IEICE TRANS. INF. & SYST., VOL.E94–D, NO.5 MAY 2011

PAPER

# Perceptual-Based Playout Mechanisms for Multi-Stream Voice over IP Networks

**Chun-Feng WU**[†], ***Nonmember***, **Wen-Whei CHANG**[†a)], ***Member***, ***and*** **Yuan-Chuan CHIANG**[††b)], ***Nonmember***

**SUMMARY** Packet loss and delay are the major network impairments for transporting real-time voice over IP networks. In the proposed system, multiple descriptions of the speech are used to take advantage of the packet path diversity. A new objective method is presented for predicting the perceived quality of multi-stream voice transmission. Also proposed is a multi-stream playout buffer algorithm, together with an adaptive parameter adjustment scheme, that maximizes the perceived speech quality via delay-loss trading. Experimental results showed that, compared to FEC-protected single-path transmission, the proposed multi-stream transmission scheme achieves significant reductions in delay and packet loss rates as well as improved speech quality.

***key words:*** *voice over IP, E-model, playout buffer*

## 1. Introduction

Quality of Service (QoS) has been one of the major concerns in the context of real-time voice communication over unreliable IP networks. Iterative audio applications such as telephony and audio conferencing require high constraints on packet loss and end-to-end delay. When packet loss rates exceeds 10% and one-way delay exceeds 150 ms, the perceived conversational speech quality can be quite poor. There has been much interest in the use of packet-level forward error correction (FEC) [1] to mitigate the impact of packet losses. Most current FEC mechanisms send additional information along with the media stream so that the lost data can be recovered in part from the redundant information. In FEC schemes, however, loss recovery is performed at the cost of increased end-to-end delay. Multiple description (MD) coding [2]–[4] is another method to gain robustness by taking advantage of the largely uncorrelated loss and delay characteristics on different network paths. In MD coding, multiple descriptions of the speech are created in such a way 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. With multiple voice streams, the network delay experienced may vary with each packet depending on the paths taken by different streams and on the level of congestion along the path. 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 buffer 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 offset by the accompanying increase in the end-to-end delay. Thus, there is a need to develop playout buffer algorithms aiming at a balance between the two opposing impairment factors, end-to-end delay and late packet loss, for optimal transmission quality.

A number of adaptive playout buffer algorithms [5]–[8] have been proposed that react to changing network conditions by dynamically adjusting the playout delay. Most of them work by taking measurements on the network delays and either compressing or expanding silent periods between consecutive talkspurts. Although there are methods which focused on the delay-loss performance [5], better algorithms have been proposed along with voice quality prediction models for perceptual optimization of playout buffer [6]–[8]. The commonly used E-model [9] does not consider the dynamics of transmission impairments because it relies on the static transmission parameters such as average packet loss and average end-to-end delay. Thus, the E-model may make invalid predictions in dealing with the overall quality issues that MD transmission is focused on. For example, the E-model may only suit single-path transmission with two conceivable playout scenarios; i.e., total loss vs. no-loss of packets. A third scenario, partial loss, however, would rise with MD transmission. That is, with multiple streams sent along two paths, if packets from one path experience erasure or excessive delay, packets from the other path can often be used to conceal the lost packets. Although the partial loss is concealed, the resulting degraded playout quality may not be. In dealing with such reconstruction scheme, 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. For multi-stream voice transmission, Liang et al. [3] proposed a perceptual-based playout buffer algorithm which uses the Lagrangian cost function to trade delay versus loss. It assumes that the human perceptual experience is more strongly impaired by high latency than packet loss, and thus their algorithm uses a play-first strat-

Manuscript received June 11, 2010.
Manuscript revised December 21, 2010.

[†]The authors are with the Department of Communications Engineering, National Chiao-Tung University, Hsinchu, Taiwan.
[††]The author is with the Department of Special Education, National Hsinchu University of Education, Hsinchu, Taiwan.
a) E-mail: wwchang@cc.nctu.edu.tw
b) E-mail: ychiang@mail.nhcue.edu.tw

DOI: 10.1587/transinf.E94.D.1018

egy at the receiver; namely, that the receiver plays out the description which arrives earlier while discarding the other one most of the time. The assumption in fact follows the observation that while packet losses occurring during a voice conversation impair the intelligibility, high latency affects the interactivity between communicating parties. Research findings are available, showing delay lengths and the corresponding effects on communication; i.e., for delays less than 150 ms, conversations occur naturally, whereas at delays in excess of 150 ms conversations begin to strain and breakdown, often degenerating into simplex communications at high delay values [10], [11]. ITU-T G.114 [12] states that the generally-accepted limit for high-quality voice connection delay is 150 ms and 400 ms as a maximum tolerable limit. Although the assumption is valid, the play-first strategy needs to be examined, for it neither considers the quality degradation due to frequent switch of playout scenarios nor tries to optimize the perceived speech quality. In this work, a new playout buffer algorithm in combination with adaptive parameter adjustment scheme is introduced that automatically balances the delay, packet loss, and speech reconstruction quality for multi-stream voice transmission.

