Journal of Network and Computer Applications (1998) **21,** 203–218
Article No. ma980073

Enhancing the quality of Internet voice communication for Internet telephony systems

K. V. Chin, S. C. Hui and S. Foo

School of Applied Science, Nanyang Technological University, Nanyang Avenue, Singapore 639798. E-mail: {asschui|assfoo}@ntu.edu.sg

Internet real time voice communication involves transmitting digitized voice signals which can be lost in the packet-switched network environment of TCP/IP, thereby causing intermittent voice losses and quality degradations. Various methods such as silence substitution, waveform substitution, sample interpolation, Xor mechanism, embedded speech coding, and the combined rate and control mechanism have been proposed to enhance voice delivery and to minimize the quality impact caused by these losses. This paper proposes a quality-based dynamic voice recovery mechanism that combines network transmission control and voice recovery to deliver voice signals with optimal intelligibility and quality. This is accomplished by considering the subjective rating of different codecs that are used in the coding and transmission of digital audio and network packet loss conditions. The dynamic mechanism results in voice delivery that, at minimum, satisfies voice intelligibility while tolerating moderate packet loss caused by network congestion. This mechanism has been successfully incorporated into the Internet Telephone Software System developed at the School of Applied Science, Nanyang Technological University.

1. Introduction

Internet telephony systems are basically synchronous distributed systems whereby two users who are physically separated are able to carry out real time voice communication over the Internet. Currently, a wide range of academic and commercial Internet telephony systems such as IPhone [1], NetMeeting [2] and NeVoT [3] have been developed. This growing interest is mainly motivated from huge potential cost savings by making it possible to make transcontinental telephone calls at the prices of local telephone calls plus nominal standard Internet connectivity charges. However, the quality of communication of the existing Internet telephony systems is not comparable to those offered by telephone companies.

The inferior quality is primarily due to the high transmission delay and packet loss of the Internet environment which is characteristic of a packet-switched network without resource reservation mechanisms. The most frequent cause of this packet loss is an overwhelming number of packets arriving at intermediate nodes resulting in network congestion and thereby causing nodes to discard packets which it could not service [4]. Transmission control protocol/Internet protocol (TCP/IP) [5] is the communication protocol of the packet-switched global Internet. It provides two kinds of network services to application processes: transport control protocol (TCP), which is a connection-oriented protocol with guaranteed delivery of data; and user datagram protocol (UDP) which is a connectionless-oriented protocol with no guarantee of arrival of data.

1084–8045/98/030203 + 16 $30.00/0

The TCP/IP protocol applies well for non-temporal data transmission but does not have any provision to support the real time nature of audio data. The on-time delivery of data is generally dependent on network performance. This is in turn largely related to the kind of network technologies implemented by various systems which unfortunately are not a controllable factor. Under varying network load conditions, the audio data packets will suffer varying degrees of delay. The variance in delay produces jitters that are undesirable for real time services as a suitable amount of delays [6] have to be introduced at the receiver to accommodate the late arrivals of packets.

To resolve these problems, two approaches are possible. The first is to develop a new real time protocol, while the second is to modify and extend the existing TCP/IP to support real time communications. The first approach requires the upgrade of the network to support quality of services such as the need for admission control, scheduling and reservation mechanisms. However, this is a long-term solution that will take time to implement. Currently, most existing telephony systems have taken the second approach to extend the TCP/IP protocol with new mechanisms and descriptors to deliver real time voice communication. However, this approach will at best simulate real time, but does not guarantee real time delivery due to the underlying nature of the existing protocol. Different mechanisms have been developed for this approach to handle the delay jitters and data packet loss problems. The play-out time of arriving audio packets can be adjusted at the destination to minimize the impact of delay jitters using a buffering mechanism [3]. Various voice recovery methods such as silence substitution, waveform substitution, sample interpolation, embedded speech coding [7], Xor mechanism [8], and combined rate and control mechanism [9] have been proposed to eliminate or minimize the impact of packet loss.

Currently, most voice recovery mechanisms only aim at minimizing packet loss without considering how best the intelligibility and quality of voice signals can be delivered. While it is important to seek high speech continuity and clarity, it is also necessary to maintain a balance between the additional delay and overhead incurred, and the quality of the received voice signals that will at least guarantee its intelligibility. In addition, dynamic transmission control should be applied to adjust the bandwidth usage dynamically when different levels of network congestion are encountered in order to avoid network congestion collapse [10].

In this research, a quality-based dynamic voice recovery mechanism is proposed and developed to enhance the quality and reliability of real time voice communication for Internet telephony systems. The dynamic recovery mechanism integrates the dynamic transmission control [11] with voice recovery [7,9] using a quality-based measurement. It minimizes packet loss by controlling the transmission rate dynamically from the source based on the network congestion condition with a quality-based voice recovery at the destination. The quality of voice signals delivered is measured based on the subjective ratings of different voice codecs. Multiple redundancies are used to enable better reception and recovery of voice signals during congested network conditions. The dynamic recovery mechanism has been incorporated into the Internet Telephone Software System (ITSS) [12] that was developed at the School of Applied Science, Nanyang Technological University.

