MULTI-STREAM VOICE OVER IP USING PACKET PATH DIVERSITY

Yi J. Liang, Eckehard G. Steinbach, and Bernd Girod
Information Systems Laboratory, Department of Electrical Engineering
Stanford University, Stanford, CA 94305, USA
{yiliang, steinb, bgirod}@stanford.edu

Abstract - We propose multi-stream transmission of real-time voice over best-effort packet networks such as today's Internet, where multiple redundant descriptions of the voice stream are sent over independent network paths. At the receiver, multi-stream adaptive playout scheduling is employed to improve the tradeoff among delay, late loss rate, and speech quality. Experiments over the Internet suggest largely uncorrelated statistical characteristics, such as erasure probability and delay jitter, for different network paths, which leads to a noticeable path diversity gain. We have obtained significant reductions in mean end-to-end latency and loss rates compared to FEC protected single-path transmission at the same data rate. The speech quality perceived by the receiver is evaluated using the recently standardized objective quality measure PESQ. In our experiments, we observe gains of more than 0.4 PESQ score for voice transmission with packet path diversity.

INTRODUCTION

High quality real-time voice communication over the Internet requires low end-to-end delay and low loss rate. Best effort networks such as the Internet, however, are characterized by highly varying delay and loss characteristics. One way to reduce the effective packet loss is to add redundancy to the voice stream at the sender. This is possible without imposing too much extra network load since the data rate of voice traffic is very low compared to other types of traffic. A common method to add redundancy is forward error correction (FEC), which transmits redundant information of each packet in subsequent packets (Fig. 1 (a)). In this sender-based scheme, loss recovery is performed at the cost of higher latency [1]. In many cases, however, the losses of successive packets are correlated and a packet loss may be followed by a burst packet loss, which significantly decreases the efficiency of FEC [2].

In this work we look at the problem of reliable voice communication over best-effort networks from a different angle. Instead of restricting our transmission to one network path, we send multiple redundant descriptions of the voice stream over different independent paths and take advantage of the largely uncorrelated statistical characteristics of loss and delay. As a result, the probability of a negative disturbance, such as packet erasure or increasing delay, impacting all channels at the same time will be small.

The multiple streams to be delivered via different paths are formed by multiple description coding (MDC), which generates multiple descriptions of the source signal

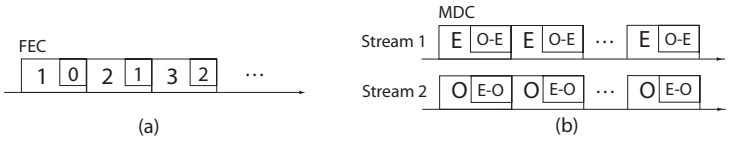

Figure 1: Source encoding: a) single-stream with FEC; b) MDC.

of equal importance that can be decoded independently at the receiver (Fig. 1 (b)). If all descriptions are received, the source signal can be reconstructed in full quality. If we do not receive all descriptions, the quality of the reconstruction is degraded, but still better than the quality resulting from losing all the descriptions. With MDC, the overall data rate of the payload does not necessarily increase, depending on the encoding scheme and whether redundancy is added.

In previous work, packet path diversity has been proposed for reliable video communication over lossy networks using multiple state encoding, where odd and even frames of a video sequence are transmitted on different network paths [3]. Multi-path transmission also alleviates the problem that the default path determined by the routing algorithm is not always optimal which might be often the case according to [4].

In this work we improve the delay - loss tradeoff for real-time voice communication over the Internet by taking advantage of the largely uncorrelated behavior of the delay variation (also known as *jitter*) on multiple independent network paths. Packet loss in delay-sensitive applications such as interactive VoIP is a result of not only packet erasure, but also delay jitter. Due to the stringent delay budget and the need to output speech periodically and continuously, packets experiencing sudden high delay have to be discarded at the receiving end if they arrive later than the scheduled playout deadline (which results in *late loss*). With multiple streams, if the packets from one path experience erasure or excessive delay, packets from the other stream can often be used in substitute to avoid late loss. Latency can also be reduced by playing out the voice description with lower delay if full audio quality is not a priority.

