JOINT PLAYOUT AND FEC CONTROL FOR ENHANCING PERCEIVED QUALITY OF MULTI-STREAM VOICE COMMUNICATION

Yung-Le Chang, Chun-Feng Wu and Wen-Whei Chang

Institute of Communications Engineering, National Chiao-Tung University
Hsinchu,Taiwan
phone: + (886) 5731826, email: wwchang@cc.nctu.edu.tw

ABSTRACT

A new objective method is presented for predicting the perceived quality of multi-stream voice transmission. Also proposed is a joint playout buffer and FEC adjustment scheme that maximizes the perceived speech quality via delay-loss trading. Experimental results showed that the proposed scheme achieves significant reductions in delay and packet loss as well as improved speech quality.

1. INTRODUCTION

Quality of Service (QoS) is of prime importance in real-time voice communication over IP networks. In addition to packet loss and delay, the delay jitter obstructs the timely reconstruction of the speech signal at the receiver. A playout buffer is often used to store recently arrived packets before playing them out at scheduled intervals. By increasing the buffer size, the late loss rate is reduced, but the resulting improvement in voice transmission is off-set by the accompanying increase in the end-to-end delay. In balancing the impairment due to delay and packet loss, two current coding strategies, single and multiple description transmissions, have used different playout buffer algorithms. In single description (SD) coding, adaptive algorithms [1]-[2] have been proposed along with the E-model [3] for perceptual optimization of playout buffer. Taking a different approach, multiple description (MD) coding [4]-[5] exploits the packet path diversity such that each description can be individually decoded for a reduced quality reconstruction, but if all descriptions are available, they can be jointly decoded for a better quality reconstruction. For multi-stream voice transmission, Liang et al. [4] proposed an algorithm which uses the Lagrangian cost function to trade delay versus loss by following a play-first strategy. They neither consider the quality degradation due to frequent switching among playout scenarios nor try to optimize the perceived speech quality. In predicting the overall quality of MD transmission, the E-model is expected to show two limitations. First, it may fail to register impairments due to reconstruction based on information from a single path as opposed to from both paths, when no packets from either path are lost. Moreover, the resulting detrimental effects that accompany the change in the playout scenarios may thus be ignored and harm its prediction of the overall quality.

As a further step toward perceptual optimization, our study also attempts to strengthen the error concealing capabilities of MD by including into our proposed MD scheme an forward error control (FEC) mechanism [6]. Previous efforts toward linking FEC with playout buffer for single-stream transmission can be found in [7], but the assumption on which their algorithm was based may limit its applicability. Specifically, it was assumed that the single-stream network over which the voice packets are sent delivers packets in sequence. This line of reasoning has been challenged by a number of related studies [8] that addressed the possibility of packets delivered out of sequence because of network jitter. In this paper, a multi-stream voice quality prediction model is presented to develop a joint FEC and playout control scheme which will ignore the constraints imposed by the no-reordering assumption made in [7].

2. SYSTEM IMPLEMENTATION

The implementation procedure consisted of description generation and description transmission over two independent network paths. Fig. 1 shows a block diagram of the system with the first two components, MD speech coder and channel coder, responsible for description generation and the rest, for transmission and signal reconstruction. For description generation, the MD-G.729 based speech packetization scheme described in [5] was used to generate two descriptions from the bitstream of the ITU-T G.729 codec [9]. Afterwards, packet-level Reed-Solomon $RS(N,K)$ codes [6] are used for channel coding of individual descriptions. The channel encoder takes a codeword of K speech packets and generate $N-K$ additional FEC check packets for the transmission of N packets over the network. During description transmission, the best-effort nature of IP networks results in packets experiencing varying amounts of loss and delay due to different levels of network congestion. To characterize this, we used the ns-2 network simulator to generate the traces of VoIP traffic for different network topologies and varying network load.