This paper is organized as follows. An overview of the Internet telephony system is first given. This is followed by a discussion of the proposed quality-based dynamic voice

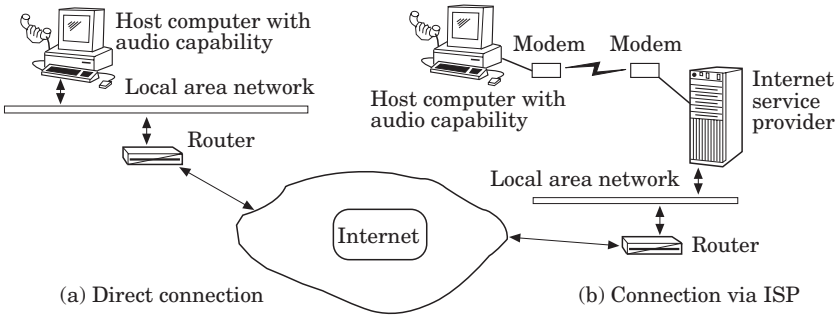


Figure 1. The Internet telephony system.

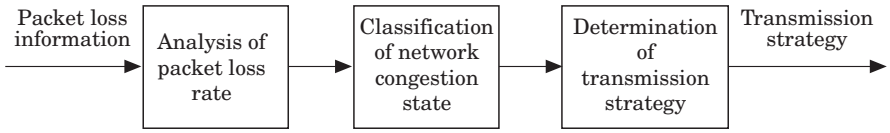


Figure 2. Quality-based dynamic voice recovery mechanism.

recovery mechanism that is supported by some experimental results. The architecture of the Internet Telephone Software System that incorporates the proposed mechanism is then presented. Finally, the conclusion and future work are given.

2. Internet telephony system

Figure 1 shows the basic components of an Internet telephony system. Two host computers acting as a caller and as a recipient are required. In using the standard TCP/IP, each host computer is identified by a unique IP address, Two modes of connection to the Internet are possible. Users can connect to the Internet either directly or via an Internet service provider (ISP). The host computer can either be a workstation or a personal computer with sufficient computation power and audio capabilities. The telephony system that resides on each host computer facilitates the real time voice communication across the Internet.

In the basic communication process, the caller's telephone system will acquire the real time voice data through an audio input device and convert the analogue signals into a digitized form which is then compressed and optionally encrypted before being transmitted to the recipient through the Internet using the TCP/IP protocol. Compression is necessary to reduce the bandwidth requirement of the voice data. At the recipient's end, the Internet telephony system carries out the reverse process. Incoming data is first decrypted, decompressed and played back in real time on the audio output device of recipient's computer. Communication can either be half or full duplex although the second form is desired since it emulates the conventional telephone system.

3. Quality-based dynamic voice recovery

Figure 2 shows the quality-based dynamic voice recovery mechanism. It first analyses the packet loss data received from incoming receiver reports and uses a low-pass filter

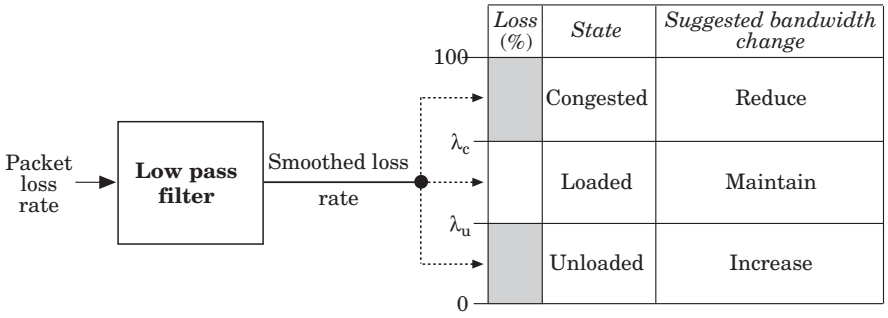


Figure 3. Classification of the network congestion state.

to smooth the packet lost rate statistics. The smoothed loss rate is used for the determination and classification of the network congestion state. Three network states, namely unloaded, loaded, and congested, can be defined according to some pre-defined thresholds. Depending on the network congestion state, the smoothed loss rate is then used for measuring the quality of the expected voice signals for different transmission strategies. Finally, the transmission strategy with the best quality rating is selected for transmitting voice signals.

3.1 *Analysis of packet loss rate*

When voice packets are transmitted to the destination, the receiver at the destination calculates the number of paskets lost from the number of packets received according to their sequence numbers for a given period of time. The number of packets received is then reported in the receiver report and transmitted to the source. Based on these loss statistics, the source determines the loss rate. The source then uses a low-pass filter to smooth out the loss rate. The new smoothed loss rate λ_{new} is computed as follows:

$$\lambda_{new} = (1-\alpha)\lambda_{old} + \alpha b \tag{1}$$

where b is the new loss rate and α is a constant between 0 and 1 which is used to indicate the influence of the new loss rate on the final smoothed loss rate. λ_{old} is the old smoothed loss rate. A moderate value of 0·3 for α has been used as discussed in [11].

3.2 *Classification of network congestion state*

The new smoothed loss rate (λ_{new}) is then used to measure and determine the network congestion state. As shown in Fig. 3, three network states are defined based on two pre-defined threshold values: unloaded, loaded or congested. They are defined according to the user's perception towards different degrees of voice packet loss. The upper threshold λ_c gives the upper limit where voice quality will be unacceptable if this threshold is exceeded. The lower threshold λ_u is defined such that a packet loss rate below this limit will give good voice quality. For a loss rate between these two thresholds, it is considered that acceptable voice quality can be delivered. In addition, the network congestion state classification also suggests an action on whether to increase, maintain

Table 1. *MOS ratings for voice coding algorithms*

CODEC	MOS
PCM (G.711)	4·3
ADPCM (G.721)	4·1
LD-CELP	4·0
GSM	3·47
CELP	3·2
LPC-10e	2·3

or reduce the current bandwidth according to the unloaded, loaded or congested state accordingly. In this case, the linear increase and multiplicative decrease [11] are used for the proportion of bandwidth change as given below:

Linear increase: new bandwidth = old bandwidth + μ

Multiplicative decrease: new bandwidth = old bandwidth*v

The value of 0·875 is used for v and 2·4 kbps is used for μ for voice transmission. The values of λ_c and λ_u will be determined based on experimental results.

