Speech/Audio Coding for IP networks

Alan Duric, Roar Hagen

Global IP Sound AB

ABSTRACT

From a user's point of view, long delay and packet loss severely disrupts speech quality and intelligibility, even for rather low loss rates.

This paper examines the existing codecs used for real time speech/audio applications through their design and depicts problems due to their properties which were tailored for usage in circuit switched and wireless networks. Speech coders which are intrinsically designed for packet networks score clearly better results in perceived quality tests. This is achieved by, enabling graceful degradation of perceived speech quality, sustaining its high quality levels for as high as 30% of packet loss ratio, unlike any of the existing speech coders. Additionally, certain techniques and methods of the new approach to speech/audio coding for IP networks are being discussed in the paper.

INTRODUCTION

The value of integrating voice and data networks is well known and is one of the major drivers for IP telephony or Voice over IP (VoIP).

As Internet connectivity becomes increasingly omnipresent, its use for conferencing applications and services becomes ever more appealing. A number of factors are changing the motivations for this technology, and its move onto the new converged network. Improvements in delivered quality of service, due to increasing bandwidths and more reliable equipment, are improving the quality of internet telephony. As well as application of network QoS advances, such as application of Diffserv [1] and Intserv [2]. With these changing factors, it is becoming more and more important to consider how to deliver the highest quality voice on the Internet and focus on the endpoint QoS improvements. The Internet as a ubiquitous transmission channel introduces a number of unique difficulties, which in addition to the network side, need to be addressed by the edge devices (edge QoS).

SPEECH CODING RUDIMENTS

Speech codecs compress the voice in such a way as to minimally affect voice quality and in turn the quality of service (QoS) delivered by IP telephony systems. The speech encoder converts the digitized speech signal (after A/D conversion) to a bit-stream, which is packetized and sent over the IP network. The speech decoder then reconstructs the speech signal from the packets received. The reconstructed speech signal is, therefore, an approximation of the original signal. Speech codecs are deployed at end points and, so, determine the achievable end-to-end quality.

A speech codec has several important features, including speech quality, bit or compression rate, robustness, delay, sampling frequency, and complexity.

CODEC FEATURES

The quality of speech produced by the speech codec will define the upper limit for achievable end-to-end quality. This will determine sound quality for perfect network conditions – no

packet loss, delay, jitter, echo, or other quality-degrading factors. Other factors affecting the overall sound quality include handling of different voices as well as the quality of non-speech signals such as background noise.

Historically, a number of speech codecs, through their design have been addressing BE (Bit Error) occurrence in the communication channel. However, in today's communication networks, predominated by IP networks, the speech codec must be able to deal with lost packets. This robustness determines the sound quality in a loaded network and also in congested situations where packet loss is likely to happen.

The delay introduced by the speech coder can be divided into algorithmic and processing delay. The algorithmic delay occurs because of framing for block processing, since the encoder produces a set of bits representing a block of speech samples. Furthermore, many coders using block processing also have a look-ahead function that requires a buffering of future speech samples before a block is encoded. This adds to the algorithmic delay. Processing delay is the time it takes to encode and decode a block of speech samples.

The complexity of a speech-coding algorithm dictates the computational effort required and the memory requirements. Complexity is an important cost factor for implementing a codec and generally increases with decreasing bit rate.

Increasing the sampling frequency from the 8 kHz used for telephony band products to the 16 kHz used for wide-band speech coding produces distinctly more natural, comfortable, and intelligible speech. This far, wide-band speech coding has found limited use in applications such as videoconferencing because speech coders mostly interact with the public switched telephone network (PSTN). There is no such limitation in voice over IP (VoIP) when the call is initiated and terminated within the IP network. Therefore, because of the dramatic quality improvement attainable, the next generation of speech codecs for VoIP will be wide-band.

The bit-rate (number of bits per second) delivered by the speech encoder determines the bandwidth load on the network. The packet headers (IP, UDP, RTP) also add a significant portion to the total bandwidth. Speech quality generally increases with increased bit rate.

ECHO CANCELATION

Echo is a severe distraction, if the round trip delay is longer than 40 msec. Since the delays in IP telephony systems are significantly higher, the echo is clearly audible to the speaker. Therefore, canceling echo is essential to maintain high quality. Two types of echo can deteriorate speech quality: network echo and acoustic echo.

Network echo originates from impedance mismatch in PSTN hybrids. The echo path is stationary, except when the call is transferred to another handset or when a third party is connected to the phone call. This results in an abrupt change in the echo path.