The receiver end features a joint playout and FEC adjustment scheme which is formulated as an optimization problem on the basis of a minimum overall impairment criterion. As a prerequisite for obtaining impairments estimation on which the joint design could be based, a delay distribution model was established as it could provide a direct link to late loss rate in the presence of jitter. Previous work in [2] has found that the delay characteristics of VoIP traffic can be represented by statistical models which follow Pareto and Exponential distributions depending on applications. Finally, the MD-G.729 bit stream is decoded and degraded speech is generated. The decoder performs differently in dealing with the three description arrival situations: If both descriptions are lost, the error concealment algorithm of G.729 [9] is used, while in other situations, speech packets are reconstructed depending on how many descriptions are received by the playout deadline. If both descriptions are received, the central decoder performs the standard G.729 decoding process after combining the two descriptions into one bitstream. If only one description is lost, the side decoder substitutes the

missing information by using received parameters from the other description or information from the most recent correctly received frame [5].

3. MULTI-STREAM VOICE QUALITY PREDICTION MODEL

The E-model [3] combines the delay impairment I_d and equipment impairment I_e into a single factor $R = 94.2 - I_d - I_e$. The task of defining the R-factor for multi-streams voice transmission lies in the difference that any subset can be used for signal reconstruction, and that the transmission quality improves with the size of the subsets. Thus, in addition to delay and packet loss, our prediction model aims to address the issue of impairments due to dynamic size allocations during the speech playout. For two-path transmission, we need to consider two kinds of playout scenarios at the receiver end. Specifically, a packet is 1) fully restored with two descriptions and thus played with high quality; and 2) partially restored with one description and thus played with degraded quality. For brevity, let $I_{e,k}$ denote the equipment impairment corresponding to the playout scenario S_k as a result of playing out k received descriptions. Conditioned on the event that the packet can be restored, we let r_k be the probability to play out the packet using k descriptions. Formally, it is given by $r_k = P(S_k)/(P(S_1)+P(S_2))$.

From the perceived QoS perspective, the MD-G.729 codec may be viewed as operating at two coding rates: 4.6 kbps for S_1 and 8 kbps for S_2. By taking frequent switch of coding rates into account, we define the average equipment impairment due to MD-G.729 coding as $I_e(e) = r_1 I_{e,1}(e) + r_2 I_{e,2}(e)$, where e is the packet-erasure rate in percentage. Following the work of [1], we derived the $I_{e,k}$ model for scenario S_k in form of $I_{e,k}(e) = \gamma_{1,k} + \gamma_{2,k}\ln(1+\gamma_{3,k}e)$, where $(\gamma_{1,k},\gamma_{2,k},\gamma_{3,k})$ for S_1 are $(52.61, 7.52, 10)$ and $(21.96, 17.02, 16.09)$ for S_2. Fig. 2 shows that impact of transmission scenario S_k and packet-erasure rate e on the equipment impairment $I_{e,k}$ with a packetization of one frame per packet.

4. FEC IN A GILBERT-MODEL LOSS PROCESS

Assume that multiple descriptions of the speech are transmitted over independent network paths and each path is characterized by a Gilbert-model loss process. The Gilbert model is a two-state Markov chain model in which state B represents a network loss and state G represents a packet reaching the destination. For each stream l, the parameters $p^{(l)}$ and $q^{(l)}$ denote respectively the probabilities of transitions from G to B states and from B to G states. A packet is said to be missing so long as the packet is either dropped in the network or discarded due to its late arrival. Let $R_l(m,n,D_i)$ denote the probability that $m-1$ packets are missing (dropped or received late) in the next $n-1$ packets following the network loss of packet i, and let $S_l(m,n,D_i)$ denote the probability that $m-1$ packets are missing in the next $n-1$ packets following the late loss of packet i. Similarly, let $\tilde{R}_l(m,n,D_i)$ and $\tilde{S}_l(m,n,D_i)$ denote the probability that $m-1$ missing packets occur in the last $n-1$ packets preceding packet i which is dropped and received late, respectively. In [10], we showed that these probabilities can be computed by recurrence as follows:

