PACKET VOICE RECOVERY TECHNIQUES FOR REAL-TIME INTERNET VOICE COMMUNICATION

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

School of Applied Science, Nanyang Technological University
Nanyang Avenue, Singapore 639798

ABSTRACT

The Internet’s phenomenal growth, worldwide coverage and increasing data delivery capabilities have made possible its use as an alternative communications medium for real-time voice messages. In addition, using the Internet for voice communication circumvents the high costs of long-distance phone rates. However, the Internet is a relatively harsh environment for real-time voice communication. Packets transmitted over the Internet are subjected to late deliveries and losses. This paper discusses the various voice packet recovery techniques to enhance voice delivery for Internet voice communication. These include both receiver only, and source and receiver only techniques.

INTRODUCTION

Internet telephony circumvents expensive international telephone toll charges by using Internet as an intermediary medium to transport voice signals between two users. In doing so, transcontinental telephone calls are achieved at the mere costs of local telephone calls and nominal Internet connectivity charges. A wide range of Internet telephony systems has been developed and marketed. However, the quality of communication using these products is still not comparable to those offered by telephone companies. The inferior quality is mainly due to the high transmission delay and packets lost of the Internet environment which is characteristic of packet-switched network without resource reservations mechanisms.

To resolve these problems, different mechanisms are developed to handle the delay jitters and packet loss problems. For examples, the play-out time of arriving audio packets can be adjusted at the destination to minimise the impact of delay jitters using a buffering mechanism. Voice recovery and adaptive rate control mechanisms can be used to eliminate or minimise the impact of packet loss.

In this paper, we will briefly describe and discuss the different voice packet recovery techniques to enhance voice delivery for Internet voice communication. These include silence substitution, waveform substitution, sample substitution, Xor mechanism, embedded speech coding and dynamic voice recovery mechanism.

INTERNET TELEPHONY ENVIRONMENT

Constructing an Internet telephony system that offers good performance is a challenging problem. Figure 1 shows the basic components of an Internet telephony environment.
Two host computers acting as caller and recipient are required. In using the standard Internet Transmission Control Protocol/Internet Protocol (TCP/IP) [1], each host computer is identified by a unique Internet Protocol (IP) address. 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 telephony system will acquire the real-time voice data through an audio input device and convert the analogue signals into digitised 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 telephony system carries out the reverse process. Incoming data is first decrypted, decompressed and played back in real-time on the audio 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.

PACKET VOICE RECOVERY TECHNIQUES

Two main types of voice recovery are available when packets containing voice signals are lost: Automatic Repeat Request (ARQ) and Forward Error Correction (FEC). With ARQ, all packets originated from the source will arrive at the destination without errors. Errorless transmission is achieved through acknowledgements issued by the destination for each arriving packet. Any erroneous or missing packets will be re-sent by the source. Although such end-to-end acknowledgements and retransmission guarantee the delivery of all packets, they incur a time delay that can be excessive for time-sensitive applications such as Internet telephony systems. On the other hand, voice recovery using FEC reconstructs the lost packets directly from the original data transmissions without re-sending lost packets. This technique increases the resilience to packet loss at the expense of less accurate reproduction of voice data. In contrast with ARQ, it does not incur additional time lag re-transmitting missing packets. Therefore, it is more suitable for time-sensitive applications.

Various techniques on voice recovery using FEC have been proposed. Hardman et al. [2] classify these techniques into either receiver only, or source and receiver only. In the receiver only technique, only the receiver is responsible for the reconstruction of missing voice segments by using whatever available voice information. Common receiver only techniques are silence substitution, waveform substitution and sample interpolation. In the source and receiver only technique, both source and destination are responsible for recovering the missing voice segments. Examples of the source and receiver only
techniques include the Xor mechanism [3] and the embedded speech coding technique [2].

**Receiver Only**

*Silence Substitution*

This is the simplest method among all the techniques, however, it is not able to maintain an acceptable quality of playback audio in the event of high lost rate and large packet size. In this technique, lost packets are simply replaced with silence. Hence, applications using this technique do not incur much additional processing power and is therefore suitable for an environment where the probability of losing packets is low and the computers do not have much processing power.

*Waveform Substitution*

In this method, when a segment of audio fails to arrive at the destination on time, the previous segment of audio is used to replace the missing segment of audio. The assumption of this technique is that the speech characteristics have not changed much from a preceding speech segment and it is logical to use the previous segment of speech to reconstruct the missing portion.

This method does not work for large packet size as the speech characteristics are most likely to change noticeably from one previous packet to the next. Moreover, it also does not guard against the continuous lost of multiple packets where speech characteristics do not remain the same over the duration of packets loss. As with Silence Substitution, it does not demand lots of processing power. Hence, it is used in some of the interactive voice communication applications.

