Voice -over -IP (Voip) Bandwidth Optimization: A Survey of Schemes and Techniques

Daniel Uchenna Peter Ani
Dept. of Computer Science, Federal University Lokoja
Kogi State. Nigeria.
ucsoil@yahoo.com or uchenna.daniel@fulokoja.edu.ng

Abstract—
Voice over Internet Protocol (VoIP) is a voice communication system that proffers the transmission and reception of audio (voice) signal and (or) data over the internet. One very crucial problem in a VoIP network is inefficient bandwidth utilization which in most case is caused by header overhead resulting from the attachment of a big (40 bytes) to small payloads (10 to 30 bytes). Given the need to reduce the amount of bandwidth being wasted or perhaps improving the utilization of bandwidth which ultimately could enhance network performance and VoIP quality of service, several schemes and techniques have been proposed. The improvement and (or) optimization of bandwidth can be achieved through; packet header reduction and defines a process of multiplexing and de-multiplexing packet headers to curb overflows. Silence suppression offer a technique of repressing the silent portions (packets) in a voice conversation using Voice Activity Detection algorithm, as a result, the transmission rate during the inactive periods of speech is reduced, and thus, the average transmission rate can be reduced. Voice/ Packet Header compression is considered the most productive of all the technique, and it presents an event whereby VoIP packets are compressed from the conventional 40 bytes of size to a smaller byte size (2 bytes in the best case). Bandwidth save is reached using compression and decompression codecs of varying data and bit rates. It is envisaged that an improvement in the performance of codecs would yield a better result in terms of optimizing bandwidth in a VoIP network.

Keywords-component*; VoIP Optimization, Bandwidth Utilization, Packet Header Reduction, Packet Header Compression, Silence Supression, and De(Multiplexing.)*

1. INTRODUCTION

Voice over Internet Protocol (VoIP) is a voice communication system that proffers the transmission and reception of audio (voice) signal and (or) data over the internet. Just as the name implies, VoIP employs Internet Protocol (IP) as its most basic transportation method, over which both TCP and UDP are utilized. VoIP mechanism is such that voice is transmitted digitally in data packets using designated codecs that convert the voice into bits and bytes [1]. The resultant signal is transmitted via IP network infrastructure over the internet. Each packet retains information about its destination and timestamp amongst others which are necessary for effective packet reconstruction at the end.

It is envisaged that in time, VoIP technology will replace the conventional Public Switched Telephone Network (PSTN) Technology [2]. VoIP's emergence and dominance over PSTN is driven by numerous factors; one is that it utilizes the existing internet infrastructure, which reduces call cost tremendously. VoIP proffers higher reliability given that it automatically bypasses physical network issues of over congestion. It possesses the capability of executing calls irrespective of location by using a computer or any network device like PDA's [3].

Quality of service literally explains the satisfaction level of the experience felt by users while using any VoIP application. Technically, a user will identify break-ups in voice transmission and poor quality if there is high delay, loss or distortion. Hence, packet size choices are made based on delay through a network, through available bandwidth and packet loss within the network. One very crucial problem in a VoIP network is inefficient bandwidth utilization which in most case is caused by header overhead resulting from the attachment of a big (40 bytes) to small payloads (10 to 30 bytes). Data that are not useful are thus transmitted; causing a waste of bandwidth. There is thus a valuable need to reduce the amount of bandwidth being wasted, or perhaps improving the utilization of bandwidth which ultimately could enhance network performance and VoIP quality of service. Numerous techniques have been put

forward as likely solutions towards optimizing bandwidth usage, thus, the objective of this paper is to present the varying techniques for improving bandwidth utilization on VoIP with a view to identifying respective downsides.