$$R_l(m,n,D_i) = \tag{1}$$
$$\begin{cases} q^{(l)}(1-p^{(l)})^{n-2}\prod_{h=1}^{n-1}(1-e^{(l)}_{b,i+h}), & m=1, n\geq 1 \\ \sum_{j=1}^{n-m}\{\{p^{(l)}R_l(m-1,n-j-1,D_{i+j+1})+A_l\} & \\ \cdot q^{(l)}(1-p^{(l)})^{j-1}\prod_{h=1}^{j}(1-e^{(l)}_{b,i+h})\} & \\ +(1-q^{(l)})R_l(m-1,n-1,D_{i+1}), & 2\leq m\leq n \end{cases}$$

$$S_l(m,n,D_i) = \tag{2}$$
$$\begin{cases} e^{(l)}_{b,i}(1-p^{(l)})^{n-1}\prod_{h=1}^{n-1}(1-e^{(l)}_{b,i+h}), & m=1, n\geq 1 \\ \sum_{j=0}^{n-m}\{q^{(l)}R_l(m-1,n-j-1,D_{i+j+1})+A_l\} & \\ \cdot e^{(l)}_{b,i}(1-p^{(l)})^{j}\prod_{h=1}^{j}(1-e^{(l)}_{b,i+h}), & 2\leq m\leq n \end{cases}$$

$$A_l = S_l(m-1,n-j-1,D_{i+j+1})\cdot(1-p^{(l)})e^{(l)}_{b,i+j+1} \tag{3}$$

where D_i is the FEC delay and $e^{(l)}_{b,i}$ is the estimated late loss probability of packet i in stream l.

RS (N,K) code can recover any missing packet in the block if and only if at least K out of N packets in this block are received before their playout time. Viewed from this perspective, the probability to recover a dropped packet is

$$P^{(l)}_1(i) \tag{4}$$
$$= \sum_{L=1}^{N-K}\sum_{m=0}^{\hat{M}}\tilde{R}_l(m+1,i,D_i)R_l(L-m,N-i+1,D_i)$$

where $\hat{M} = min(L-i, i-1)$ and the probability to recover a late lost packet is given by

$$P^{(l)}_2(i) \tag{5}$$
$$= \sum_{L=1}^{N-K}\sum_{m=0}^{\hat{M}}\tilde{S}_l(m+1,i,D_i)S_l(L-m,N-i+1,D_i)$$

Using these probabilities, we can compute the residual loss probability (after FEC is used) as follows:

$$P^{(l)}_L(i) = e^{(l)}_n(1-P^{(l)}_1(i)) + (1-e^{(l)}_n)e^{(l)}_{b,i}(1-P^{(l)}_2(i)) \tag{6}$$

where $e^{(l)}_n$ represents the network loss probability measured in stream l. The packet-erasure probability e_i is defined as the probability that none of the descriptions of packet i arrives on time, and is given by $e_i = P^{(1)}_L(i)P^{(2)}_L(i)$.

5. JOINT FEC AND PLAYOUT CONTROL

We formulated the joint playout and FEC control as a perceptually motivated optimization problem and the criterion relies on the proposed multi-stream quality prediction model. First, we applied an autoregressive algorithm [1] to estimate

the mean $\hat{d}$ and variance $\hat{v}$ of network delay, and use them to calculate the buffer delay $d_b = \hat{d} + \beta\hat{v}$. Waiting for the FEC check packets results in additional delay and, consequently, the playout delay is given by $d_{play} = \hat{d} + \beta\hat{v} + (N-1)T_p$, where T_p is the packet generation interval. In this work, a β-adaptive algorithm is instead used to control the buffer size so that the reconstructed voice quality is maximized in terms of delay and loss.

Our general problem can be stated as follows: Given estimates of the parameters characterizing the packet loss and delay distribution, find the optimal values of β and $\{N,K\}$ so as to minimize the overall impairment function subject to the rate constraint. Let d_i be the end-to-end delay experienced by the ith packet, which consists of encoding delay d_c and playout delay d_{play}. Now, we define an overall impairment function I_m with the following form

$$I_m(d_i,e_i) = I_d(d_i) + \frac{1}{K}\sum_{j=1}^{K}\sum_{l=1,2} r_l I_{e,l}(e_j). \qquad (7)$$

