Packet Loss Concealment in a Secure Voice over IP Environment

Christopher M. White, Keith A. Teague, and Edward J. Daniel

Speech and Audio Communications Laboratory
Oklahoma State University
Stillwater, OK 74078

ABSTRACT

Quality of Service (QoS) in a Narrowband Secure Voice over data over IP (SVoIP) system is impacted by changes in network behavior, particularly packet loss. Several methods exist for mitigating these effects. This paper analyzes the performance of packet loss concealment using forward error correction, frame replication, adaptive playout buffering, and dynamic buffer selection in a SVoIP system. These methods are shown to improve to varying degrees overall QoS.

I. INTRODUCTION

Packet loss occurs as a result of buffer overflow(s) at network nodes due to heavy loads or bit errors incurred by packets during transit. There have been numerous studies on Internet packet loss statistics including ([1] – [5]). Packet loss is known to have some correlation with packet size, time of day (congestion), and network delay [6]. According to [7], packet loss of 10% is unexceptional, and losses of up to 40% are possible on the Internet [5]. Internet packet loss is bursty and correlated meaning that if packet n is lost then there is high probability that packet n+1 will be lost [4] [1] [6]. Borella et al. [4] found that packets lost in bursts account for the majority of the packets lost over a long period of time; however, overall packet loss gaps are usually close to one or two packets [1].

In a VoIP (Voice over data over Internet Protocol) application packet loss corresponds to a loss of voice data (seen at playback). This can also result from packets arriving at the destination with varying delay (between packets) referred to as 'jitter'. Under ideal network situations, packets will be received at their destination at a constant inter-arrival rate, equal to the send rate. However, as packets are transmitted on packet networks they may experience different delays as they travel from source to destination. When the inter-arrival delay is too large it leads to starvation of the audio playback system. If the inter-arrival delay between packets is too small it leads to a buffer overrun where the application can not service the packet in time.

There are several techniques for recovery of lost packets in voice applications including [8]. Insertion techniques such as frame replacement have low complexity and add no delay. If a packet loss is detected, silence, noise, or adjacent frames are inserted in place of the lost packet. This study details receiver based packet loss recovery techniques using adjacent frame replication, forward error correction, and adaptive playout buffering with network identification for narrowband SVoIP systems over heterogeneous networks.

II. PACKET LOSS AND SECURE SYNCHRONIZATION

Secure signaling adds an additional constraint to Quality of Service (QoS) for a VoIP system. In a typical VoIP application, a lost packet results in an equivalent amount of lost speech. Secure signaling in a SVoIP system adds another "layer" to the application due to its framing requirements and cryptographic synchronization. For example, consider a SVoIP system signaling plan which includes a super/sub frame structure (of length n frames) with a cryptographic synchronization frame (sync-frame) following by n-1 encrypted and encoded low rate voice frames. This SVoIP system contains a distributed key counter mode encryption scheme with synchronization information distributed across m sync-frames. The underlying VoIP system fragments each superframe into several packets for transmission.

Packet loss in a SVoIP environment can significantly degrade performance since at least one and perhaps as many as m sync-frames frames must be received before synchronization can be regained. However, for a small sequential packet loss of one to two packets the remainder of the packets can be decrypted given the index of the encryption key (based on the location in the superframe of the lost packet). Nevertheless, any packet loss results in lost voice frames, and the loss is propagated to the low rate voice codec.

Loss of synchronization, as a result of the loss of a sync-frame, severely impacts the SVoIP system and deserves special attention for packet loss recovery. The Sync Loss Model is a probabilistic model that provides a means to study, evaluate, and improve SVoIP behavior in the presence of packet loss, particularly as it relates to loss of synchronization. The loss of a sync-frame (containing encryption counter information) prevents the decryption of subsequent data resulting in voice loss until synchronization can be regained, up to m*(n-1) frames.

III. SETUP, NETWORK AND SYNC LOSS MODELS

The remainder of this paper is devoted to packet loss recovery for a hypothetical Secure Voice over data over IP system, and its recovery of lost packets due to Internet models. Encryption is based on the American Encryption

Standard (AES) with a counter mode implementation. This system considers the MELP voice codec (e.g. MIL-STD-3005 [9]) where n = 24 and m=3. To evaluate the quality of MELP in the presence of packet loss the Gilbert model is used to simulate packet loss on an IP network and a Sync Loss Model is used to model the impact of encrypted loss. These packets containing MELP frames are deleted (dropped) according to the probabilistic Gilbert model. The p and q values used for each simulation are calculated based on the steady state probabilities in Equation 1:

$$p = \frac{ulp(1-clp)}{1-ulp}, \quad q = 1 - clp \qquad (1)$$

where ulp is the unconditional loss probability and clp is the conditional loss probability.