*Sample Interpolation*

This technique is similar to Waveform Substitution, however, it does not directly replace all missing audio segments with the previously received segments. It modifies the previous audio packets before substituting the missing audio segments with it.

The method assumes that the audio characteristics change slightly over a short period of time. In order to use previously received samples to replace the missing audio segments and at the same time accommodating the slight change in audio attribute, the missing samples are estimated based on the previous samples' characteristics. A simple form of sample modification is linear interpolation of audio.

In comparison, it requires more processing power than the previous two methods, but it offers a better contingency solution. As with Waveform Substitution, it is not usable in a prolonged duration of packets loss as it is likely that the audio characteristics will change significantly.

**Source and Receiver Only**

*Xor mechanism*

Xor mechanism uses an exclusive-or mechanism to provide a RAID style redundancy to tackle packet loss. The idea is based on the work done by Shacham and McKenney [4]. For every n packets of voice segments, it generates a parity frame, as the exclusive-or of each of the n packets. This parity frame is piggybacked into the next audio packet. When a single packet of voice segments is lost, the missing voice segments can be recovered by computing the exclusive-or of the parity frame with the other n-1 correctly received packets. This is a relatively simple mechanism for a single packet loss in an n packet message; however, it does not guard against multiple consecutive packet loss. Moreover, an additional frame of parity is sent which increases the send rate from the source.
**Embedded Speech Coding Technique**

The embedded speech coding technique is introduced by Hardman et al. [2]. This technique implements voice recovery by adding redundancies to the audio packets sent by the source. In order to implement this technique without incurring excessive bandwidth, both toll and non-toll quality voice coding algorithms are used in the primary and redundancy transmissions respectively. Therefore, the redundant voice segment is of lower quality than the primary voice data. The use of different coding algorithms is necessary, as better quality voice coding algorithms demand higher network bandwidth and more processing power. In this way, the output speech waveform will consist of periods of toll quality speech, interspersed with periods of synthetic quality speech. The synthetic quality speech coding algorithm, Linear Predictive Coding (LPC), is used for redundant voice encoding.

![Diagram: Positioning of Redundancy in the Packet](image)

Figure 2: Positioning of Redundancy in the Packet.

a) for low loss rates b) for higher loss rates.

Hardman et al. further proposed that the placement of redundancies be dependent on the network loads condition. According to Bolot et al. [5], in a light and intermediate loads condition, losses are essentially non-consecutive for an audio stream, and for heavy loads, the behaviour is similar, but consecutive losses are more prevalent. Hence, [2] proposed that redundancies placed immediately in the following packet works well in a light and medium network loads. Whereas, redundancies placed in a number of packets later is more suitable for heavy loads. Figure 2 is a pictorial description of the packet structure.

Bolot and Vega-Garcia [6] further suggest using multiple coding algorithms for redundant data encoding. Apart from LPC, other coding algorithms such as Adaptive Delta Modulation (ADM) and Global System for Mobile communication (GSM) are also used. It extends the concept of redundancy transmission by embedding redundancies of multiple voice segments instead of a single voice segment into a packet to enhance the ability to recover from multiple consecutive losses and high packet loss rate. Adding redundant voice segments increases the CPU and the bandwidth requirements. In addition, using multiple redundancies will cause a higher play-out delay and the possible wastage of bandwidth if redundancy is not utilised. However, this technique is more resilient to packet loss than the other methods discussed so far.

Furthermore, Bolot and Vega-Garcia [6] propose a combined rate and error control mechanism which couples the embedded speech coding technique with a rate control mechanism. In this combined mechanism, the amount of redundant voice segments added in audio packets at the source is based on feedback information about the loss rate as measured at the destination. The mechanism chooses one of the pre-defined combinations which contain information about a primary coding algorithm, redundant coding algorithm(s), a CPU cost, bandwidth requirements, a delay and a reward (auditory quality)
depending on the state of the network congestion. This mechanism focuses on maintaining speech continuity and quality using multiple transmissions. However, the quality of the received voice signals might not be optimal. It depends on which combination was chosen for use during transmission. The rate control is only used when the network is congested. Upon detection of congestion, the amount of redundant information transmitted will increase without decreasing the source send rate. This will give the priority to the audio streams over the network. However, this may result in network congestion collapse [7]. During network congested condition, congestion collapse occurs when an application attempts to improve its performance by using more and more bandwidth without success. On the other hand, the increase of such bandwidth usage forces other congestion-conscious applications to react by using less bandwidth.