3.3 *Determination of the transmission strategy*

3.3.1 *Quality ratings.* Various measurement methods, both objective and subjective, have been used to measure the voice coding algorithms. Objective measurement [13] refers to the mathematical comparisons of waveforms by giving an 'undisputed value' that is supposed to correspond to the quality as perceived by users. Although a myriad of objective measurements such as signal-to-noise ratio (SNR) and segmental signal-to-noise ratio (SSNR) have their scientific basis, it is subjected to debate as there is no unambiguous definition of 'quality'. Such methods are inappropriate for measuring voice signals produced by various coding techniques [14]. On the other hand, subjective measurement [13] is based on the feedback gathered from a group of listeners. Quality tests such as mean opinion score (MOS) and diagnostic acceptability measure (DAM) determine the quality of voice signals based on the listener's subjective perceptions. It measures the naturalness in addition to spoken word recognition.

As voice data streams are made up of voice signals digitized using various voice coding algorithms, the quality of the voice signals can be measured using MOS which is applicable to various voice coding techniques. A list of MOS scores for several different voice coding algorithms is given in Table 1.

3.3.2 *Quality measurements.* As packet loss results in momentary loss of voice, it is possible to estimate the final voice quality in terms of the function of the expected quality of the voice and the proportion of audio packets arriving at the destination. This accounts for the quality degradation for the period of time when voice signals are not available. Therefore, assuming that subjective measurement of silence is null, then

$$Q(\textit{voice signals received}) = (1 - L)*Q \text{ (voice signals sent)} \quad (2)$$

where the function Q represents the quality rating using MOS and L is the network packet loss rate.

In order to minimize data packet loss that can occur during the process of transmitting voice signals to the destination, redundancy transmission can be used. The redundant voice segments can be transmitted at different intervals relative to the primary transmissions. However, a total elimination of the problem may not be possible if consecutive packet loss occurs. In this case, redundant voice segments in the packets following the primary voice segments will also be lost. Hence, transmitting the redundant voice segments multiple times can further increase the resilient to the effects of packet loss.

When multiple redundancies transmissions and simple random packet loss characteristics are considered, the quality rating (Q) for the voice signals received can be derived from eqn (2) as follows:

$$\mathrm{Q} = (1-L)*\mathrm{P} + L^{1}*(1-L)*\mathrm{R}_1 + L^{2}*(1-L)*\mathrm{R}_2 + \ldots + L^{\mathrm{n}}*(1-L)*\mathrm{R}_n \tag{3}$$

where L is the network packet loss rate reported by the receiver; P is the quality rating of voice coding algorithm used for primary transmission and R_n is the quality ratings of voice coding algorithms used for the n^{th} redundancy transmissions.

3.3.3 *Primary and secondary voice streams.* For voice recovery mechanism using redundancy transmissions, a single stream of redundancy transmission will be sufficient to recover from low packet loss rate. Increasing the number of redundancy transmissions will only be justifiable if packet loss rate is escalating. However, multiple transmissions will consume higher bandwidth. If this is not kept under control, it will even cause higher losses. Moreover, multiple redundancy transmissions will increase the play-out delay as redundancies are delivered after the primary stream has been transmitted. Thus, the proposed mechanism limits the number of redundancy transmissions to two.

The number of streams (both primary and redundant data) to be transmitted is determined according to the loss rate at a particular moment. The loss rate used to determine the number of streams to be transmitted is based on the largest value of the two computed loss rates: the current loss rate and smoothed loss rate [i.e. b and λ_{new} from eqn (1)]. The current loss rate reflects a short-term reception condition. If a high current loss rate is reported, it could have indicated the start of a congestion period. Hence, it is necessary to respond immediately to the impending situation. Similarly, a sudden decrease in the current loss rate could mean a temporary recovery from high losses, but not a long-term trend. Since smoothed loss rate is computed based on accumulated past loss rates, it reflects a long-term reception condition. A decreasing current loss rate will have the effect of reducing the smoothed loss rate. Therefore, in order to increase the robustness of the mechanism, the number of streams to be transmitted is determined based on the highest value of the two loss rates.

As discussed earlier, increasing the number of data streams (or redundancies) can reduce packet loss, but it will cause a higher play-out delay and possible wastage of bandwidth. As there is a trade-off in increasing the number of streams, it is necessary to decide on how much voice recovery is needed based on the current network condition. Figure 4 shows the determination of the number of voice streams to be transmitted that is dependent on two values, the upper loss limit (U) and the lower loss limit (L). The dynamic recovery mechanism uses redundancies to reduce the packet loss rate to

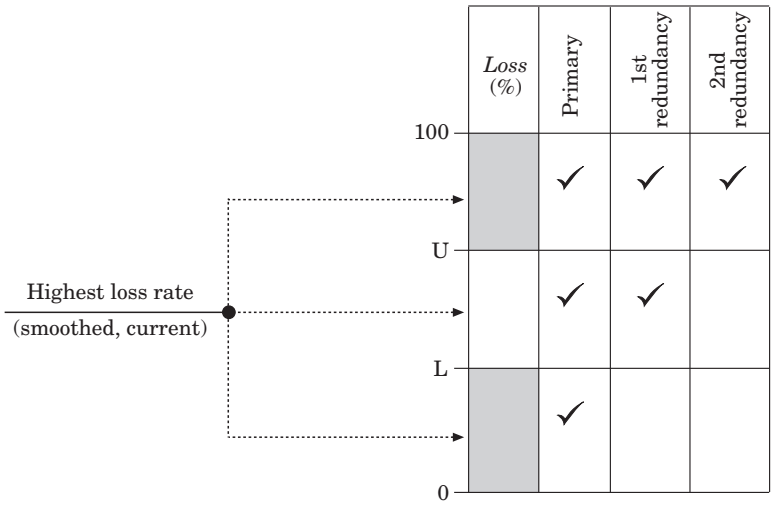


Figure 4. Determination of the number of voice streams.

be within λ_u. When the loss rate is below the threshold, λ_u, it indicates good voice quality.