where $r_1 + r_2 = 1$ and the probability to receive both descriptions is given by $r_2 = (1-P_L^{(1)}(i))(1-P_L^{(2)}(i))/1-e_i$. Our optimization framework requires an analytic expression for the packet erasure probability e_i as a function of the parameter β. Notice that $e_{b,i}^{(l)}$ and the playout delay d_{play} are strongly correlated, and to find out their relationship, the network delays of stream l are assumed to follow a Pareto distribution which is defined as $F_D^{(l)}(d) = 1-(g_l/d)^{\alpha_l}$. Then, the late loss probability of packet i in stream l can be computed as $e_{b,i}^{(l)} = 1-F_D^{(l)}(D_i) = (g_l/D_i)^{\alpha_l}$, where $D_i = d_{play} - (i-1)T_p$.

Finally, we summarize the proposed multi-stream joint playout and FEC adjustment algorithm as below.

1. At the beginning of each talkspurt, update network delay records for the past 200 packets in every stream l ($l = 1,2$), and use them to calculate the Pareto distribution parameters (α_l, g_l).
2. Use the values of (α_l, g_l) to compute the late loss probability $e_{b,i}^{(l)}$ and the packet erasure probability e_i. Find the minimizer $(\hat{\beta}_i^{(l)}, \hat{N}^{(l)}, \hat{K}^{(l)})$ of the overall impairment function in (7) subject to the code rate constraint $\frac{N}{K} \times \frac{9.2}{8} \leq 2$.
3. Set $d_{play} = \hat{d}^{(l^*)} + \hat{\beta}_i^{(l^*)}\hat{v}^{(l^*)} + (\hat{N}^{(l^*)}-1)T_p$ and $(N,K) = (\hat{N}^{(l^*)}, \hat{K}^{(l^*)})$, where $l^* = \arg\min\{I_m(\hat{\beta}^{(l)}, \hat{N}^{(l)}, \hat{K}^{(l)}), l = 1,2\}$

6. EXPERIMENTAL RESULTS

Computer simulations were carried out to evaluate the performances given by the four MD voice transmission schemes, MD1-4, which all used the MD-G.729 for source coding and RS(N,K) code for channel coding. The speech data fed into the simulations were two sentential utterances spoken by one male and one female, each sampled at 8 kHz and 8 seconds in duration. Among the four schemes, MD1 had its parameters $\{\beta,N,K\}$ dynamically adjusted according to the proposed voice quality prediction model, while MD2-4 shared a fixed $\beta = 4$ with (N,K) set at (3,2), (5,3), and (10,6) respectively. It should be pointed out that the last two (N,K) sets allowed MD3 and MD4 to perform at the same FEC coding ratio but with different lengths of delay, which gave us the opportunity to evaluate in our test environment the effect of packet loss vs. delay.

Fig. 3 plots the perceived speech quality for the four schemes as a function of link loss rate. Among the four schemes, MD4, with the longest end-to-end delay, yielded the lowest *R*-factors, while MD3, with the same FEC coding ratio but shorter delays than those set for MD4, yielded higher *R*-factors than MD4, but lower *R*-factors than MD2. MD2 with the lower delay impairment allowed it to outperform MD3 and MD4, but its strength of packet recovery receded faster as the link loss rate was increased. The best results were obtained with the currently proposed scheme MD1.

7. CONCLUSIONS

We presented a perceptually motivated optimization criterion and a practically feasible new algorithm for multi-stream voice transmission. Experimental results show that the proposed multi-stream voice transmission scheme can achieve a better delay-loss tradeoff and thereby improves the perceived speech quality.

REFERENCES

[1] L. Sun and E. Ifeachor, "Voice quality prediction models and their application in VoIP networks," *IEEE Transactions on Multimedia*, August 2006.

[2] K. Fujimoto, S. Ata, and M. Murata "Adaptive Playout Buffer Algorithm for Enhancing Perceived Quality of Streaming Applications ," in *Processings of IEEE Globecom*, Nov 2002.

[3] International Telecommunication Union, "The E-model, a computational model for use in transmission planning, " *ITU-T Recommendation G.107*, July 2000.

[4] Y.J. Liang, E.G. Steinbach, and B. Girod, "Multi-stream voice over IP using packet path diversity," in *Multimedia Signal Processing IEEE Fourth Workshop*, 2001, pp. 555–560.