The Sync Loss Model has 5 states, corresponding to a loss, a loss of a sync frame, and a state for each sync frame that could have been lost (our example has 3 sync frames). In the above Gilbert model, the simulation packets that are lost are propagated to the Sync Loss Model that determines if the lost packet contains a sync-frame (state S2) or a MELP frame (state S1). If the loss is a sync-frame (from state S1 to S2) then the model determines if it is sync-frame 1, 2, or 3 (states S3, S4, or S5 respectively). Equation 2 shows the probability transition matrix for the model.

Sync-frames contain the encryption counter information. Due to narrow bandwidth constraints imposed by the SVoIP system, the counter value is spread across m (m=3) sync-frames with the start of the counter (partially) being placed in the first sync-frame (SF1). The second sync-frame (SF2) contains partial counter values that allow the counter to pick up where the last counter left off and continue to decrypt the next superframe data. Finally, the third sync-frame (SF3) contains values of the counter to continue from the last superframe (SF2) and decrypt the final superframe in the three superframe set. A new counter value is provided in the next SF1, and the process continues. If SF1 is lost the SVoIP system must wait for three superframes until the next SF1 is received with a new counter value, thus achieving encryption synchronization. If SF1 is received and SF2 is lost then encryption synchronization is achieved when the next SF1 is received, two superframes later. Similarly, if SF3 is lost then encryption synchronization is regained at the next superframe, SF1. The transitional probability matrix for the Sync Loss Model is provided below in Equation 2

$$P_{sync} = \begin{bmatrix} 1-\lambda_2 & \lambda_2 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & \lambda_4 & \lambda_5 \\ \mu_3 & 0 & 0 & 0 & 0 \\ \mu_4 & 0 & 0 & 0 & 0 \\ \mu_5 & 0 & 0 & 0 & 0 \end{bmatrix}. \qquad (2)$$

The probability that a lost frame is a sync-frame given the sample space is a superframe is λ_2 = pSF. If a frame is lost there is a 1 in n (n=24) probability the lost frame will be a sync-frame, assuming for the time being that a packet contains only one frame of data. Therefore pSF = 1/n, where n is the number of frames in a superframe. The counter values are spread across 3 consecutive sync-frames, therefore if a sync-frame is lost there is a 1 in 3 chance that SF1(λ_3 = pSF1), SF2 (λ_4 = pSF2), or SF3 (λ_5 = pSF3) will be the lost frame (pSF1 = pSF2 = pSM3 = 1/3). The probability of a MELP frame loss is 1-λ_2 = 1-pSF, and $\mu_3 = \mu_4 = \mu_5 = 1$. The loss model can be used to determine the effect on QoS of lost frames.

IV. QUALITY MEASURE

Perceptual Evaluation of Speech Quality (PSEQ) [10] has been adopted by the ITU-T as an objective measure for evaluating narrowband speech. It considers input filtering, variable delays, coding distortions, and channel errors, and produces an objective PESQ score as well as a prediction of the subjective Mean Opinion Score (MOS) which is traditionally obtained by subjective listening tests.

PESQ uses a perceptual model to compare the original signal and the degraded output signal of a system under test to measure one-way quality. PESQ can be implemented non-intrusively comparing a reference signal with a system output degraded signal. The original signal and the degraded signal are level aligned to a normal listening level used in subjective tests and filtered with an input filter that models telephone headsets. To compensate for variations in delay imposed on the signal by the network, the signals are time aligned. The delay results from the time alignment are used in an auditory transform to compensate for delay changes in both silence and active speech. A perceptual model is used to compare the reference and time aligned distorted signals. The comparison, resulting in a disturbance measure, gives an indication of the absolute audible error as well as audible errors significantly louder than the reference. These disturbance parameters are converted to a PESQ score and mapped to a predicted MOS score. The PESQ score ranges from -1 to 4.5 and the MOS-like scores are between 1 and 4.5, with 4.5 corresponding to very good quality in both cases.