The two boundary loss limits, upper loss limit and lower loss limit, are defined in relation to λ_u as follows:

$$L = \lambda_u \tag{4}$$

and

$$(1-U) + U(1-U) = 1 - \lambda_u$$

$$\Rightarrow U = \sqrt{\lambda_u} \tag{5}$$

Equation (4) defines the lower loss limit where only one stream is necessary to constrain the losses to λ_u. Above this loss limit, two streams (i.e. one primary and one redundant) are needed to help recover from the higher losses. However, when packet loss exceeds the upper loss limit as defined by eqn (5), three streams (i.e. one primary and two redundancies) will be required to maintain the losses within λ_u.

3.3.4 *Transmission strategies.* Finally, according to different network states and the number of streams to be used, the quality ratings of the expected voice signals received will be calculated for different transmission strategies. The calculation uses the value of the new smoothed loss rate λ_{new} for L and the MOS ratings for different coding algorithms based on eqn (3). The transmission strategy with the best quality rating will then be used for voice transmission.

3.4 *Example on use of recovery mechanism*

In this section, an example is given to illustrate the underlying concept of the quality-based dynamic voice recovery mechanism.

Table 2. *A list of transmission strategies*

Number of streams	Primary stream	Redundant stream 1	Redundant stream 2	Contributed bandwidth (kbps)
1	μ-Law	—	—	64
...	...	—	—	...
1	LPC-10e	—	—	2·4
2	μ-Law	μ-Law	—	128
2	μ-Law	ADPCM	—	96
2	μ-Law	LD-CELP	—	80
2	μ-Law	GSM	—	77·2
2	μ-Law	CELP	—	68·8
2	μ-Law	LPC-10e	—	66·4
2	ADPCM	ADPCM	—	64
2	ADPCM	LD-CELP	—	48
2	ADPCM	GSM	—	45·2
2	ADPCM	CELP	—	36·8
2	ADPCM	LPC-10e	—	34·4
2	LD-CELP	LD-CELP	—	32
2	LD-CELP	GSM	—	29·2
2	GSM	GSM	—	26·4
2	LD-CELP	CELP	—	20·8
2	LD-CELP	LPC-10e	—	18·4
2	GSM	CELP	—	18
2	GSM	LPC-10e	—	15·6
2	CELP	CELP	—	9·6
2	CELP	LPC-10e	—	7·2
2	LPC-10e	LPC-10e	—	4·8
3	μ-Law	μ-Law	μ-Law	192
...	...	...	...	...
3	LPC-10e	LPC-10e	LPC-10e	7·2

Assuming $\lambda_c = 0{\cdot}13$, $L_u = 0{\cdot}06$, $\mu = 2{\cdot}4$ kbps and $v = 0{\cdot}875$, then $U = \sqrt{\lambda_u} = \sqrt{0{\cdot}06} = 0{\cdot}25$ and $L = 0{\cdot}06$. In addition, if the current transmission bandwidth = 32·0 kbps, $\lambda_{old} = 0{\cdot}12$, $b = 0{\cdot}20$ and $\alpha = 0{\cdot}3$, then

$$\lambda_{new} = (1+\alpha)\lambda_{old} + \alpha b = (1-0{\cdot}3)*0{\cdot}12 + 0{\cdot}3*0{\cdot}20 = 0{\cdot}14$$

Since $\lambda_{new} > \lambda_c$, then a smaller bandwidth will be used for transmission. The revised bandwidth to be used is calculated as:

$$\text{New bandwidth} = 32*v = 32*0{\cdot}875 = 28{\cdot}0 \text{ kbps}$$

As the highest loss rate (current, smoothed) = 0·20, which falls between the lower loss limit and the highest loss limit, the mechanism will use a transmission strategy with two streams (one primary and one redundant). It will look up all the available strategies that use two streams for transmission with bandwidth limit of 28·0 kbps from a table of strategies given in Table 2.

The selected strategies are then computed for the estimated quality as shown in Table 3. As illustrated in the first two strategies of Fig. 3, the higher bandwidth requirement does not necessarily mean better quality. The one with the highest quality score is then

Table 3. *Transmission strategies with bandwidth and contributed quality*

Strategies	Bandwidth	Contributed quality
GSM, GSM	26·4	0·86*3·47 + 0·4*0·86*3·47 = 3·4
LD-CELP, CELP	20·8	0·86*4·0 + 0·14*0·86*3·2 = 3·8
LD-CELP, LPC-10e	18·4	0·86*4·0 + 0·14*0·86*2·3 = 3·7
GSM, CELP	18	0·86*3·47 + 0·14*0·86*3·2 = 3·4
GSM, LPC-10e	15·6	0·86*3·47 + 0·14*0·86*2·3 = 3·3
CELP, CELP	9·6	0·86*3·2 + 0·14*0·86*3·2 = 3·1
CELP, LPC-10e	7·2	0·86*3·2 + 0·14*0·86*2·3 = 3·0
LPC-10e, LPC-10e	4·8	0·86*2·3 + 0·14*0·86*2·3 = 2·3

selected as the transmission strategy for voice delivery. Therefore, the second strategy with LD-CELP as its primary stream and CELP as its redundant stream will be used for transmitting voice signals.