**Dynamic Voice Recovery**

Embedded Speech Coding technique demands the most processing power among all the techniques described so far. However, it achieves a far better performance than all other methods.

While it is important to seek high speech continuity in voice communication for better comprehension, 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 Embedded Speech Coding, the most notable overhead arises when additional bandwidth is needed for transmitting redundant information. Although redundant information may help to achieve better voice reception and continuity, it causes additional load to network resources. In addition, it is necessary to apply dynamic transmission control to adjust the bandwidth usage dynamically when different levels of network congestion are encountered in order to avoid congestion collapse.

Therefore in the dynamic voice recovery approach, the dynamic transmission control [8] is integrated with the voice recovery [2, 6] using a quality-based measurement. The quality measurement is derived from the coding algorithms used to encode both primary and redundant data. The dynamic transmission control is incorporated to provide the bandwidth adjustment function when different levels of congestion are encountered. In addition, multiple redundancies are used to enable better reception and recovery of voice signals during congested network condition.

![Figure 3: Quality-Based Dynamic Voice Recovery Mechanism.](image)

Figure 3 shows the quality-based dynamic voice recovery mechanism. The mechanism consists of three phases as follows:

**Analysis of Packet Loss Rate.** This phase analyses the packet loss data received from incoming receiver reports and uses a low-pass filter to smooth the packet lost rate statistics.

**Classification of Network Congestion State.** In this phase, the smoothed loss rate generated from the first phase is used for the determination and classification of the network congestion state. Three network states, namely, Unloaded, Loaded and Congested have been defined according to some pre-defined
thresholds as shown in Figure 4. The upper threshold $\lambda_c$ gives the upper limit where voice quality will be unacceptable if this threshold is exceeded. The lower threshold, $\lambda_u$, is defined such that packet loss rate below this limit will give good voice quality. For 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 or reduce the current bandwidth according to the unloaded, loaded or congested state accordingly. In this case, the linear increase and multiplicative decrease [8] are used for the proportion of bandwidth change.

![Figure 4: Classification of Network Congestion State.](image)

**Determination of Transmission Strategy.** 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. The transmission strategy with the best quality rating is selected for transmitting voice signals.

As packet loss results in momentarily 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

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

where the function Q represents the quality rating measured using MOS [9] and L is the network packet loss rate.

In order to minimise 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 resilience to the effects of packet loss.

When multiple redundancies transmissions are considered, the quality rating (Q) for the voice signals received can be derived from equation (1) as follows:
\[ Q = (1-L)P + L_1^1 (1-L) R_1 + L_2^2 (1-L) R_2 + \ldots + L_n^n (1 - L) R_n \quad -- (2) \]

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.

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. The current loss rate reflects a short-term reception condition. If 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 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 computed loss rates.

![Diagram showing the determination of the number of voice streams.](image)

Figure 5: Determination of the Number of Voice Streams.

Figure 5 shows the determination of the number of voice streams to be transmitted that is dependent on two values, the Upper Loss Limit \((U)\) and Lower Loss Limit \((L)\). The dynamic mechanism uses redundancies to reduce the packet loss rate to be within \(\lambda_u\). When the loss rate is below the threshold, \(\lambda_u\), it indicates good voice quality.

The two boundary loss limits, Upper Loss Limit and Lower Loss Limit, are defined in relation with \(\lambda_u\) as follows:

\[ L = \lambda_u \quad -- (3) \]

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

\[ \Rightarrow U = \sqrt{\lambda_u} \quad -- (4) \]

Equation (3) defines the Lower Loss Limit where only one stream is necessary to constrain the losses to
\( \lambda_u \). Above this loss limit, 2 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 equation (4), three streams (i.e. one primary and two redundancies) will be required to maintain the losses within \( \lambda_u \).

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 new smoothed loss rate for \( L \) and the MOS ratings for different coding algorithms based on equation (2). The transmission strategy with the best quality rating will then be used for voice transmission.

Performance results have shown that the voice recovery mechanism is able to exercise bandwidth control to reduce loss rate. However, at low transmission rates, loss characteristics are influenced more heavily by the total network usage of all network users. In order to improve voice reception, voice signals are compressed using lower bandwidth voice encoding algorithms and transmitted as primary and secondary voice streams to aid the voice recovery process at the destination.

**CONCLUSION**

In this paper, we have described the various voice packet recovery techniques to enhance voice delivery for Internet voice communication. Simple and low complexity receiver only techniques are sufficient for improving the quality of received voice when packets loss is minimal. On the other hand, more complex source and receiver only techniques must be used when packets loss rate is high.

**REFERENCES**


