PERCEPTUAL-BASED PLAYOUT MECHANISMS FOR MULTI-STREAM VOICE OVER IP NETWORKS

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

Department of Communications Engineering, National Chiao-Tung University
Hsinchu, Taiwan, ROC
wwchang@cc.nctu.edu.tw

ABSTRACT

Packet loss and delay are two essential problems to real-time voice transmission over IP networks. In the proposed system, multiple descriptions of the speech are transmitted to take advantage of largely uncorrelated delay and loss characteristics on different network paths. Adaptive playout scheduling of multiple voice streams is formulated as an optimization problem leading to a better delay-loss tradeoff. Also proposed is a perceptually motivated optimization criterion based on a simplified version of the ITU-T E-model. Experimental results show that the proposed multi-stream playout algorithm improves the delay-loss tradeoff as well as speech reconstruction quality.

1. INTRODUCTION

In Voice over IP (VoIP) applications, packet loss and delay are the main network impairments that affected perceived voice quality. Recent research [1-2] proposed the use of multiple description (MD) coding to exploit the largely uncorrelated delay and loss characteristics on different network paths. In MD coders, the source is encoded into multiple redundant descriptions that are separately transmitted over independent network paths. Each description can be individually decoded for a reduced quality reconstruction of the source, but if all descriptions are available they can be jointly decoded for a higher quality reconstruction. For the multi-stream voice transmission, the network delay experienced may vary for each packet depending on the paths taken by different streams and on the level of congestion along the path. Packets could get lost due to their late arrival resulting from excessive network delays.

The variation in network delay, referred to as jitter, must be smoothed out since it obstructs the proper and timely reconstruction of the speech signal at the receiver end. The most common approach is to store recently arrived packets in a jitter buffer before playing them out at scheduled intervals. By increasing the buffer size, the number of late packet loss can be reduced at the cost of increased end-to-end delay. Thus, there is a need to develop playout buffer algorithms for a better balance between the end-to-end delay and packet loss. Most of the proposed playout buffer algorithms [1-3] focused on the delay-loss performance, but not the quality perceived by end users. Recent work [4-7] has considered new models for perceived voice quality prediction and their applications in playout buffer optimization for single-stream voice transmission. In this work, we will extend the concept of perceptual optimization to adaptive playout scheduling of multiple voice streams.

2. SYSTEM IMPLEMENTATION

A block diagram of the proposed multi-stream voice transmission system is shown in Fig. 1. The system has four major components: MD speech coder, network simulator, delay distribution modelling and adaptive playout buffer. Following the work of [8], we used the MD speech coder to generate two side descriptions from the bitstream of the G.729 codec [9]. The coder operates in a way that each description is of the same rate 4.6 kbps and speech decoded from either description is of similar quality. The best-effort nature of the Internet results in packets experiencing varying network delay due to different levels of network congestion. To characterize this, we used the ns-2 network simulator to generate different categories of network delay traces for performance evaluation. We simulate sending two constant bit rate (CBR) voice streams from source to destination via two paths, with TCP data traffic contending for network resources at the same time. Fig. 2 shows a multi-hop topology for network simulation. Each CBR stream is transmitted in 10-ms UDP packets at a rate of 9.2 kbps. The first stream follows the route from node N1 through N3 to the destination, while Stream 2 follows route from N4 through N6 to the destination. The intermediate nodes N1 through N6 represent access points on the routes for data traffic. Each of these nodes has a number of data sources attached, with a large amount of incoming TCP traffic heading for different destinations.

Delay jitter can be removed by buffering the received packets for a short period of time before playing them out at scheduled intervals. Before the arrival of packet i, we have to determine the playout time for that packet according to the most recent delays we recorded. This task is accomplished by using an adaptive playout buffer algorithm that achieves the optimum perceived voice quality in the presence of jitter. To proceed with this,

it is important to establish the delay distribution model as it is directly related to late packet loss rate. Previous work in [6] has found that the delay characteristics of voice over Internet is better characterized by a Pareto distribution than a normal or an exponential distribution.