Acoustic echo occurs commonly in hands-free equipment and small devices. The echo occurs when the loudspeaker's sound reflects back to the microphone in an enclosed environment. This echo type also occurs because of the proximity of the loudspeaker and the microphone. The acoustic echo path is non-stationary.

IP NETWORKS

The influence of the data network on the perceived quality of the IP telephony call is significant. The quality of real-time voice communication over IP networks (above all being best-effort) is mainly determined by the delay and loss characteristics observed along the network path.

The delay experienced in a call occurs on the transmitting side, in the network and on the receiving side. Most delay on the transmitting side is due to codec (packetization and look-ahead) and processing delay. In the network, most delay stems from transmission time (serialization and propagation) and queuing time. Finally, the jitter buffer depth, processing, and, in some implementations, polling intervals add to the delay on the receiving side.

Jitter occurs because packets have varying transmission times. It is caused by different queuing times in the routers and by possible

different routing paths. Jitter results in unequal time spacing between the arriving packets, and requires a jitter buffer to ensure smooth, continuous playback of the voice stream.

Most packet loss occurs in the routers, either due to high router load or high link utilization. In both cases, packets in the queues might be dropped. Packet loss also occurs when there are problems in the transmission links. The result is data link layer error and the incomplete or damaged (CRC) packet being dropped. Furthermore, so-called "human factor" or configuration errors are a number of times reason for a packet loss or excessive jitter or delay.

Phenomenon of exponential growth of peer-to-peer networking, such as applications for file sharing, is putting even more pressure on access links and poses a number of issues for inelastic traffic of real time nature, such as Internet Telephony.

Moreover, waist deployment of 802.11 networks [3] and certain security mechanisms (such as IPSEC, which confronts a number of QoS mechanisms, by not allowing any fine granularity of the flows due to hidden port numbers etc.) are making even more stress on holistic approach to solve QoS issues from both, top-down and horizontal perspective, putting edge QoS in the spotlight for both perspectives.

LEGACY - CIRCUIT SWITCHED SPEECH CODING IN IP TELEPHONY

The most commonly used codecs for IP telephony today are G.711, G.729, and G.723.1 (at 6.3 kbps). These were designed for (or based on technology designed for) circuit switched telephony.

Mobile telephony has been the major driver for development of speech coding technology in recent years. All the codecs used in mobile telephony, as well as G.729 and G.723.1, are based on the code excited linear prediction (CELP) paradigm. Due to design for use in circuit switched networks, these codecs were intended to handle bit errors rather than packet loss.

ITU G.711

G.711 has been designed for usage in circuit switched telephony, as such it does not include any means to counter packet loss. The common remedy is to insert 'zeros' (zero stuffing) whenever packet loss occurs, leading to voice break-up and steep degradation of quality

Error concealment can be introduced by extrapolating/interpolating received speech segments. An example is the method described in Appendix I to G.711 [4], which, however, is an heuristic procedure and does not guarantee robust operation.

ITU G.723.1 and G.729

G.729 and G.723.1 belong to the CELP coder class, which are based on a coding paradigm that was designed for circuit switched telephony (especially mobile telephony). Basic speech quality is worse than PSTN quality (similar to mobile telephony).

The CELP coding process is based on inter-frame dependencies that lead to inter-packet dependencies. Error propagation resulting from such dependencies leads to poor performance when packets are lost or delayed. Speech quality degrades rapidly with increasing packet loss

Despite the fact that those coders have built-in error concealment methods, they suffer immensely from inter-frame dependencies.

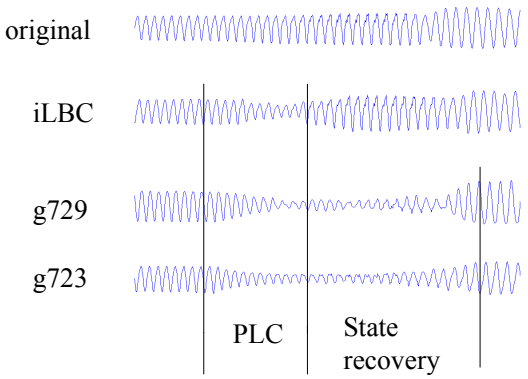

Figure 1 – advantage of speech coding with non inter-frame dependencies

As Figure 1 shows, the convergence times for resynchronizing encoder and decoder state are

noticeably greater then for iLBC codec [5], which is currently a work item of the IETF avt (audio/video transport) Working Group. Higher perceived quality, when iLBC and other new paradigm coders (which do not bear inter-frame dependency) deployed on such networks are shown through results of independent lab tests (which are contained in a later section of this paper).