[5] J. Balam and J. D. Gibson "Multiple Descriptions and Path Diversity for Voice Communications Over Wireless Mesh Networks, " *IEEE Transactions on Multimedia*, August 2007.

[6] S. Lin and D.J. Costello, Error Control Coding, Pearson Prentice Hall, New Jersey, 2004.

[7] C. Boutremans and J. Boudec, "Adaptive Joint Playout Buffer and FEC Adjustemnt for Internet Telephony," in *Processings of IEEE INFOCOM*, 2003.

[8] Chia-Chen Kuo, Ming-Syan Chen, and Jeng-Chun Chen, "An Adaptive Transmission Scheme for Audio and Video Synchronization based on Real-time Transport Protocol, " in *IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, August 2001.

[9] International Telecommunication Union, "Coding of Speech at 8kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP), " *ITU-T Recommendation G.729*, Nov. 2000.

[10] Yung-Le Chang, "Adaptive Joint Playout Buffer and FEC Adjustment for Multi-Stream Voice Over IP Networks, " Master Thesis, Institute of Communications

Engineering, National Chiao Tung University, Republic of China, June 2009.

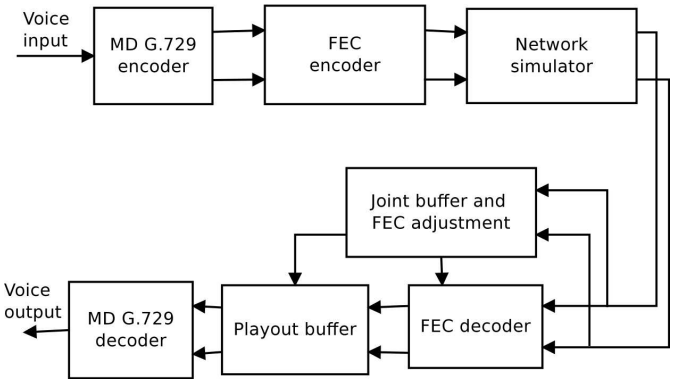

Figure 1: A multi-description voice transmission system.

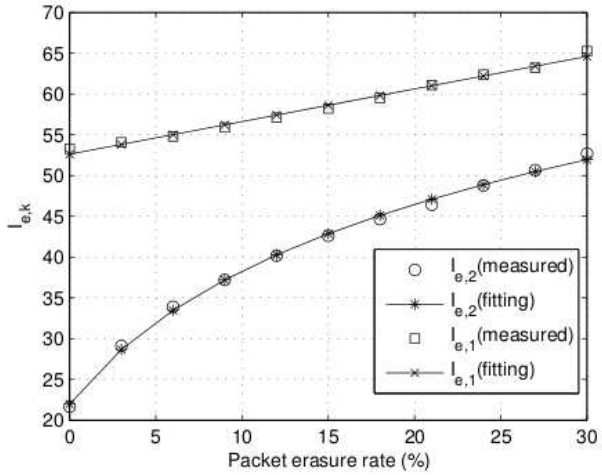

Figure 2: $I_{e,k}$ vs. packet erasure rate e.

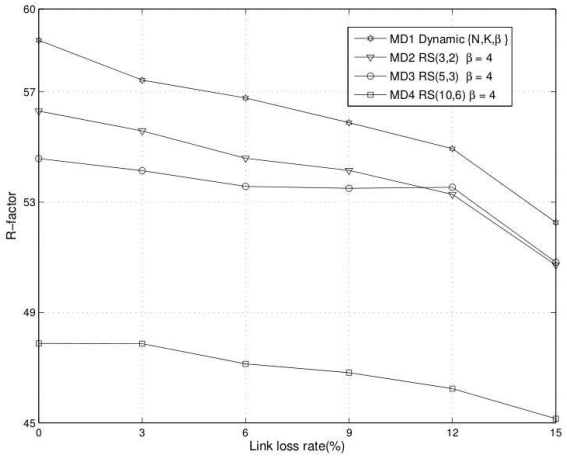

Figure 3: Performance comparison for different MD schemes.