3. MULTI-STREAM VOICE QUALITY PREDICTION MODEL

The E-model, defined in the ITU-T Recommendation G.107 [10], is an analytic model of conversational speech quality used for network planning purposes. It combines individual impairments due to both the signal's properties and the network characteristics into a single R-factor ranging from 0 to 100. In VoIP applications [11], the R-factor may be simplified as follows: $R = 94.2 - I_d - I_e$, where I_e is the equipment impairment factor and I_d is the delay impairment factor. The R-factor is related to the conversational mean opinion score (MOS_c) through a fixed mapping in [10]. The delay impairment factor can be derived by a simplified fitting process in the form

$$I_d(d) = 0.024d + 0.11(d - 177.3)H(d - 177.3) \quad (1)$$

where d is the end-to-end delay and $H(x)$ is the step function. Although the ITU E-model is commonly used, the derived I_e model is applicable to a restricted number of codecs and this hinders its use in new applications. To address this, objective methods for deriving the model parameters have been proposed in [12], but this is limited to a consideration of only the single-description voice transmission. Recognizing this, we will extend the objective methods to predict perceived quality of multi-stream voice over IP networks.

For two-stream voice transmission, each channel can either deliver all its bits or deliver none of its bits, so the two channels will always be in one of three possible states: no loss, loss in exactly one channel, and loss in both channels (packet erasure). Let S_1 be the channel state that both descriptions are received, and S_2 be the channel state that only one description is received. Corresponding to each channel state S_k, the MD decoder reconstructs the source signal with an equipment impairment factor $I_{e,k}$. We next present the objective method for deriving the $I_{e,k}$ model for each channel state. The reference speech signal is first MD-encoded and then processed in accordance with the network loss characteristics to and generate the degraded speech. The degraded speech and reference speech are then fed to the PESQ to obtain a measurement of speech quality due to loss and codec. For each speech sample in the data set, a MOS score for a packet erasure rate is obtained by averaging over 30 different packet erasure locations. Further, these MOS scores are averaged over all speech samples recorded by eight males and eight females. Values of MOS obtained from PESQ are transformed to R-factor and then to $I_e = 94.2 - R$. The curves for measured $I_{e,k}$ versus packet erasure rate e are shown in Fig. 3. From the figure a nonlinear regression model can be derived for each channel state by the least squares method and curve fitting. The derived $I_{e,k}$ model has the following form: $I_{e,k}(e) = \gamma_{1,k} + \gamma_{2,k} \ln(1 + \gamma_{3,k} e)$, where the fitting parameters $(\gamma_{1,k}, \gamma_{2,k}, \gamma_{3,k})$ are codec- and state-dependent.

4. PERCEPTUAL OPTIMIZATION OF SCALE FACTOR

Although there are methods which use fixed playout algorithms, better algorithms have been proposed that react to changing network conditions by dynamically adjusting the playout delay. Here we focused on adaptive playout algorithms and adjust the buffer between talkspurts. The basic adaptive playout algorithm operates by estimating two statistics characterizing the network delay, and uses them to calculate the playout delay as follows:

$$d_{play,i} = \hat{d}_i + \beta \hat{v}_i. \quad (2)$$

where $\hat{d}_i$ and $\hat{v}_i$ are running estimates of the mean and variance of network delay seen up to the ith arriving packet. Here β is the safety factor which can be used to set the playout time to be far enough beyond the delay estimate; so that only a small fraction of packets will arrive too late to be played out. A higher value of β results in a lower late loss rate as more packets arrive in time, however the end-to-end delay increases. The safety factor β has a critical impact on the adjustment of playout delay, which in turn influences the delay-loss tradeoff. Compared with fixed β in existing perceptual-based playout algorithms [4-6], further enhancement is expected with dynamic setting of β_i for every packet i. In this work, β_i is adapted according to the observed delay distribution and the adopted criterion relies on the use of a simplified version of the conversational-quality E-model.