Studies [7] indicate that these convergence times do not depend strongly on the number of consecutively lost frames and that convergence times are roughly 70 to 100 ms in duration, which affects perceived quality exceedingly. Furthermore, that time interval is shorter than typical network round trip times eliminating resynchronization protocols from being useful for G.729 or other CELP coders.

The average energy in the error signal (between the decoded errored bitstream and the decoded unerrored bitstream) is found to increase with increasing numbers of consecutively erased frames. This increase is sharp when the number of consecutive erasures jumped from one to two, and more gradually increasing from there. Subjective and objective evaluations confirmed this, and demonstrated [7], [8] that the G.729 concealment algorithm works fair for single frame erasures, but not more.

Finally, subjective evaluations confirmed (Figure 3) that it is more important to restore the state of the decoder after a frame erasure than to attempt to restore the speech that was lost during the frame erasures.

VoIP SPEECH CODING

Global IP Sound (GIPS) speech processing software for real time communications creates a paradigm shift, where the nature of the real-time, interactive communications over packet networks is taken into consideration, providing an edge-device QoS solution with very high voice quality, even under severe network degradations due to jitter and packet loss. Its design is characterized by the following principles:

- Speech quality in IP telephony must generally be equal-to or better-than PSTN.
- Speech quality must degrade gracefully with increasing packet loss and delay. Moderate packet loss should be inaudible

CODEC ENHANCEMENTS

The new paradigm for speech processing is (among other specifics) characterized by the fact that no inter-packet dependencies are introduced, precluding error propagation. Furthermore, such speech processing algorithms support diversity (Figure 2), which results in minimal loss of speech from packet loss. The speech signal is managed in a new and advanced way and the method does not add redundancy or send the same information as it is the case with FEC methods [9]. Multiple description coding (MDC) [10], generates multiple descriptions of the source signal of equal importance. These descriptions can be decoded independently at the receiver. If all descriptions are received, the source signal can be reconstructed in full quality. If received only a subset of the descriptions, the quality of the reconstruction is somewhat affected, however it is still evidently better than the quality resulting when all descriptions are lost.

Consequently, it is bandwidth efficient and ensures that packet loss leads to a graceful degradation of perceived quality. Diversity inevitably leads to increased delay. However, other developed and deployed solutions (Playout Controller [11]), mentioned in the following paragraph, utilize novel signal processing to dynamically minimize delay, making it similar to non-diversity capable solutions.

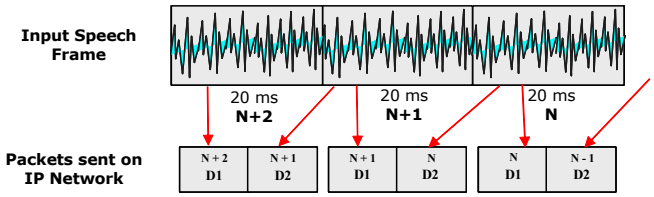

Figure 2 - MDC concept overview

Subjective MOS tests, conducted by independent labs show vast difference in perceived quality, by technology based on the new paradigm. As Figure 3 depicts, GIPS EG.711 (Enhanced G.711) is able to sustain toll quality even for adverse conditions such as 30 pct packet loss.

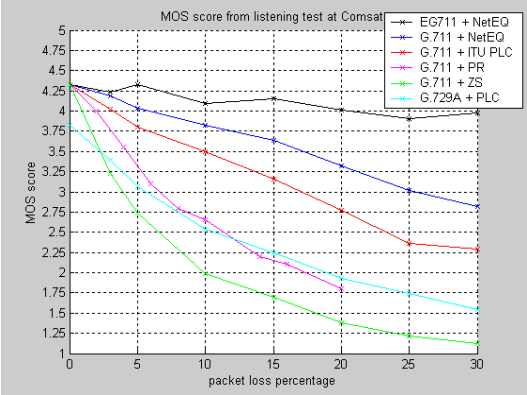

Figure 3 - MOS results from LMGT labs (Comsat) for narrowband codecs

Packet loss and jitter model in the test are based on a live network recording. GIPS EG.711 is bit exact to ITU G.711 for the cases of no packet loss. It is a low complexity coder and has a variable bit rate which is in average (depending on the periods of silence) slightly lower then for ITU G.711.