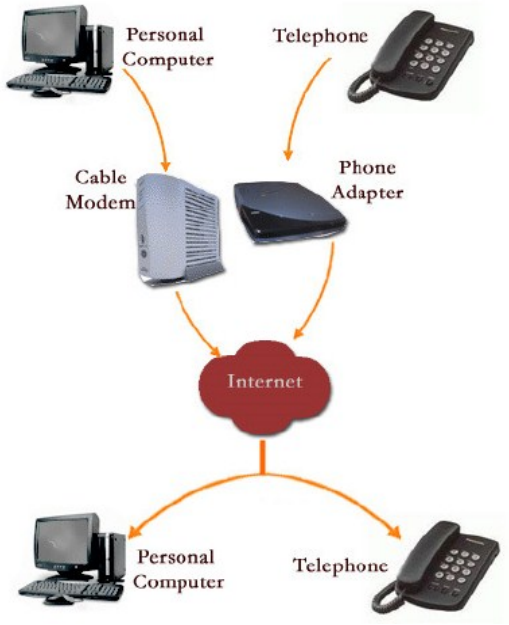

Figure 1: VoIP Layout [1]

2. VOIP BANDWIDTH OPTIMIZATION TECHNIQUES.

There are essentially three major techniques for improving bandwidth utilization in a VoIP network/service, which include VoIP packet header reduction and VoIP header compression and Silence suppression techniques.

2.1 Packet Header Reduction

This explains the process of reducing the size of the Packet header from its conventional size to a smaller size such that overall overhead is reduced for transmitted packets. This is achieved using varying techniques

1) Multiplexing

Multiplexing VoIP packets to increase the payload size considerably reduces overhead. [4]. The concept here presents a technique whereby the usual 40 bytes of VoIP packet header that combines with the usual payload (10 – 30 bytes) is reduced. The combination of the huge packet header with payload usually causes overhead. This is thus reduced by multiplexing related payloads in one header. By this, bandwidth is saved from wastage.

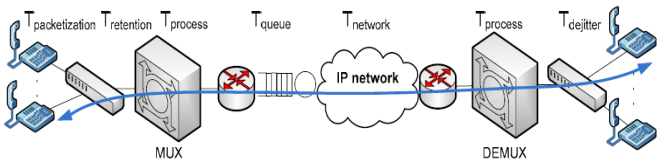

Figure 2 : Multiplexing technique [4]

2) RTP Multiplexing Techniques and Applications

Hoshi and his colleagues [5], proposed a Voice Multiplexing scheme that could be used to reduce the number of RTP packets transported over the IP Network. Their scheme combines voice packets from varying streams into a single UDP packet for transmission. Although this technique improves transmission efficiency, it does not handle the idea of compressing the RTP headers. [5]

Sze et al [6], presents a concept of combining packet multiplexing and header reduction that includes having a number of RTP packets in a single UDP packet. The RTP packets are compressed for which context-mapping tables are created in the multiplexer and de-multiplexer. This is to ensure that original packets are rebuilt or regenerated at the destination end. Through experimental test (simulation), the proposed system showed an effective utilization of Wide Area Network trunk and an increase in the number of supported simultaneous calls severally, as compared to the earlier (conventional) scheme.

A multiplexing scheme presented by [7] brings to light the concept of RTP multiplexing using translators and mixers. The mixer is responsible for collecting and adding up different or numerous voice streams, changes the data format and retransmits to the receiver. The translator re-encodes multiple packets into one. The point is that at the end, both send only a combined or translated RTP flow. Thus, the TC RTP defines the

combination of protocols for efficient management of voice packets. ECRTP header compression scheme compresses the IP, UDP and RTP headers into one new header. The PPP multiplexing is thus used for final transmission of packets through its tunnelling scheme.

3) Delta-Multiplexing

Delta-multiplexing techniques are another technique proposed by [8] which is used for improving band utilization. Noting inefficiency bandwidth utilization and network overloads as the two major problematic issues in a computer network circle. Delta-multiplexing provides solutions by combining header overhead reductions and payload size decrease [8].

The delta-multiplexing architecture consist of two entity; the multiplexer (Mux), located in the sender gateway and performs payloads size reduction and packet multiplexing. The second entity is the D-Multiplexer (DMux) which is located at the receiver gateway, and performs packet de-multiplexing, returning the payloads to its original size. Performance analysis on bandwidth efficiency in multiplexing 10 users in each stream showed that aggregate level of up to 68% - 72% of bandwidth is saved. This depicted an improved network performance in terms of network traffic, overload and packet congestion. Running VoIP packets over network are reduced and voice quality is enhanced. This puts Delta-multiplexing to be adaptable in SIP and H.323 systems conveniently [8].