Perceptual-based buffer design must take into account the tradeoff between delay, packet loss, and speech reconstruction quality. We formulated this tradeoff as an optimization problem which involves finding the best value of the decision variable β_i for every packet i. By the best variable we mean the one that results in smallest value of the utility function defined by

$$\begin{aligned} I_{m,i}(d_i, e_i) &= I_d(d_i) + I_e(e_i) \\ &= I_d(d_i) + \textstyle\sum_{k=1,2} P\{\frac{S_k}{S_1 \cup S_2}\} I_{e,k}(e_i). \end{aligned} \quad (3)$$

where the end-to-end delay d_i is a summation of the encoding delay d_c and the playout delay $d_{play,i}$, i.e. $d_i = d_c + \hat{d}_i + \beta_i \hat{v}_i$. The erasure probability of packet i can be expressed as

$$\begin{aligned} e_i = \; & e_n^{(1)} e_n^{(2)} + e_n^{(1)}(1 - e_n^{(2)}) e_{b,i}^{(2)} + e_n^{(2)}(1 - e_n^{(1)}) e_{b,i}^{(1)} \\ & + (1 - e_n^{(1)})(1 - e_n^{(2)}) e_{b,i}^{(1)} e_{b,i}^{(2)} \end{aligned} \quad (4)$$

where $e_n^{(l)}$ and $e_{b,i}^{(l)}$ represent the network loss probability and the late loss probability in stream l, respectively. The probability of channel state S_1 is given by

$$P\{\frac{S_1}{S_1 \cup S_2}\} = \frac{1}{1-e_i}(1-e_n^{(1)})(1-e_n^{(2)})(1-e_{b,i}^{(1)})(1-e_{b,i}^{(2)}) \quad (5)$$

Through the delay distribution modelling, as described in the following section, the packet erasure probability e_i can be represented in terms of the safety factor β_i. This reduces the expression of the utility function to be $I_{m,i}(d_i, e_i) = I_{m,i}(\beta_i)$. By differentiating it with respect to β_i, we get the following equation for the gradient:

$$\begin{aligned} I_{m,i}^{'}(\beta_i) = \; & c\hat{v}_i + \textstyle\sum_{k=1,2}\{P\{\frac{S_k}{S_1 \cup S_2}\}\frac{\gamma_{2,k}\gamma_{3,k}}{1+\gamma_{3,k}e_i}\frac{de_i}{d\beta_i} + \\ & \frac{dP\{\frac{S_k}{S_1 \cup S_2}\}}{d\beta_i} I_{e,k}\}. \end{aligned} \quad (6)$$

where

$$c = \begin{cases} 0.024, & \beta_i < (177.3 - d_c - \hat{d}_i)/\hat{v}_i; \\ 0.134, & \beta_i > (177.3 - d_c - \hat{d}_i)/\hat{v}_i. \end{cases} \quad (7)$$

$$\begin{aligned} & \frac{dP\{\frac{S_1}{S_1 \cup S_2}\}}{d\beta_i} = \\ & \frac{\hat{v}_i}{d_{play,i}(1-e_i)}(1-e_n^{(1)})(1-e_n^{(2)})[\alpha_1 e_{b,i}^{(1)}(1-e_{b,i}^{(2)}) \\ & +\alpha_2 e_{b,i}^{(2)}(1-e_{b,i}^{(1)})] + \\ & \frac{1}{(1-e_i)^2}\frac{de_i}{d\beta_i}(1-e_n^{(1)})(1-e_n^{(2)})(1-e_{b,i}^{(1)})(1-e_{b,i}^{(2)}) \end{aligned} \quad (8)$$

Proceeding in this way, the secant method [11] is then applied to find the perceptual optimum value of β_i. Starting with two initial values $\beta_i(-1)$ and $\beta_i(0)$, the iterative formula for the secant algorithm has the form

$$\beta_i(j+1) = \beta_i(j) - \frac{\beta_i(j) - \beta_i(j-1)}{I_{m,i}^{'}(\beta_i(j)) - I_{m,i}^{'}(\beta_i(j-1))} I_{m,i}^{'}(\beta_i(j)). \quad (9)$$

The new value $\beta_i(j+1)$ is then used in the next iteration and the estimation process is repeated until the difference $|\beta_i(j+1) - \beta_i(j)|$ is smaller than a threshold.