With today's Internet protocols, multi-path transmission can be realized by a dedicated overlay network of relay servers [3] or by exploiting future peer-to-peer architectures. In these schemes, different routes can be made by explicitly sending packets to intermediate relay nodes and then forwarding them to the destination. With the next-generation IP protocol IPv6, the source will have a better control over each packet's route, which makes future implementation of multi-path transmission even simpler.

PLAYOUT SCHEDULING OF MULTIPLE STREAMS

One important functionality to be implemented at the receiver is playout scheduling of the voice packets, or in other words, setting the time when to play out the received packets. Under the stringent delay requirements, packets could be discarded due to their late arrival resulting from increasing network delay.

Before the arrival of each packet i, we set the playout deadline for that packet according to the most recent delays we recorded. The playout deadline of packet i is

denoted by d^i_{play}, which is the time from the moment the packet is delivered to the network until it has to be played out. When determining the playout deadlines, we have to consider the tradeoff among delay, losing both MDC coded packets (we refer to this as packet *erasure* in order to distinguish it from the next case), and losing only one description. The latter two cases result in a speech distortion. This tradeoff can be stated as follows. Given a certain acceptable speech distortion, minimize the average latency $\mathcal{E}(d^i_{play})$, which is a constrained problem. We convert this constrained formulation into an unconstrained one by introducing a Lagrange cost function for packet i as

$$\begin{aligned} C^i &= d^i_{play} + \lambda_1 \cdot probability(\text{both descriptions lost}) \\ &\quad + \lambda_2 \cdot probability(\text{only one description lost}) \\ &= d^i_{play} + \lambda_1 \hat{\varepsilon}^i_{S_1} \hat{\varepsilon}^i_{S_2} + \lambda_2 (\hat{\varepsilon}^i_{S_1}(1 - \hat{\varepsilon}^i_{S_2}) + \hat{\varepsilon}^i_{S_2}(1 - \hat{\varepsilon}^i_{S_1})), \end{aligned} \tag{1}$$

where $\hat{\varepsilon}^i_{S_1}$ and $\hat{\varepsilon}^i_{S_2}$ are the estimated loss probabilities of the packet from stream 1 and 2, respectively, given d^i_{play}. The estimate of $\hat{\varepsilon}^i_{S_1}$ and $\hat{\varepsilon}^i_{S_2}$ is based on d^i_{play} and order statistics of past delay values that are recorded for the two streams. The higher d^i_{play} is, the lower the loss probabilities since the likelihood of playing out late packets is better. The Lagrange multipliers λ_1 and λ_2 are predefined parameters to trade off delay and the two loss probabilities.

The playout deadline is obtained by searching for the optimal d^i_{play} which minimizes the cost function (1). Perceptually, high latency and degraded speech quality resulting from packet loss are "orthogonal" experiences. The multiplier λ_1 is used to trade off total delay and packet erasure probability. Greater λ_1 results in lower erasure rate at the cost of higher latency.

The multiplier λ_2 is introduced to give penalty to speech distortion as a result of playing only one description. The greater λ_2 is, the better the quality of the reconstructed speech signal at the cost of higher delay. Note that although packet erasure (the second term in (1)) and quality degradation due to the loss of one MDC description (the third term in (1)) are different perceptual experiences, they are not "orthogonal" measures. From (1) it can be deduced that increasing λ_2 also leads to lower erasure probability. However, with zero or very small λ_2 only packet erasure is given concern. In this case good reconstruction quality is not a priority but low latency is given more emphasis, with the tradeoff between delay and erasure determined mainly by λ_1. In practice, this is usually desirable since the human perceptual experience is most strongly impaired by high latency, while speech distortion resulting from losing one description only increases the quantization noise in the MDC scheme we use here (will be described in the next section) and is usually tolerated as a minor impairment.

When switching between streams during speech playout, the playout schedule needs to be dynamically adjusted and adapted to the delay statistics of each individual stream. The dynamic setting of each packet's playout schedule is achieved by an adaptive playout technique proposed in our earlier work [5]. In such a scheme, proper reconstruction of continuous output speech is achieved by scaling individual voice packets using a time-scale modification technique which modifies the rate of speech while preserving its pitch.