3.5 *Performance analysis*

This section illustrates the performance of the dynamic voice recovery mechanism. The aim is to show the effects of network packet loss on the quality of the play-out when the dynamic recovery mechanism is used for bandwidth adjustment and transmission control. In the experiment, the values of 0·3, 0·06, 0·13, 0·875 and 2·4 kbps for the variables α, λ_u, λ_c, v and μ are used, respectively. These values correspond to the numerical example in the previous section. The experiment has been carried out as follows. The transmitter and the receiver are separated by eight hops. The experiments were performed during two separate periods of the day. One set of results was collected between 2–3 p.m. when the network is deemed heavily utilized. During this period of time, a packet loss rate of 18% is experienced when a 15·6 kbps voice stream with a packet size of 39 bytes is transmitted. The second set of results was collected between 9–10 p.m., when the network is deemed lightly utilized; the packet loss rate is below 2%. In the experiment, one computer transmits a voice stream to the receiving computer. The computers used are Pentium II PCs running the Linux operating system. Each transmission lasts for 6 minutes. Over this period of time, the receiving computer constantly sends back reception condition to the transmitting computer. The feedback period is 5 s. Based on each feedback, the transmitting computer readjusts its transmission strategy accordingly.

Figure 5 shows the packet loss characteristics when the dynamic voice recovery is used during heavy network load condition. The experiment was started with a 128 kbps transmission with two streams of μ-law signals. This high bandwidth transmission soon put the network at stress and the packet loss was increased to 35%. The transmitter reacted to the high loss rate by readjusting its transmission strategy by using a lower bandwidth transmission strategy. As shown in Fig. 5, the network had initially reacted well to this lowering of bandwidth as the smoothed loss rate was reduced to less than 10%. However, this did not last for long as losses were increased and fluctuated around 20%, even though the least bandwidth strategy had been used for the transmission. This could be due to an increase in network usage by other network users since the proportion of network usage by the transmitter was not so significant when compared

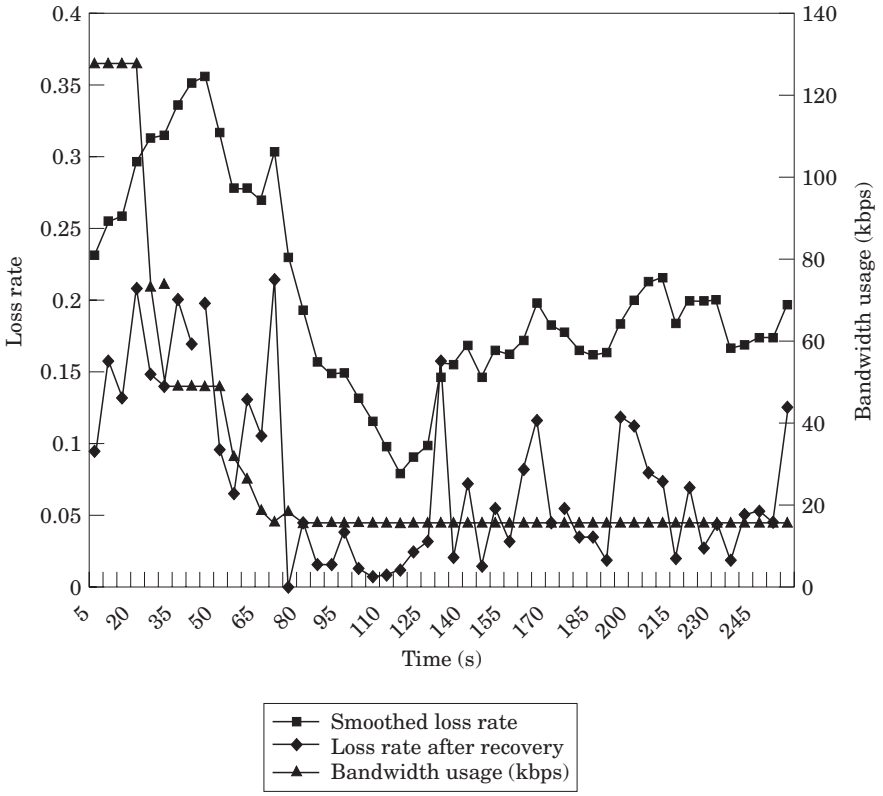


Figure 5. Dynamic recovery during heavy network load condition.

to the total network usage. Thus, the quality of the audio transmission was still very much dependent on the aggregate network usage as contributed by other users. Throughout the transmission, redundant streams were transmitted to aid recovery from high losses when losses exceeded λ_u. As illustrated in Fig. 5, packet loss after recovery was less than 10% of that which was approximately half of the smoothed packet loss rate.

As MOS ratings and smoothed loss rate have been used to calculate the expected quality of different transmission strategies, the transmission strategy used should correspond to the one having the 'best' expected quality. In order to evaluate the overall quality of voice play-out, a comparison should be made with a normal voice conversation where the signals can be characterized by its continuity and clarity. This can only be achieved if there are few packet losses and the voice codecs used for compression/decompression do not excessively degrade voice signals.

Figure 6 shows the average quality rating and loss rate of the received voice signals after all possible voice recovery from the redundancy audio packets has taken place during the period of heavy network load condition. As can be seen, the average quality rating varied between 2·5 and 4, with the quality decreasing over time. The quality rating plotted is based on the amount of voice packets received and played-out. As the voice signals contained in voice packets are of different grades of quality as a result of encoding using different codecs, the voice signals played-out are of varying quality.