5. PERCEPTUAL-BASED MULTI-STREAM PLAYOUT ALGORITHM

The main attraction of multi-stream voice transmission arises from its flexibility to trade off the end-to-end delay, losing both descriptions, and losing only one description. The latter two cases results in different degrees of speech quality degradation. For this investigation, we will extend the concept of perceptual optimization to adaptive playout scheduling of multiple voice streams. We first applied an autoregressive algorithm [3] to estimate the mean delay $\hat{d}_i^{(l)}$ and delay variance $\hat{v}_i^{(l)}$ for individual stream l ($l = 1, 2$) as follows:

$$\hat{d}_i^{(l)} = \alpha \hat{d}_{i-1}^{(l)} + (1-\alpha) n_i^{(l)}. \quad (10)$$

$$\hat{v}_i^{(l)} = \alpha \hat{v}_{i-1}^{(l)} + (1-\alpha)|n_i^{(l)} - \hat{d}_i^{(l)}|. \quad (11)$$

where $n_i^{(l)}$ is the actual network delay and $\alpha = 0.998002$ is a weighting factor for convergence control.

The next issue to be addressed is how to associate the safety factor β_i with the packet erasure probability e_i, which in turn influences the calculation of the gradient $\frac{de_i}{d\beta_i}$ in equation (8). Notice that $e_{b,i}^{(l)}$ and $d_{play,i}$ are strongly correlated, and to find out their relationship, the characteristics of network delay in stream l are assumed to follow a Pareto distribution which is defined as $F_l(x) = 1 - (g_l/x)^{\alpha_l}$. Pareto distribution parameters $\{\alpha_l, g_l\}$ can be estimated from a network trace using the maximum likelihood estimation method [6]. Given a playout delay $d_{play,i}$, the late loss probability in stream l can then be calculated as $e_{b,i}^{(l)} = 1 - F_l(d_{play,i}) = (g_l/d_{play,i})^{\alpha_l}$. With this delay distribution modelling, we can find that the gradient of the packet loss probability e_i with respect to β_i is

$$\begin{aligned} \frac{de_i}{d\beta_i} = \; & \frac{-\hat{v}_i}{d_{play,i}}\{(1-e_n^{(1)})(1-e_n^{(2)})e_{b,i}^{(1)}e_{b,i}^{(2)}(\alpha_1+\alpha_2) \\ & +e_n^{(1)}(1-e_n^{(2)})e_{b,i}^{(2)}\alpha_2 + e_n^{(2)}(1-e_n^{(1)})e_{b,i}^{(1)}\alpha_1\} \end{aligned} \quad (12)$$

Finally, we summarize the proposed multi-stream playout algorithm as below.

1. 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) by the maximum likelihood estimation method.
2. Estimate the delay mean and variance $\hat{d}_i^{(l)}$ and $\hat{v}_i^{(l)}$.
3. Use the values of (α_l, g_l) in the secant method to determine the minimizer $\hat{\beta}_i^{(l)}$ of the utility function,

$$\begin{aligned} & I_{m,i}^{(l)}(\beta_i^{(l)}) = I_d(d_c + d_{play,i}^{(l)}) + I_e(e_i(d_{play,i}^{(l)})), \\ & d_{play,i}^{(l)} = \hat{d}_i^{(l)} + \beta_i^{(l)}\hat{v}_i^{(l)} \end{aligned} \quad (13)$$

4. Set the playout delay to

$$\begin{aligned} & d_{play,i} = \hat{d}_i^{(l^*)} + \hat{\beta}_i^{(l^*)}\hat{v}_i^{(l^*)}, \\ & l^* = \arg\min\{I_{m,i}^{(l)}(\hat{\beta}_i^{(l)}), l = 1, 2\} \end{aligned} \quad (14)$$