V. SYSTEM QUALITY ENHANCEMENTS

5.1 (Un) Secure MELP Distortion Analysis

As an initial basis for comparison, consider an unencrypted VoIP system using MELP at 2400 bps as the codec. Frame replication is considered as a baseline loss recovery technique from which to analyze the impact of encryption and additional techniques. This experiment uses a file containing 2 minutes and 8 seconds of digitized speech corresponding to 3 male and 3 female talkers in a quiet environment. The file is encoded using the MELP encoder (producing approximately 5,714 frames). The

output (parameterized) MELP frames are grouped to form packets; in this case 2 MELP frames are included per packet. Voice Activity Detection (VAD) is not used in this simulation.

Frame replication replaces the missing frame with an adjacent frame. The resulting speech file, with errors introduced by the Gilbert model, is compared to the original file synthesized by MELP.

The potential for sync loss increases as network loss rates increase with an accompanying impact on quality (performance). Simulation over multiple network conditions (parameter values) and resulting loss rates were recorded for each simulation along with the quality score and spectral distortion. There were 103 simulation scenarios run for Clear MELP and Secure MELP. The simulated output quality evaluation is shown in Figure 1.

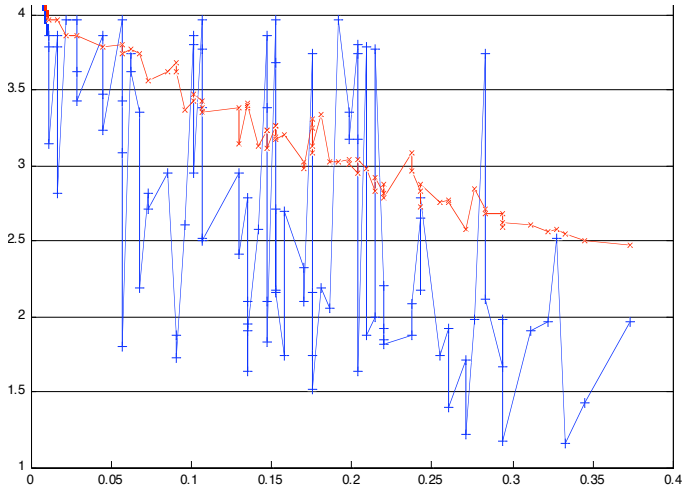

Figure 1: PESQ Score vs. Loss Rate for Clear MELP (.) and Secure MELP (+)

Figure 1 shows a decrease of PESQ score for Clear MELP down to 2.5 at a loss rate of 0.35. However, the PESQ score of Secure MELP falls as low as 1.2 at a loss rate of 0.33. Several areas have PESQ quality values for Secure MELP swing from 1.5 to 3.7 for a particular loss rate. If a sync-frame is lost, a large number of consecutive voice frames is lost, therefore, the loss of a sync-frame at any given time in the simulation causes the quality to drop very quickly. The minimum quality values show that Secure MELP has difficulty maintaining synchronization, with quality decreasing rapidly from good to poor.

5.2 SF Forward Error Correction (FEC) Analysis

Media-specific FEC is an implementation that applies to secure sync-frame recovery. It involves sending multiple copies of each frame (not whole packets) in successive packets. FEC can be used to prevent the low PESQ scores by avoiding synchronization loss. Depending on the number of frames included in a packet, the sync-frame can be included in every packet, every other packet, every 5 packets, every 10 packets, etc., within a superframe.

Bolot et al. [11] introduced an adaptive FEC and rate control algorithm for real-time audio data over the Internet. This scheme adapts FEC to varying loss conditions in the network where varying amounts of redundancy are determined to minimize loss rate and conserve bandwidth. Mean loss rate feedback from the receiver system is used to determine the amount of redundancy and combination needed at the transmitter system to achieve a loss rate closest to a predetermined target loss rate at the receiver.

The network loss process is modeled as a 2-state Gilbert model where the mean (steady state) loss rate is $\pi_1 = \frac{p}{p+q}$ without redundancy. Bolot et al., show that by adding a single redundant packet, the loss rate after recovery becomes $\pi_2 = \frac{p(1-q)}{p+q}$. A single redundant packet assumes packet n contains redundant information about packet n-1 only, and data is only lost in a burst loss greater than 1 packet. This showed that by adding a single redundant frame the loss rate is reduced proportional to (1-q), i.e. if q = 0.70 that means a 70% loss reduction.