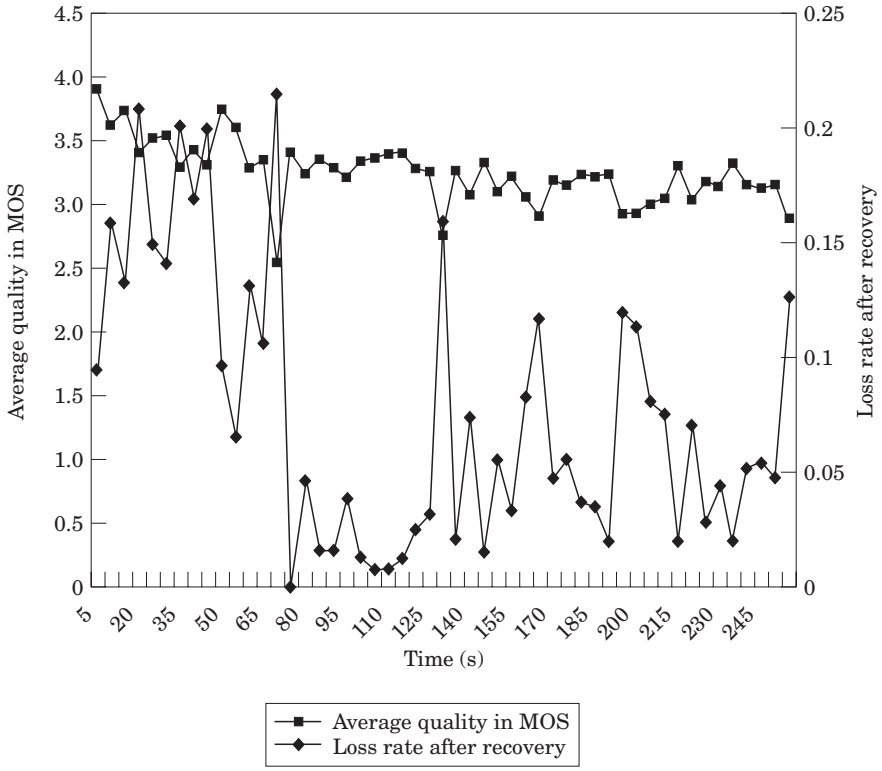


Figure 6. Average quality rating of received voice signals.

Subjective measurements had been carried out to measure the quality of the voice play-out from a group of five listeners. The results showed that the initial high packet loss had caused difficulty in listening. However, as voice segments of good quality are intermixed with those of poor voice quality, they were able to infer and understand the signals. As shown in Fig. 6, during the initial period of time the group of listeners reported a poor voice quality, but they did not have much difficulty in interpreting the signals when packet loss was reduced after the 80^{th} s. Smoother voice signals were played-out at low packet loss rate. The transmission strategy used during this period of low loss comprised GSM (MOS rating of 3·47) and LPC-10e (MOS rating of 2·4) voice signals. The poor voice quality during high packet loss was the result of transmitting lower quality voice signals in order to reduce the overall bandwidth usage. As discussed before, the reduction in transmission bandwidth was an attempt to secure a higher delivery rate by conserving bandwidth usage.

4. Internet telephone software system

The dynamic voice recovery mechanism has been incorporated into the Internet Telephone Software System (ITSS) for real time voice communication over the Internet. Figure 7 shows the system implementation of the ITSS system. As shown in Fig. 7, the recording and playback processes relay voice signals between the transmitter and

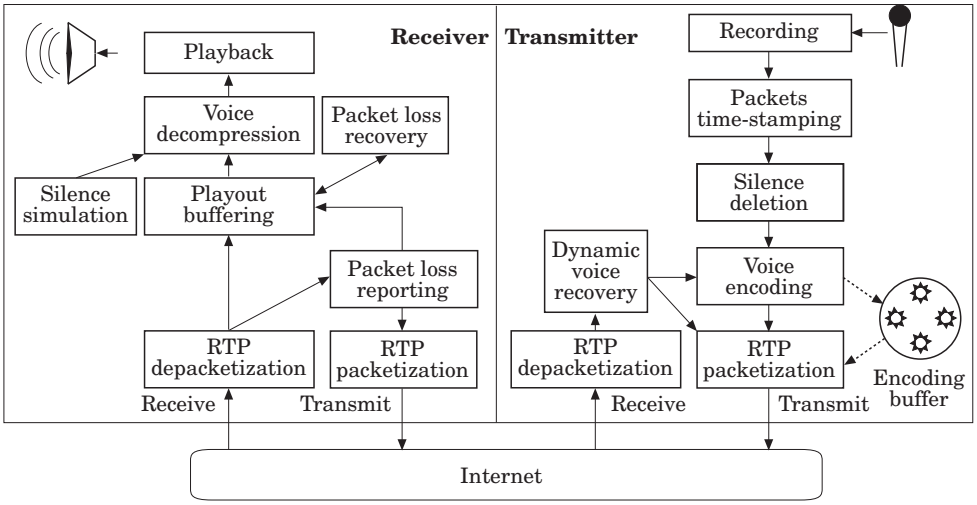


Figure 7. The Internet Telephone Software System.

receiver. The voice transmission is periodic with voices sampled at regular time intervals. The voice samples go through a series of processing activities before being transmitted. It is time-sensitive in nature. UDP/IP is used to transport the voice packets. Although the occasional non-delivery of voice packets will reduce the quality of communication, it will not be disastrous as dynamic voice recovery mechanism is incorporated to recover lost packets.