6. EXPERIMENTAL RESULTS

Experiments were carried out to investigate the potential advantages of using the perceptual-based playout algorithm for multi-stream voice communication. Our efforts began with the simulated delay traces for use in two different voice streams. The speech database for these studies consisted of two sentential utterances spoken by one male and one female, each 8 seconds in duration and sampled at 8 kHz. The reference speech signal is encoded and then processed in accordance with the delay and loss characteristics of the trace data to generate the degraded speech. Perceived speech quality of various playout buffer algorithms were evaluated in terms of the predicted R factor, based on a combination of E-model and ITU-T P.862 PESQ algorithm. Fig. 4 also compares the average R factor evaluated by

simulation for various values of β (4, 6, and dynamic). For purpose of comparison, we also investigate a single description (SD) voice communication scheme based on the G.729 codec at 8 kbps.Compared with SD coding schemes, the better speech quality of resulting from the MD coding scheme is clearly illustrated. Table 1 compares the ratio of the full-quality speech and averge end-to-end delay under packet erasure rate = 5%. The results are given for various values of β (4, 6, and dynamic) in multi-stream voice transmission system. From this table, the β-adaptive MD scheme yielded the highest ratio of 90.66% and the lower average playout delay of 132.64 ms, compared with 85.52% and 134.51 ms for fixed $\beta = 6$ MD schemes. Subjective listening tests also indicate that the proposed β-adaptive scheme can enhance perceived speech quality for multi-stream voice transmission.

7. CONCLUSIONS

In this paper, we proposed a perceptually motivated optimization criterion and a practically feasible new algorithm for multi-stream playout buffer design. We formulate the perceptual-based buffer design as an optimization problem leading to a better tradeoff between packet loss and end-to-end delay. We also compared the perceived speech quality using the E-model methodology for playout algorithms with fixed and dynamic setting of the safety factor. Experimental results show that the proposed multi-stream playout algorithm can achieve a better delay-loss tradeoff and thereby improves the perceived speech quality.

8. ACKNOWLEDGEMENTS

This study was jointly supported by MediaTek Inc. and the National Science Council, Republic of China, under contract NSC 96-2221-E-009-031-MY3.

REFERENCES

[1] W. Jiang and A. Ortega, "Multiple description speech coding for robust communication over lossy packet networks," in *International Conference on Multimedia and Expo*, New York, USA, August . 2000, vol. 1,pp. 444–447.

[2] 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.

[3] S.B. Moon, J. Kurose, and D. Towsley, "Packet audio playout delay adjustment: Performance bounds and algorithms," *Multimedia Systems* , vol. 6, no. 1, pp. 17–28, Jan. 1998.

[4] L. Sun and E. Ifeachor, "New Models for Perceived Voice Quality Prediction and their Applications in Playout Buffer Optimization for VoIP Networks," in *Proceedings of ICC 2004* , June 2004.

[5] L. Atzori, and M.L. Lobina "Speech playout buffering based on a simplified version of the ITU-T E-Model, " *IEEE Signal Processing Letters* , June 2004.

[6] K. Fujimoto, S. Ata, and M. Murata "Adaptive Playout Buffer Algorithm for Enhancing Perceived Quality of Streaming Applications ," in *Processings of IEEE Globecom2002*, Nov 2002.

[7] Chun-Feng Wu, and Wen-Whei Chang "Perceptual Optimization of Playout Buffer in VoIP Appications, " in *Communications and Networking in China, Chinacom 2006*, Oct. 2006.

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

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

[11] R. Cole and J. Rosenbluth, "Voice over IP performance monitoring," in *Journal on Computer Communication Review*, vol. 31, no. 2, Apr. 2001.

[12] L. Sun and E. Ifeachor, "Prediction of Perceived Conversational Speech Quality and Effects of Playout Buffer Algorithms," in *Proceedings of ICC 2003* , 2003.