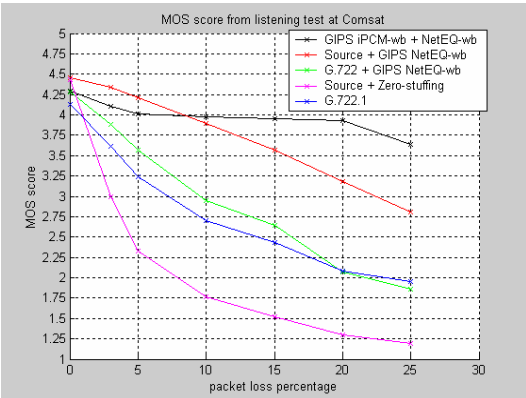

Figure 4 - MOS results from LMGT labs (Comsat) for wideband codecs

Compensation for network impairments is more difficult for wideband speech coders, due to wider frequency range that those coders are handling. Hence, all imperfections are very audible. However, by applying, methods which some are described in this section, GIPS iPCM-wb (Internet PCM) wideband coder was able to sustain very high quality level of perceived speech even for conditions with 20 pct packet loss (see Figure 4). Beside, high quality and high packet loss robustness iPCM-wb is characterized by low complexity and effortlessness when mixing with G.711 codec, making it very attractive for use on conference bridges where some of the endpoints involved in conversation are connected through PSTN, while others are connected directly through IP. iPCM-wb is a variable bit rate codec and its average bitrate is in range of 75 kbps (depends on periods of silence during conversation).

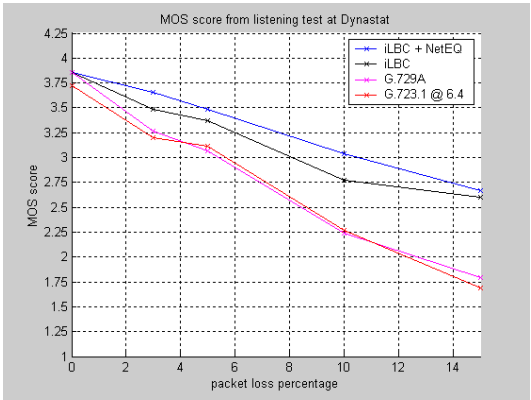

Figure 5 - MOS results from Dynastat labs for low bitrate narrowband codecs

As discussed in previous chapter, iLBC [5], which is a royalty free codec, and work item of IETF avt WG, does not endure of inter-frame dependency vulnerabilities and is distinguished by relatively high packet loss robustness, as seen on the Figure 5. iLBC is an algorithm for the coding of speech signals sampled at 8 kHz, which has a bit rate of 13.33 kbit/s and it is using a block-independent linear-predictive coding (LPC) algorithm. Codec operates at block lengths of 30 ms and produces 399 bits per block, which should be packetized in 50 bytes [6]. Support for 20 ms frames is being included. The described algorithm results in a speech coding system with a controlled response to

packet losses. At the same time, iLBC enables fixed bit rate coding with a very enviable quality-versus-bit rate and complexity tradeoff.

The essence of the codec is LPC and block based coding of the LPC residual signal. For each 240 sample block, the following major steps are done: A set of LPC filters are computed and the speech signal is filtered through them to produce the residual signal. The codec uses scalar quantization of the dominant part, in terms of energy, of the residual signal for the block. The dominant state is of length 58 samples and forms a start state for dynamic codebooks constructed from the already coded parts of the residual signal. These dynamic codebooks are used to code the remaining parts of the residual signal. By this method, coding independence between blocks is achieved, resulting in elimination of propagation of perceptual degradations due to packet loss. Additionally, the method facilitates good packet loss concealment (PLC).

VoIP END POINT ENHANCEMENTS

As it was mentioned previously, delay variation is a vital factor in real-time voice applications which obstructs the proper and timely reconstruction of the speech signal at the receiving end. Under the stringent delay requirements, packets could get lost due to their late arrival resulting from excessive network delay. One of the important functionality to be implemented at the receiver is the Playout Controller [11] of the voice packets. With the existence of delay jitter, the playout scheme greatly affects the tradeoff between loss and latency. Effective implementation of the Playout Controller, decreases end to end delay and additionally increases perceived quality by means of more effective packet loss concealment mechanism. Furthermore, it mitigates effect of clock skew which can occur in VoIP system due to unsynchronized sender and receiver.