4.1 *Recording/playback*

Recording/playback provides the audio interface for recording and playing back of voice samples. It can also be used to modify the characteristics of the audio recording and playing back devices. These characteristics include the number of channels, sampling rate and sampling resolution. In the recording process, digitized voice samples are retrieved at constant intervals from the audio input device. The corresponding playback process is responsible for scheduling the playing out of voice samples.

4.2 *Packets time-stamping*

Although audio transmission is carried out in sequential order, the packets may arrive out of order due to the different network paths traversed and varying network conditions. Therefore, each packet of audio samples obtained from the audio input device is linearly time-stamped to indicate the instant of sampling. This allows the packets arriving at the receiver to be ordered in the correct time sequence before being played out. Each audio packet contains 20 ms of audio samples. The time-stamp mechanism uses 20 as the base value with increments of 20 for each new audio packet generated. The time-stamp is stored in the time-stamp field of the RTP header of each audio packet.

4.3 *Silence deletion*

There will invariably be silence periods between talk spurs. This process filters out silence packets from transmitting to the receiver. It reduces both the CPU and network

load. A number of silence deletion algorithms [15–18] such as the magnitude and zero-crossing rate algorithm, HAM algorithm, voiced algorithm, absolute algorithm, differential algorithm and exponential algorithm are available. The magnitude and zero-crossing rate algorithm is selected for implementation for its good performance. It works by first determining the average magnitude and zero-crossing rate of background noise signals as thresholds; after which, silence periods can be determined by examining the voice packets. If the average energy of a voice packet is less than the magnitude threshold and the zero-crossing rate is less than the threshold, then the voice packet is a silence packet.

4.4 *Dynamic voice recovery*

The proposed dynamic voice recovery mechanism is incorporated in this process. It receives regular packet loss reports from the receiver, analyses the packet loss rate, classifies the network state and finally determines a transmission strategy to be used for transmission based on the available bandwidth. The data encoding and redundancy information of the selected transmission strategy are then passed to the voice encoding process for further processing.

4.5 *Voice encoding*

Voice encoding reduces the bit rate requirement of digital voice samples imposed on the network. The voice samples received from the recording process are in pulse code modulation (PCM) which is the raw audio format. This format is the direct digital representation of the strength of audio signals at each sampling instant. The coding algorithms reduce the voice sample's size by representing the voice signal in another format such as A-law, μ-law, ADPCM, LD-CELP, GSM and PLC.

In addition, as voice samples may be transmitted a multiple number of times as redundancies, additional encoding is needed for the redundancies. This may result in the unnecessary need of computational power if the previously compressed voice samples for primary transmission are not utilized. Thus, an encoding buffer has been created to contain compressed voice samples to enable efficient reuse for secondary transmissions.

4.6 *RTP packetization*

This process packetizes the processed voice information according to the RTP format [19–21]. It computes all the required information necessary to complete the RTP headers, and in particular, it generates a sequence number for each RTP packet. The sequence number serves two purposes. First, it can be used to order the incoming RTP packets, which may have arrived out-of-sequence due to different network paths and delivery delays. Moreover, it allows the receiver to monitor the packets loss rate and jitters experienced by the packets.

4.7 *RTP depacketization*

This process is the reverse process of RTP packetization. Information carried by incoming RTP packets is separated into two data structures: RTP header, and voice data.

4.8 *Play-out buffering*

The purpose of play-out buffering is to cushion the out of order, late delivery and jitters experienced by the packets. The play-out buffering process uses a linked-list of buffers to store incoming voice data. When a new packet arrives, its time-stamp will be compared with the time-stamp of the last packet played-out. If the new packet's time-stamp is smaller, then it indicates that a late loss has occurred and the new packet will be discarded. Otherwise, the voice packet will be used for playing out and it will then be placed into the appropriate position within the linked-list according to its time-stamp. Besides containing the voice data, each buffer also contains the play-out time for each voice data.

4.9 *Packet loss recovery*

Any loss of voice data during transmission can be discovered according to the sequence numbers of the received voice packets. The missing voice data can then be identified using the time-stamping information that is associated with each voice packet. Packet loss recovery processes will then try to recover the missing voice data from the redundant voice data transmitted. Once recovered, the voice data will be inserted into the play-out buffer for playing out.

4.10 *Simulation of silence packets*

This process simulates the silence effect from the voice packets transmitted from the transmitter. It identifies silence data from examining the time-stamps of successive voice packets. As the difference between two successive voice packets should be 20, it indicates the period of silence data if the difference is greater than 20. Silence periods will then be simulated between two packets accordingly.

4.11 *Packet loss reporting*

The RTP header extracted from the depacketization process will serve as an input to this process. It computes the packet loss and jitters and passes them to the transmitter. The jitter information is also fed back to the play-out buffering process for adjusting the play-out time. The play-out time for each talk spur is constantly re-adjusted according to the current reception condition to allow more packets to be played-out that will otherwise be lost due to late arrival. As a consequence of this adjustment, there will be an artificial elongation of silence periods.

4.12 *Voice decompression*

Voice decompression reverses the compressed voice samples into the uncompressed format before sending it to the playback process. At a constant time interval, voice decompression will obtain and uncompress a voice packet from the play-out buffers.

5. Conclusion and future work

In this paper, a quality-based dynamic voice recovery mechanism that enhances the quality and reliability of real time voice communication for Internet telephony systems

has been described. The dynamic mechanism aims to provide a controlled mechanism for achieving higher speech continuity and speech quality through bandwidth control and the use of redundancy transmissions. This mechanism has been successfully incorporated into the Internet Telephone Software System (ITSS) and verified to have shown improved audio qualities.