Although this technique is used for FEC of audio data, it has potential for use with secure sync-frames. This adaptive FEC technique sends redundant copies of sync-frames to improve synchronization loss rate while minimizing bandwidth. However, a slight modification to the above loss rate equations must be made. The network loss process above is assumed Gilbert, but the synchronization loss process based on the Sync Loss Model is Bernoulli, where $p = \lambda_2 /(\lambda_1 + \lambda_2)$, and $\lambda_1 + \lambda_2 = 1$. So, loss rate without redundancy is therefore equal to p and the loss rate with one redundant packet is p(1-q).

Redundancy of up to 3 redundant sync-frames per superframe is considered and tested in this implementation. The potential for burst losses in the Internet and wireless networks warrants spacing the FEC sync-frames for greater potential recovery due to burst losses. To minimize the susceptibility of loss due to consecutive burst losses, FEC frames can be spaced evenly throughout the superframe. For a superframe size of 24, 1 redundant sync-frame can be sent in the middle of the superframe between frames 12 and 13. For the case of 2 redundant frames, FEC frames can be sent to effectively divide the superframe into 1/3rds, after frame 8 and again after frame 16. Likewise, 3 redundant frames can be sent to effectively divide the superframe into 1/4ths, after frames 6, 12, and 18.

The mean synchronization loss rate from the receiver system is fed back periodically. The number of redundant packets and combination is chosen based on the minimum redundant packets containing sync-frames required to achieve the lowest average synchronization loss rate.

The maximum bandwidth increase with 2400 bps Secure MELP is 12% or 2728 bps for 3 redundant sync-

frames which is well below modern communication bandwidth standards, particularly with IP networks. Even with the addition of overhead due to IP packet headers this adaptive FEC scheme will satisfy bandwidth requirements over narrow bandwidth IP compatible networks such as IS-95 CDMA and GPRS cellular data which have channel bandwidths as low as 9600 bps and 9000 bps, respectively.

In Figure 2 the results of redundant packets with consecutive and burst spacing is shown for the same simulation. The burst spacing of redundant packets reduces the loss rates 5 times more than with consecutive spacing. Figure 2 shows robust burst spacing to burst loss for the adaptive FEC algorithm with burst spacing.

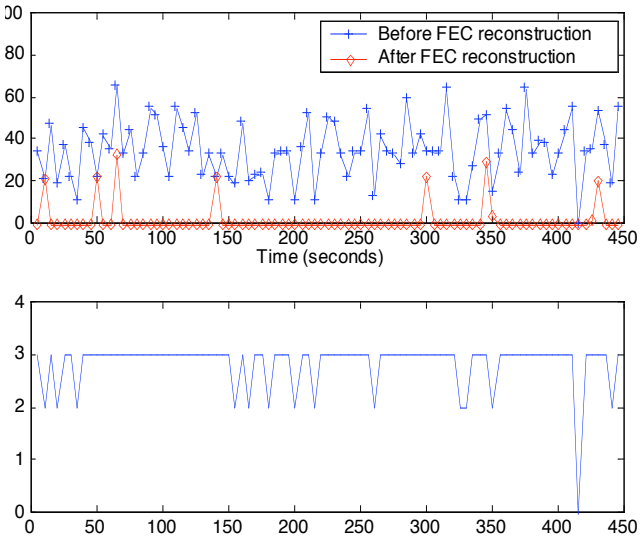

Figure 2: Adaptive FEC of SF with Spacing of FEC Frames to Avoid Burst Losses (Above: Time vs. Loss Rate, Below: Time vs. # Redundant SF)

5.3 Adaptive Buffering and Selection with Network Identification

The necessity for loss concealment schemes such as FEC has much to do with the network configuration under test. Identification of significant network characteristics can improve the types and parameters of such schemes. To illustrate this impact consider the characteristics of a IS-95 CDMA cellular data network shown in Figure 3. Figure 3 highlights fixed and adaptive buffering schemes for a mean jitter 81 ms, and standard deviation of 208 using Algorithm 4 in Ramjee et al. [12] It shows the jitter output values and the corresponding buffer length for each jitter value. Simulated packets with jitter values greater than the buffer length are considered late and are discarded.

Figure 3 presents an extreme case of jitter delay; under these jitter conditions real-time communication would be difficult. Large lengths of the adaptive buffer enable the conversation to get through in contrast to the fixed buffer which loses the majority of the packets to late arrivals based on buffer length. Discarding packets which arrive after twice the send rate, the total late loss and loss rate for the fixed buffer is 107 out of 354 total packets and 30.2%, respectively. The adaptive buffer adjusts to the jitter delay trends in high burst delay areas yielding no late losses for the entire trace. To match the packet loss savings of the adaptive buffer, the equivalent fixed buffer length would be approximately 800 ms. A fixed buffer of 800 ms would give favorable late loss results, but the delay savings of the adaptive buffer would go unmatched. Consequently, the additional delay imposed on the overall system would be too great for practical real-time communication.