[13] E.K.P. Chong and S.H. Zak, *An Introduction to Optimization*, John Wiley & Sons, Inc., 2001.

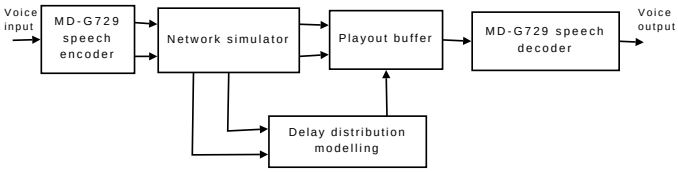

Figure 1: A block diagram of proposedmulti-descriptionvoice transmissionsystem.

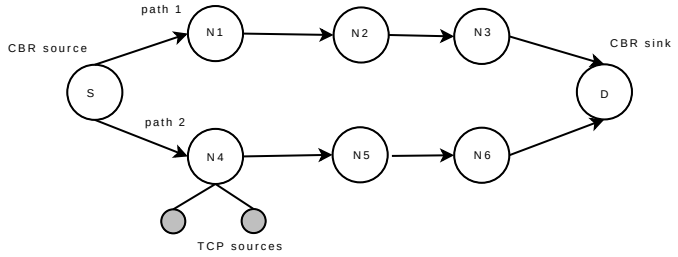

Figure 2: A multi-hop topology for network simulations. Each of the intermediate nodes has a number of TCP data sources attached.

MD schemes	$\beta = 4$	$\beta = 6$	dynamic β
$P\{\frac{S_1}{S_1 \cup S_2}\}(\%)$	82.45	85.52	90.66
Average $d_{play,i}$(ms)	133.13	134.51	132.64
R-factor	61.98	63.42	65.37

Table 1: The ratio of full quality speech and averge end-to-end delay comparison for different playout algorithms under packet erasure rate = 5% .

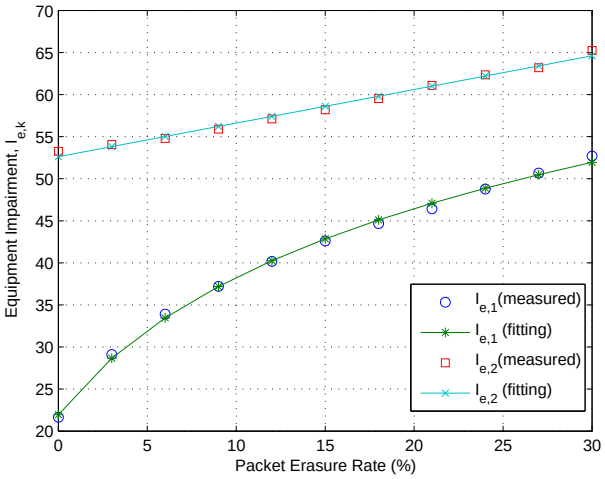

Figure 3: $I_{e,k}$ v.s. Packet erasure rate

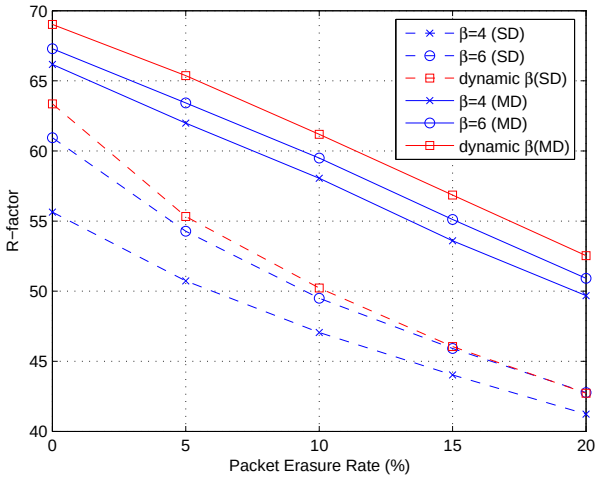

Figure 4: Performance comparison for different playout algorithms.