems which will be discussed later.

VoIP problems have gained worldwide interest and several studies have been made to improve packet loss recovery methods in packet based communication networks [7–19]. Also, the network point of view has been studied, for example, in [20,21], and several solutions have been proposed for ensuring the quality of service over the congested network. But still there are major problems in speech transmission and decoded speech quality despite of the developed solutions.

When considering new inventions in speech coding from a practical point of view, the adaptation of new technology to the existing equipment is a slow process, no matter how clever and simple the methods are. Also, interoperability must be guaranteed with the older VoIP phone versions. In this paper, we compare two commonly used VoIP coders, ITU-T G.723.1 and G.729a, widely used in most VoIP systems, and their ability to handle lost packets with and without a lost packet concealment system. The main purpose is to evaluate the usefulness and appropriateness of the built-in error concealment system in the coders in question. We made the codec evaluations using Objective Mean Opinion Score (OMOS) that gives an opportunity to make a more detailed analysis when compared to ordinary subjective MOS [22]. Later, we discuss different aspects that may have influenced the VoIP penetration level.

2. Codecs and network

In VoIP-based calls, the speech is encoded and packed to a RTP/IP frame for transmission. Nowadays three codecs (codec = encoder/decoder) are common in VoIP, the first is ITU-T G.711, because it is mandatory in VoIP protocols, and the other two are optional codecs, G.723.1 and G.729a. The latter two codecs are more efficient than G.711 because they compress the speech before transmission in order to save bandwidth in the networks. There are several other voice codecs that can be implemented in the VoIP phones, but they are not addressed in this paper due to their limited success. The G.711 codec is a logarithmic (a-law/μ-law) pulse code modulation codec operating at 64 kbit/s with toll quality. There is no packet error concealment algorithm in this codec, because the coding is performed on a sample-by-sample basis and the bitrate is absolutely too big in order to benefit from IP-networks. This codec will not be discussed later in this paper due to its simple mechanism without any built-in recovery mechanism. The G.723.1 is a two-mode near-toll quality codec with 5.3 and 6.3 kbit/s bit rates. The G.729a is also a near-toll quality codec operating at 8 kbit/s.

Built-in error concealment mechanisms are used in both G.723.1 and G.729a codecs. The packet concealment mechanism in G.723.1 was designed for burst lost packet errors ignoring the random bit errors. In the case of a lost packet in G.723.1 the linear predictive filter coefficients are estimated from the past values, but the new information is set to zero. The excitation signal is copied from the last good packet information combined with its voiced/unvoiced classification. The unvoiced excitation is obtained from the random number generator with re-estimated gain and the voiced excitation is generated with a periodic pulse signal according to the pitch period and gain from the classifier. If the erasure lasts for the next two packets, the excitation is attenuated 2.5 dB for each packet. After three estimated speech frames the decoder and estimation are stopped [23].

In the case of G.729a, the linear predictive filter parameters are also estimated from the last good packet values. The excitation is constructed using attenuated adaptive and fixed codebook values from the previous good frame [24,25]. The overview of the packet concealment methods is presented in Fig. 1.

Speech encoding, transmission, buffering and decoding takes time. Contrary to the public switched telephone network (PSTN) phone call, the spoken VoIP message will be delayed, due to reasons mentioned earlier, before the listener can hear it at the other end. A delay exceeding a certain threshold will cause disturbances in interaction and will feel very uncomfortable during the call. That is why in ITU-T G.114 recommendation, the maximum end-to-end delay, also referred to as 'mouth-to-ear' delay, in real-time communications is recommended to be less than 150 ms. An end-to-end delay consists of data collection, encoding delay, network transmission delay, network buffering delays, and decoding delay. In ITU-T multimedia communication-related recommendation, H.225.0, it is recommended that the default-framing interval for audio should be 20 ms. The usage of longer packets than 20 ms is allowed unless the overall delay exceeds the end-to-end delay mentioned above. This means two 10 ms packets of speech for G.729a and one 30 ms speech packet for G.723.1. In the case of packet loss in transmission, the more speech information the lost packet will contain the greater the degradation will be in the decoded speech quality.

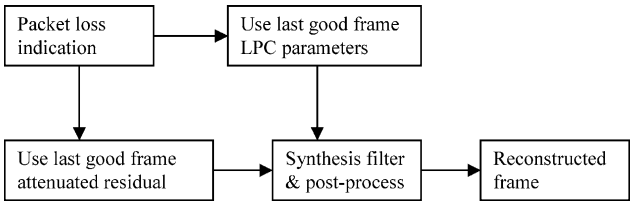

Fig. 1. The general schematic diagram of the packet reconstruction process.