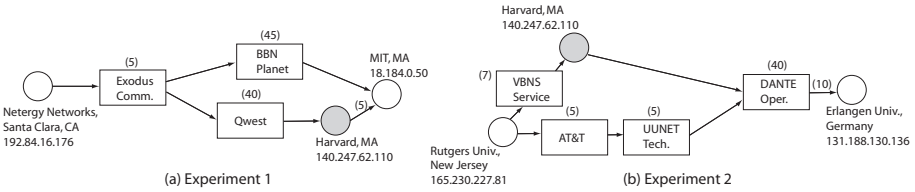

Figure 2: Experimental setup. Source and destination hosts are shown as white circles and relay servers as gray circles. Intermediate service providers are represented by boxes. The numbers in parentheses show the average time in ms required for the packets to traverse corresponding providers or other interconnected networks.

INTERNET EXPERIMENTS

We transmit two 8kHz sampled MDC coded voice streams in 30ms UDP packets from the source host to the destination host at different geographic locations. For stream 1, even samples in a packet are quantized in finer resolution (PCM, 8 bits/sample) and the difference between adjacent even and odd samples in coarser resolution (ADPCM, 2 bits/sample), as described in [6]. For stream 2, even and odd samples are quantized in the opposite way (Fig. 1 (b)). Using this scheme, the redundancy imposed when neglecting the packet headers is 25%. We compare our scheme with a scheme that uses single-stream FEC at the same payload data rate. In the FEC scheme, the source packet is coded with the same finer quantization as before, and a secondary copy of the packet is coded with the same coarser quantization and carried by the subsequent packet (Fig. 1 (a)).

In Experiment 1 (Fig. 2(a)), the source host is located at Netergy Networks in California and the destination host is at MIT. The first stream follows the direct path, following the route determined by the default routing algorithm. For the alternative path, we explicitly direct the flow to a designated relay server at Harvard and let the relay server forward the packets to the destination. We measured the correlation coefficient of the delays of the two streams to be 0.028. This suggests that the delays on the two path are largely uncorrelated. In Experiment 2 a long cross-Atlantic link is shared by the two paths. The sender host is in New Jersey and the receiver at the University of Erlangen in Germany. Again, a host at Harvard serves as the relay server. Despite the shared link, the delay correlation turns out to be only 0.034.

We first compare the delay - loss tradeoff by setting λ_2 to zero and varying λ_1 in (1) during playout. The results are plotted in Fig. 3 (a) for using Stream 1 only, using Stream 2 only, and using both streams. The *average total delay* is the average value of d^i_{play} of all the received packets in a playout session. The *erasure rate* is the percentage of erased packets (neither description played out), no matter if the loss is a result of channel erasure or late arrival. The adaptive playout technique in [5], which already achieves state-of-the-art performance for single stream transmission, has been applied to all schemes under comparison.

From Fig. 3 (a), we observe a significant reduction of the packet erasure rate for a

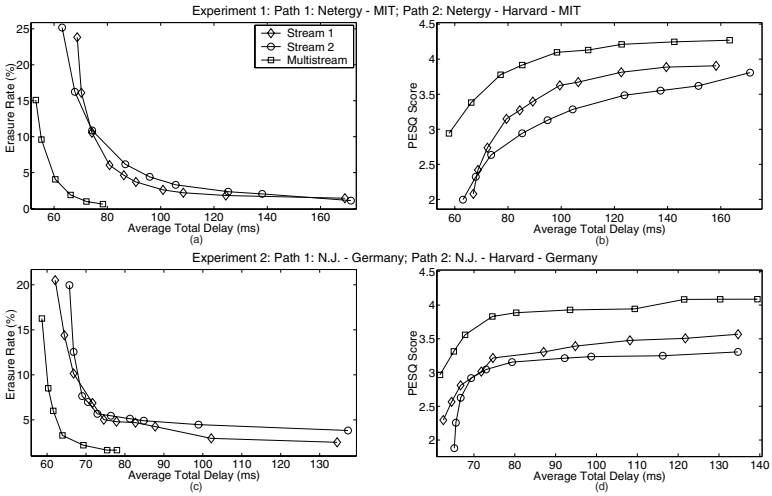

Figure 3: Internet experiments: erasure rate and PESQ score vs. average total delay.

fixed target delay when using multi-stream multi-path transmission. At the same average delay of 70 ms, the erasure rate is reduced from more than 16% to less than 2%, compared with using single-stream transmission with FEC. The majority of packet loss in this experiment is not a result of channel erasure, but late loss caused by the high delay jitter.