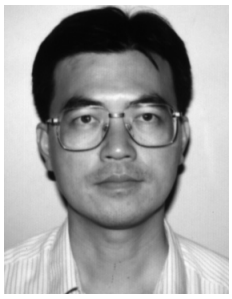
Currently, the adaptation of the same dynamic recovery mechanism for use in an Internet telephony gateway system is under investigation. In a gateway system, multiple communication sessions are needed. Network bandwidth will be shared among all the transmitting streams of voice signals. Higher bandwidth usage of one stream may then affect the voice delivery of another stream. Therefore, it is necessary to provide an appropriate amount of the bandwidth for each stream so that each stream can provide quality voice delivery to the destination.

References

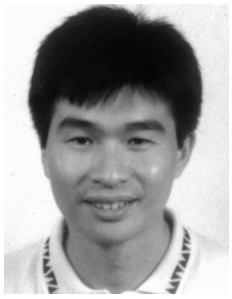
1. VocalTec Communications Ltd. Internet Phone. Online document can be found at URL: http://www.vocaltec.com
2. Microsoft Corporation. NetMeeting 2.0. Online document can be found at URL: http://www.microsoft.com/netmeeting
3. H. Schulzrinne 1992. Voice Communication Across the Internet: a network voice terminal. Department of Electrical and Computer Engineering & Computer Science, University of Massachusetts. Online document can be found at URL: ftp://gaia.cs.umass.edu/pub/hgschulz/nevot/
4. F. Fluckiger 1995. *Understanding Networked Multimedia Applications and Technology.* New Jersey: Prentice-Hall.
5. D. E. Comer 1995. *Internetworking with TCP/IP.* Vol 1. New Jersey: Prentice-Hall.
6. R. Ramjee, J. Kurose, D. Towsley and H. Schulzrinne 1994. Adaptive playout mechanisms for packetized audio applications in wide-area networks. In: Proceedings of the Conference on Computer Communications (IEEE INFOCOMM) Montreal, Canada, pp. 680–688.
7. V. Hardman, M. A. Sasse, M. Handley and A. Watson 1995. Reliable audio for use over the Internet. In: Proceedings of INET95 (Ohahu, Hawaii).
8. J. Rosenberg 1996. Reliability Enhancements to NeVoT. Bell Laboratories. Online document can be found at URL: http://www.cs.columbia.edu/~jdrosen/aisfinal/aisindex.html
9. J. C. Bolot and A. Vega-Garcia 1996. A Control Mechanisms for Packet Audio in the Internet. In: Proceedings of the Conference on Computer Communications (IEEE Infocom). San Francisco, California, pp. 232–239.
10. S. Floyd and K. Fall 1998. Promoting the Use of End-to-End Congestion Control in the Internet. Lawrence Berkeley National Laboratory (LBNL). Online document can be found at URL: http://www-nrg.ee.lbl.gov/floyd/tcp__unfriendly.html
11. J. Busse, B. Deffner and H. Schulzrinne 1996. Dynamic QoS control of multimedia applications based on RTP. *Computer Communications*, **19**, 49–58.
12. K. V. Chin, S. C. Hui and S. Foo 1997. An Internet telephone software system for real-time voice communication. *International Journal of Computer and Engineering Management*, **5**, 37–52.
13. N. S. Jayant and P. Noll 1984. *Digital Coding of Waveforms: principles and applications to speech and video.* New Jersey: Prentice-Hall.
14. P. E. Papamichalis 1987. *Practical Approaches to Speech Coding.* New Jersey: Prentice-Hall, pp. 177–198.
15. J. Riedl and M. Claypool 1994. Silence is golden? The effects of silence deletion on the CPU load of an audioconference. In: Proceedings of IEEE Multimedia Computing and Systems Conference. Boston, Massachusetts, pp. 9–18.
16. M. H. Savoji 1989. A robust algorithm for accurate endpointing of speech signals. *Speech Communication*, **8**, 45–60.

17. L. R. Rabiner and R. W. Schafer 1978. *Digital Processing of Speech Signals.* New Jersey: Prentice-Hall, pp. 130–133.
18. C. K. Un and H. H. Lee 1980. Voice-unvoiced-silence discrimination of speech by delta modulation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **28**, 398–407.
19. H. Schulzrinne, S. Casner, R. Frederick and V. Jacobson 1996. RFC 1889 RTP: a transport protocol for real-time applications (PROPOSED STANDARD). Audio-Video Transport Working Group. Online document can be found at URL: ftp://ftp.nus.sg/pub/docs/rfc/rfc1889.txt
20. H. Schulzrinne 1996. RTP Profile for Audio and Video Conferences with Minimal Control (PROPOSED STANDARD). Audio-Video Transport Working Group. Online document can be found at URL: ftp://ftp.nus.sg/pub/docs/rfc/rfc1890.txt
21. C. Perkins, I. Kouvelas, O. Hodson, V. Hardman, M. Handley, J. C. Bolot, A. Vega-Garcia and S. Fosse-Parisis. 1997. RFC 2198 RTP Payload for Redundant Audio Data (PROPOSED STANDARD). Network Working Group. Online document can be found at URL: ftp.//ftp.nus.sg/pub/docs/rfc/rfc2198.txt

Kian Vine Chin is a graduate student in the School of Applied Science, Nanyang Technological University (NTU), Singapore. He received his BASc (Hons) degree in Computer Engineering in 1996 from NTU. His research interests are communication systems over the Internet and Internet technology.



Siu Cheung Hui is Senior Lecturer at the School of Applied Science, Nanyang Technological University. His current research interests include document retrieval, database systems, Internet technology and multimedia systems.



Schubert Foo is the Head of the Division of Information Studies, School of Applied Science at Nanyang Technological University. His current research interests include multimedia technology, Internet technology, CSCW systems, information retrieval systems and digital libraries.