- *Applications*
  Multiplexing techniques was applied on IP-Telephony Gateways (IP-TG) which connect PSTN/IPBX to IP networks [8]. The technique was noted to be applicable to IP-TG which connects cellular access networks with Mobile Switching Centre (MSC). This technique multiplexes multiple VoIP packets from different sources in single RTP header. A Test application showed that overhead reduced from 50% to 80%.
  Another application of multiplexing technique was shown in [9] which involve the multiplexing of voice packets generated by Session Initiation Protocols (SIP) applications. The scheme relied on the assumption that there are multiple SIP VoIP LANS connected via one SIP WAN VoIP Gateway (SWVG). This implied an increase in traffic on the SWVG gateway thus enhancing the multiplexing process. The concept is such that the sender SWVG gateway multiplexes the packets destined to the same destination, while the receiver SWVG gateway de-multiplexed the packets and dispatches them to their various destinations.

The end result is such that header overhead is reduced, thus saving bandwidth. Even more, is that the number of packets sent is reduced which also reduces overall overhead on network hips [9].

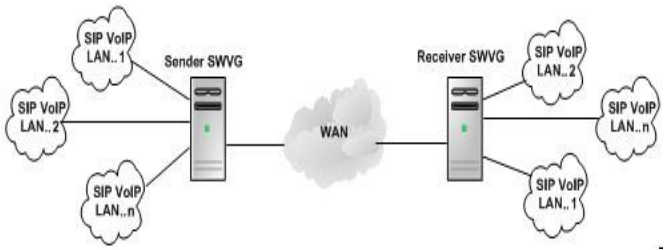

Figure 3 : SWVG Gateway Connect Multiple SIP VoIP LAN [9]

2.2 Voice / Packet Header Compression

This explains the phenomenon whereby VoIP packets are compressed from the conventional 40 bytes of size to a smaller byte size (2 bytes in the best case). The task of saving bandwidth is achieved using compression and decompression codecs. Codecs transform the voice signals from analog to digital, and further compresses the digital voice using digital compression algorithms. The compressed data are thus converted to frames (packet payload), which usually vary in sizes depending on the codec [9]. Data sizes are reduced greatly. There exist various compression codecs with varying compression rates as could be seen in table 1.

Table 1 : Most Used Codecs [9]

Codec	Frame size	Algorithm Delay	Compressed Rate(Bitrate) /kbps
G.723.1 (lr)	30	37.5	5.3
G.723.1 (hr)	30	37.5	6.3
G.729	10	15	8
G.729A	10	15	8
G.729D	10	15	6.4
G.729E	10	15	11.8
iLBC (lr)	30	30	13.33
iLBC (lr)	20	20	15.2
Speex	20	30	Various
GSM-FR	20	20	5
GSM-HR	20	20	24
GSM-EFR	20	20	18
AMR	20	25	Various

Header compression exploits the VoIP packet field characteristics of either being constant or increasing at constant ratio all through a call period. This techniques use compression/transmission protocols to effectively ensure that voice packet move from source to destination. This technique has been noted to most significantly improve bandwidth utilization.

1) Compressed Real-time Transport Protocol (CRTP) Compression

Given that IP telephony services utilize RTP protocols for their operations, it follows that the IP/UDP/RTP headers of each voice (data) packet can be compressed from the usual 40 or more bytes to a smaller size of up to 2 bytes. This is achievable using Compressed Real-time Transport Protocol, (CRTP). Although CRTP offers good compression mechanism, it has been found that packet loss rate after decompression at the receiver is too high [10]. It is thus clear that CRTP alone might not offer suitable solution to header compression in cellular & VoIP links; since packets are lost immensely. A high error-rate tolerance scheme that is at least efficient as CRTP is required. The ability to correctly produce and(or) reconstruct headers and packets is a very important requirement.