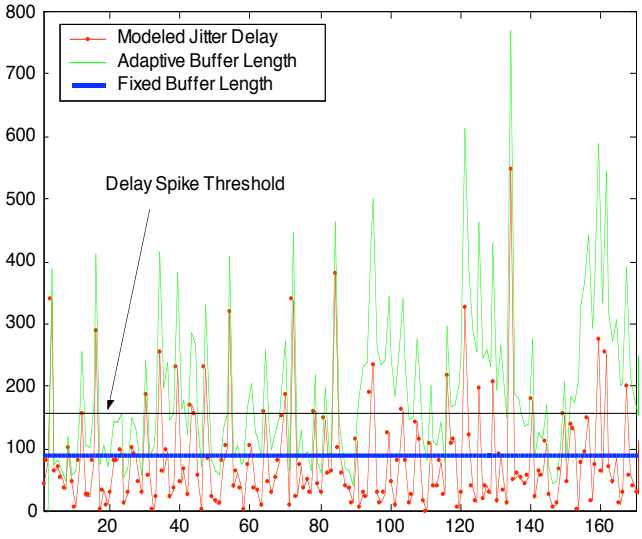

Figure 3 Inter-arrival Time (Jitter Delay) vs. Jitter Buffer Length

This network exhibits characteristic behavior which when known can be useful for adaptive jitter buffering and FEC. In order to select an appropriate jitter buffer (static, Algorithm 4 [12], etc.) or number of redundant SF the dominant leg in the mixed network must first be identified. The Kolmogorov-Smirnov (KS) test is used for network identification through comparison between delay data and data from a base knowledge set. This test is chosen as it is well known for accurately comparing two cumulative distribution functions (CDFs) to show general differences between two distributions [13][14]. The KS test uses the maximum vertical distance between two CDFs for statistical comparison. CDF analysis allows for a general description of differences in two data sets. Any significant difference between the data and library data sets will be present in their distributions and correspondingly in their CDFs. Figure 4 illustrates the CDF comparison between the LAN data and IS-95 CDMA data.

The KS computes a distance as in Equation 3:

$$D = \max\left|F1(x) - F2(x)\right| \quad (3)$$

where F1(x) is the CDF of the collected data and F2(x) is the data for a library set. This implies D is equal to the maximum difference in the two CDFs. Figure 4 shows the

CDF of a LAN data set, a CDMA data set, and an unknown mixed network configuration data set. It depicts an example network classification.

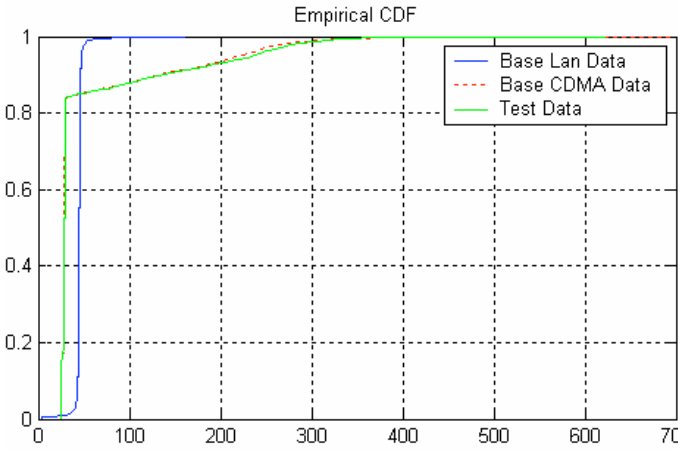

Figure 4: CDF plots with test data

This experiment tested the first 100 packets of an incoming data stream. This stream was not from the configuration of any data set in the library. However, it is clear from the CDF that an IS-95 CDMA data link was present somewhere in the mixed network configuration. This link dominated the network performance for the communications channel. This result, corresponding to a small value of D, facilitated its classification in accordance with an IS-95 CDMA data set and its corresponding jitter buffer (Algorithm 4, shown in Figure 3). Every part of the test data curve was within five percent of the IS-95 CDMA library data curve indicating a statistical similarity as shown in Figure 4. Such a comparison passes the test as both sets contain similar characteristics over the time period [13]. This result can be used to support network dependent adaptive jitter buffering.