In Fig. 3 (a), the average delay is also much lower for the multi-stream packet path diversity scheme. At the same erasure rate of 5%, the average delay is reduced by more than 20 ms. The delay reduction can be explained by the possibility to play out the description with the lower delay if obtaining full voice quality is not a priority.

One important issue to be addressed in this context is whether the improved trade-off between delay and loss is achieved at the cost of compromised speech quality, e.g., when we decide to play out the description which arrives earlier while discarding the other one most of the time. In the next experiment, we measure the quality of the reconstructed speech signal for the different schemes. We use PESQ (perceptual evaluation of speech quality), which is an objective measure for narrow-band speech recently adopted by the ITU-T [7]. Unlike previous objective measures, PESQ is applicable not only to speech codecs but also to end-to-end measurements, since it takes into consideration factors such as filtering, variable delay, coding distortions and channel errors. The range of the PESQ score is -0.5 to 4.5, but for most cases the output is a MOS-like score between 1.0 and 4.5 [7]. In our experiments, erased packets are concealed using information from a prior packet, while in other situations speech packets are reconstructed depending on how many MDC descriptions are received by the playout deadline.

In Fig. 3 (b) we have plotted PESQ score vs. delay as we vary λ_1 and λ_2 while

keeping their ratio fixed. It is obvious that as delay increases, speech quality is better due to the lower erasure rate and higher probability that both MDC copies are played out successfully. In Fig. 3 (b) and (d), the voice quality corresponding to multiple streams is better than that using single-stream FEC by more than 0.4 PESQ score for all delays. This indicates that the improved tradeoff between delay and loss using MDC path diversity transmission is obtained without compromising voice quality. This can be explained by the fact that the benefit (such as the lower erasure rate) obtained from the path diversity scheme outweighs the loss of one out of the two MDC descriptions, since packet erasure introduces much higher perceptual distortion than quantization noise. Similar results are observed in Experiment 2 (Fig. 3 (c) (d)).

CONCLUSION

In this work, we propose a scheme for real-time voice transmission using multiple independent network paths. By taking advantage of the largely uncorrelated delay and loss statistics of different paths, packets can be used and played out from the channel that currently presents superior transmission characteristics. Experiments over the Internet show that with this scheme the mean end-to-end latency and erasure rate can be greatly reduced compared with using FEC protected single-stream transmission at the same payload data rate. The overall speech quality is evaluated with an objective measure, showing typical gains of more than 0.4 PESQ score.

References

[1] J.-C. Bolot, S. Fosse-Parisis, and D. Towsley, "Adaptive FEC-based error control for Internet telephony," in *Proceedings of IEEE INFOCOM '99*, Mar. 1999, vol. 3, pp. 1453–1460.

[2] J.-C. Bolot, "End-to-end packet delay and loss behavior in the Internet," *Computer Comm. Review*, vol. 23, no. 4, pp. 289–298, Sept. 1993.

[3] J. G. Apostolopoulos, "Reliable video communication over loss packet networks using multiple state encoding and path diversity," in *Proceedings Visual Communication and Image Processing*, Jan. 2001, pp. 392–409.

[4] S. Savage, A. Collins, E. Hoffman, J. Snell, and T. Anderson, "The end-to-end effects of Internet path selection," *Computer Comm. Review*, vol. 29, no. 4, pp. 289–99, Oct. 1999.

[5] Y. J. Liang, N. Färber, and B. Girod, "Adaptive playout scheduling using time-scale modification in packet voice communications," in *Proceedings ICASSP01*, May 2001, Salt Lake City, UT.

[6] W. Jiang and A. Ortega, "Multiple description speech coding for robust communication over lossy packet networks," in *International Conference on Multimedia and Expo*, Aug. 2000, vol. 1, pp. 444–7, New York, NY, USA.

[7] 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.