2) Robust Checksum-based header Compression (ROCCO) Compression

Robust Checksum-based header Compression (ROCCO) was developed to meet the weaknesses of CRTP as noted earlier. ROCCO was developed to be adaptable both to the characteristics of packet stream compression and of the link utilized for packet loss [11]. ROCCO is directed towards local de-compressor repair, which seeks to perform several reconstruction attempts to obtain the correct header. The outcome today is that ROCCO profiles for VoIP exist which are used to compress IP/UDP/RTP headers down to a minimum size of 1 to 2 bytes. ROCCO significantly reduces the adverse effects on header compression performance that are otherwise caused by high packet loss. Even more, its compression on headers is well better than CRTP and proffers a better security against errors on header generation [10].

3) CRTP / ROCCO UDP Lite Integration.

UDP Lite protocol proffers a supple way for applications to move decision of whether or not a packet should discarded as a result of bit errors from the transport layer to the application itself. Now a combination of this scheme with CRTP could deliver twice as many packets to the application as compared to classic UDP, while an integration of UDP lite with ROCCO will yield a greater member of packets to the application. This is seen in the simulation results as presented by [10].

2.3 Silence Suppression.

This technique is born out of the observed and statistically proven fact that 40 – 60% of phone calls are characterised by silence. A silence suppression technique uses an algorithm functions consisting of an encoder and a decoder. On the encoder side, voice activity detection (VAD) engine orders the input signal of a given frame either as active or inactive speech.

In the case it is classified active, a low bit-rate encoder is engaged before packing and transmission of the information. On the other hand, if the classification is inactive, a special packet called

silence insertion descriptor (SID) packet that retains several characteristic parameters of the background noise is created and sent to the far-end. A discontinuous transmission (DTX) algorithm is what determines the frequency of the SID packet transmission. On the decoder side, a signal representative of the silence in between conversation is created by a comfort noise generation (CNG) engine. This is meant to fill in the gaps that the original silence/noise should have occupied. At such, the transmission rate during the inactive periods of speech is reduced, and thus, the average transmission rate can be reduced [13].

With the adoption of ITU-T G.729A and B, numerous silence suppression algorithms have been developed with objective of enhancing performance. However, it is not simply the performance that is important but also the implementational complexity. There is thus high demand for a low-complexity silence suppression algorithm amongst other characteristics [13]

1) Desirable aspects of VAD algorithms

Basic requirements of Voice Activity Detection algorithms proposed by [14] and [15] as follows:

- **A Good Decision Rule:** A physical property of speech that can be exploited to give consistent and accurate judgement in classifying segments of the signal into silence or otherwise.
- **Adaptability to Background Noise:** Adapting to non-stationary background noise improves robustness, especially in wireless telephony where the user is moving.
- **Low Computational Complexity:** Internet telephony is a real-time application. Therefore the complexity of VAD algorithm must be low to suit real-time applications.

In their study of the Comparison of Voice Activity Detection Algorithms for VoIP, [15] presented a solution that lies in efficient VAD scheme used for VoIP systems. The time domain VAD algorithms were discovered to be computationally less complex, although the speech quality was reduced compared to frequency domain algorithms. The frequency domain algorithms is observed to maintain better immunity to low SNR compared to time domain algorithms. As a result, some VAD algorithm had been proposed, their test results of which expressed consistent superiority of the Comprehensive VAD scheme above all other algorithms. This scheme has additively offered good speech detection and noise immunity, although their still exist performance degradation under low SNR conditions. By and large the algorithms are still considered suitable real-time applications.

3. CONCLUSION AND FUTURE WORK.

Bandwidth utilization is indeed a key characteristic for improving quality of service in VoIP. And given the varying techniques for improving bandwidth utilization (Silence suppression, Packet Header Reduction and Compression) for which reduction of service cost is a principal objective, it is important that relative approaches are able to yield maximum result, where the least delay and timely delivery of packets are achieved. Silence suppression is relative productive although leave room for improvement. Packet header reduction using multiplexing is seen to reduce both bandwidth wastage and the number of packets. The delta-multiplexing technique which combines the approach of header reduction and payload reduction yield even greater results. However, the Packet (payload) compression technique yields the optimum result using codecs. Hence it could be considered are the most efficient of the techniques. An improvement could be geared to the implementation of codecs that would take faster bit rates, reduced delays for compression and conversion.