VI. CONCLUSION

Narrowband Secure Voice over data over IP systems present particular implementation challenges. Packet loss affects a VoIP system with loss in voice data and reduction of QoS. Packet loss in a SVoIP system has the added possibility of synchronization loss which can quickly degrade QoS to an unacceptable level. Increases in packet loss rates (ulp) for the Internet enhance the possibility of synchronization loss, especially due to burst errors.

It is well known and verified in this study that VoIP data streams can achieve quality improvement with packet loss recovery techniques including frame insertion using adjacent frame replication. Adaptive Forward Error Correction with rate control feedback is shown to successfully reduce synchronization loss through the recovery of sync-frames. Synchronization loss is shown to be reduced at high loss rates, >20%, by a significant amount when redundant sync-frames are evenly spaced. Adaptively sending redundant sync-frames conserves bandwidth, with a satisfactory voice quality maintained with a maximum bandwidth increase as small as 12%. Adaptive jitter buffering reduces packet loss due to buffer over/underrun when compared to static buffering, while minimizing overall system delay. Network identification is shown to correctly classify dominant segments in a heterogeneous or unknown network within a library of known performance characteristics. This classification can improve QoS with the aid of adaptive jitter buffering specifically optimized for a particular network performance characteristic. When used together these loss recovery techniques can be used to improve overall QoS in a SVoIP system and enhance system performance at playout.

VII. REFERENCES

[1] J. C. Bolot, "Characterizing End-to-End Packet Delay and Loss in the Internet", *Journal of High Speed Networks*, pp. 305-323, Vol. 2, 1993.

[2] D. Su, J. Srivastava, and J-H Yao, "Investigating Factors Influencing QoS of Internet Phone" *IEEE International Conference on Multimedia Computing and Systems*, pp. 308-313, 1999.

[3] M. Yajnik, S. Moon, J. Kurose, and D. Towsley, "Measurement and Modeling of the Temporal Dependence in Packet Loss," *Proceedings of IEEE Conference of the IEEE Computer and Communications Societies INFOCOM '99*, pp. 345-352, June 1999.

[4] M. Borella, D. Swider, S. Uludag, and G. Brewster, "Internet Packet Loss: Measurement and Implication for End-to-End QoS," *Proc. IEEE ICPP Workshop '98,* pp. 3-12, 1998.

[5] J. C. Bolot, A. Vega-Garcia, "Control Mechanisms for Packet Audio in the Internet", *Proceedings IEEE INFOCOM '96*, Vol. 1, pp 232-239.

[6] N. F. Maxemchuk and S. Lo, "Measurement and Interpretation of Voice Traffic on the Internet", *IEEE International Conference on Communications ICC'97*, Vol. 1, pp. 500-507, Aug. 1997.

[7] M. Khansari, V. Bhaskaran, "A low-complexity error-resilient H.263 coder", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp 2737-2740, 1997.

[8] C. Peerkins, O. Hodson, and V. Hardman, "A Survey of Packet Loss Recovery Techniques for Streaming Audio," *IEEE Network*, pp. 40-48, Sept. 1998.

[9] L. Supplee, R. Cohn, and J. Collura, "MELP: The New Federal Standard at 2400bps," IEEE International Conference on Acoustic, Speech, and Signal Processing ICASSP'97, pp. 1591-4, 1997.

[10] ITU-T Recommendation P.862. "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.

[11] J. C. Bolot, S. Fosse-Parisis, D. Towsley, "Adaptive FEC-Based Error Control for Internet Telephony", *Proceedings of IEEE 8th Annual Joint Conference of the Computer and Communications Societies INFOCOM '99*, Vol. 3, pp. 1453-1460, March 1999.

[12] R. Ramjee, J. Kurose, D. Towsley, and H. Schulzrinne, "Adaptive Playout Mechanisms for Packetized Audio Applications in Wide-Area Networks," *Proceedings IEEE Networking for Global Communications INFOCOM '94*, Vol. 2, pp. 680-688, 1994.

[13] M. Sandford et. all, "Neural approach to detecting communication network events," *IEE Proc. Commum.,* Vol 149, No. 5, Oct. 2002

[14] H.R. Neave, and P.L. Worthington, "Distribution-free tests," (Unwin Hyman, 1988)