In IP telephony systems, echo cancellers must also be optimized for IP traffic and packet loss in addition to solving the echo cancellation problems of PSTN. Special care has to be been taken to ensure that new echo cancellation solutions are robust against IP-specific problems, such as packet loss and jitter. This, together with a favorable performance-complexity ratio, makes the essence of echo cancellation products well-suited for IP telephony.

OTHER APPROACHES

Other approaches to deal with packet loss are FEC schemes [9] and interleaving schemes.

A packet loss may usually be followed by a burst of loss, which significantly decreases the efficiency of FEC schemes. In order to combat burst loss, redundant information has to be added into temporally distant packets, which introduces even higher delay. Hence, the repair capability of FEC is limited by the delay budget. Additionally, if FEC applied during the conversation, when packet loss is evidenced, it may even further deteriorate quality due to increased bandwidth (for cases where packet loss is caused due to bandwidth saturation on the access link) and if a number of endpoints deploy it at the same moment, it may lead to DoS (Denial of Service) alike scenarios.

Another sender-based loss recovery technique, interleaving, which does not increase the data rate of transmission, also faces the delay dilemma. The efficiency of loss recovery depends on over how many packets the source packet is interleaved and spread. Again, the wider the spread, higher the introduced delay [12].

DISCUSSION

Internet telephony promises to combine our separate data and voice networks into a single transport mechanism. The potential benefits for corporate and home users beside the reduced cost (needed to buy a single communication line, flat and lower rates etc.), are number of new services that this converged network is offering, combining traditional communications services with presence, positioning etc. However, there are significant barriers to acceptable QoS that have to be overcome [13], despite the fact that backbone networks are over provisioned, which would ensure good flow of data traffic, but not necessarily VoIP traffic. Wider deployment of QoS mechanisms will depend on the policy decisions of the involved network operators, which is particularly cumbersome when needed

to be deployed through several operator domains that call can traverse. Additionally, deployment of new applications (peer to peer file sharing) and technologies (e.g. .802.11 and IPSEC) make the QoS picture even more difficult. Therefore, focus is shifting to end-points, and improvements which could be achieved on that side by deploying suitable speech processing technology (codecs, play out controllers, echo canceling etc.), designed and optimized for packet network based environment. Subjective tests conducted by the third party evaluation labs show radical improvements which can be achieved by deploying such technology.

REFRENCES

[1] Blake, S., et al, "An Architecture for Differentiated Services", IETF RFC 2475, December 1998.

[2] Braden, et al "Resource Reservation Protocol (RSVP) - Version 1 Functional Specification", IETF RFC 2205, September 1997.

[3] G. Xylomenos, G. Polyzos, "TCP and UDP Performance over a Wireless LAN, INFOCOM 1999

[4] ITU-T Recommendation G.711, Appendix I, available online from the ITU bookstore at http://www.itu.int

[5] Andersen, et al., Internet Low Bit Rate Codec (iLBC)", IETF draft - draft-ietf-avt-rtp-ilbc-00.txt, September 2002.

[6] A. Duric and S. V. Andersen, "RTP Payload Format for iLBC Speech", IETF draft - draft-ietf-rtp-ilbc-00.txt, October 2002.

[7] J. Rosenberg, "G.729 Error Recovery for Internet Telephony", Columbia University Computer Science Technical Report CUCS-016-01, December 19, 2001

[8] Sun L., et al "Impact of Packet Loss Location on Perceived Speech Quality". Proceedings of 2nd IP-Telephony Workshop (IPTEL '01), Columbia University, New York, pp114-122, 2-3 April, 2001

[9] C. Perkins, et al, "RTP Payload for Redundant Audio Data", IETF RFC 2198, September 1997

[10] G. Kubin and W. Kleijn. Multiple-description coding (MDC) of speech with an invertible auditory model. In 1999 IEEE Workshop on Speech Coding Proceedings, pages 81–3, June 1999. Porvoo, Finland.

[11] H.Astrom, "Playout Controller for Real Time Speech/Audio applications", ETSI STQ Workshop "Compensating for Packet Loss in Real-Time Applications", February 2003, Sophia Antipolis, France

[12] B. Wah and D. Lin. Transformation-based reconstruction for real-time voice transmissions over the Internet. IEEE Transactions on Multimedia, 1(4):342–351, Dec. 1999.

[13] A. Markopoulou et al, "Assessment of VoIP quality over Internet Backbones:, IEEE Infocom 2002