4. REFERENCES

[1] A Gkritsi, "Introduction To Voip Technology And Its Security Issues," in Proc *4th Annual Multimedia Systems, Electronics and Computer Science*, Southampton.UK,

2003, pp. 1-6.

[2] G Thomsen and Y Jani, "Internet telephony: Going like crazy," *IEEE SPECTRUM*, pp. 52-58, May 2000.

[3] H M Chong and H S Matthews, "Comparative analysis of traditional telephone and voice-over-Internet protocol (VoIP) systems," in Proc *2004 IEEE International Symposium on Electronics and the Environment*, 2004, pp. 106-111.

[4] J Saldana, J Murillo, J Fernandez-Navajas, J Ruiz-Mas, E V Navarro and J I Aznar, "Evaluation of Multiplexing and Buffer Policies Influence on VoIP Conversation Quality," in Proc *3rd IEEE International Workshop on Digital Entertainment, Networked Virtual Environments, and Creative Technology*, Zaragoza, Spain, 2011, pp. 378-382.

[5] T Hoshi, K Tanigawa and K Tsukada, "Proposal of a method of Voice Stream Multiplexing for IP Telephony systems," in *Proceedings of IWS*, 1999, pp. 182-188.

[6] H P Sze, S C. Liew, J Y B Lee, and D. C S Yip, "A Multiplexing Scheme for H.323 Voice-Over-IP Applications," *IEEE Journal on Selected Areas in Communication*, vol. 20, no. 7, pp. 1360-1368, September 2002.

[7] H Schulzrinne, S Casner, R Frederick, and V Jacobs. (2003, July) RFC 3550: "RTP: A Transport protocol for real-time applications". Document.

[8] M M Abu-Alhaj and M S Kolhar, "Delta-Multiplexing: A Novel Technique to Improve VoIP Bandwidth Utilization between VoIP Gateways," in *Proc 10th IEEE International Conference on Computer and Information Technology (CIT 2010)*, Malaysia, 2010, pp. 329-335.

[9] B Subbiah, S Sengodan, and J Rajahalme, "RTP payload multiplexing between IP telephony gateways," in *Global Telecommunications*, 1999, pp. 1121-1127.

[10] M AbuAlhaj, M S Kolhar, M Halaiyqah, and O Abouabdalla, "Multiplexing SIP Applications voice Packets between SWVG Gateways," in *Proc 2009 International Conference on Computer Engineering and Applications*, Singapore, 2009, pp. 226-230.

[11] L Larzon, H Hannu, and , L Jonsson, "Efficient Transport of Voice over IP over Cellular links," Luleii University of Technology, Sweden, Ericsson Research 2000.

[12] L Jonsson, M Degermark, H Hannu, and K Svanbro, "Robust Checksum-based header compression (ROCCO)," Sweden, Ericsson Research 2000.

[13] W Chu, M O Ahmad, and M.N. S Swamy, "Modified silence suppression algorithms and their performance tests," in *Symposium on Circuits and Systems, 2005. 48th Midwest*, vol. 1, Canada, 2005, pp. 436 - 439.

[14] A Sangwan, M C Chiranth, H S Jamadagni, R Sah, R Venkatesh Prasad, V Gaurav, "VAD techniques for real-time speech transmission on the Internet ," in *Proc 5th IEEE International Conference on High Speed Networks and Multimedia Communications* , 2002, pp. 46-50.

[15] R V Prasad, A Sangwan, H S Jamadagni, M C Chiranth, and R Sah., "Comparison of voice activity detection algorithms for VoIP ," in *Proc Seventh International Symposium on Computers and Communications, 2002. Proceedings. ISCC 2002.* , 2002, pp. 530-535.