## 2. System Implementation

A block diagram of the proposed multi-stream VoIP simulation system is shown in Fig. 1. The system has four major components: MD speech coder, Internet traffic simulator, delay distribution modelling and adaptive playout buffer. The implementation procedure consisted of description generation and description transmission over two independent network paths. For description generation, the MD-G.729 based on speech packetization scheme described in [4] was used to generate two descriptions from the bitstream of the ITU-T G.729 codec [13]. G.729 is a conjugate-structure algebraic code-excited linear prediction (CS-CELP) codec for encoding narrowband speech at the rate of 8 kbps. It operates on 10-ms speech frames and each speech frame is divided into two subframes and all the parameters except the LPC coefficients are determined once per subframe. The MD-G.729 coder is designed to create two balanced descriptions; i.e., each description is of equal rate 4.6 kbps and speech decoded from either description is of similar quality. During description transmission, the best-effort nature of IP networks results in packets experiencing varying amounts of delay and loss due to different levels of network congestion. To characterize this, we used the ns-2 network simulator [14] to generate the traces of VoIP traffic for different network topologies and varying network load. Meanwhile, traces were also extended for varying link loss rates. A value ranging from 0-30% was used to simulate losses with different degrees of severity. Figure 2 shows a two path multihop network topology for our simulation, with transmission control protocol (TCP) data traffic on both paths contending simultaneously for network resources. The three nodes situated between source and destination on each path (N1 through N3 on the top path and N4 through N6 on the bottom), represent the data access points, each with a number of data sources attached, thus channelling in a large amount of incoming TCP traffic heading for different destinations. On each path a constant bit rate (CBR) voice stream is transmitted in 10-ms UDP packets at a rate of 4.6 kbps. The running time for each simulation is 15 seconds.

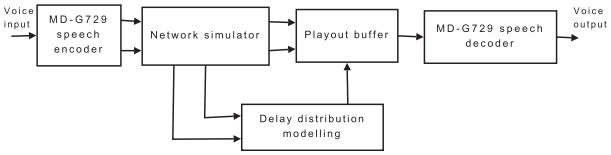


**Fig. 1** A two-channel VoIP simulation system.

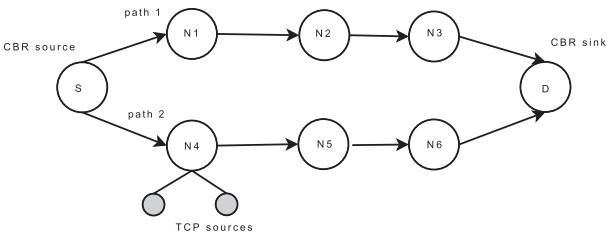


**Fig. 2** A multi-hop transmission model for network simulations.

At the receiver, a playout buffer is employed to improve the tradeoff among delay, late loss rate, and speech reconstruction quality. We focused on adaptive algorithms which adjust the playout buffer at the beginning of each talkspurt and subsequent packets of that talkspurt are played out with the generation rate at the sender. Scheduling the playout of multiple voice streams is formulated as an optimization problem on the basis of a minimum overall impairment criterion. In addition to packet loss and delay, it takes into account the dynamics of transmission impairments due to frequent switch of playout scenarios. To proceed with this, it is a prerequisite to establish a delay distribution model as it provides a direct link to late loss rate in the presence of jitter. Previous work in [8] has found that the delay characteristics of VoIP traffic can be represented by statistical models which follow Pareto, Normal and Exponential distributions depending on applications. Finally, the MD-G.729 bit stream is decoded to generate the degraded speech. In our experiments, the decoder deals with the loss of two descriptions by using the error concealment algorithm of G.729 [13], 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 [4].

## 3. Multi-Stream Voice Quality Prediction Model

Conceptually the proposed model followed the commonly

used ITU E-model [9] in defining factors that affect the perceptual quality of the MD voice transmission. As an analytical model of conversational speech quality used for network planning purposes, the E-model combines individual impairments due to the signal's properties and the network characteristics into a single R-factor, ranging from 0 to 100. In VoIP applications [15], the R-factor may be simplified as follows: $R = 94.2 - I_d - I_e$, where $I_d$ represents the delay impairment. $I_e$ is known as the equipment impairment and accounts for impairments due to speech coding and packet loss. The delay impairment can be derived by a simplified fitting process in [15] with the following 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. The E-model, originally proposed for single-stream transmission, is only applicable to a limited number of speech codecs and network conditions, since it requires time-consuming subjective tests to derive the $I_e$ model. With multiple voice streams, any subset can be used for signal reconstruction, and the transmission quality improves with the size of the subsets. In addition to delay and packet loss, a good quality prediction model should take into account the impairments due to dynamic size allocations during the speech playout.

For two-path transmission, each channel can either deliver or erase the transmitted description, so the two channels will always be in one of four possible states: no loss, loss in channel 1, loss in channel 2, and loss in both channels (packet erasure). Among them, only the speech resulting from the packet-erasure state is not affected by playout buffer operations. The receiver deals with the loss of both descriptions by using the error concealment algorithm of G.729 codec to conceal the erased packet. If, additionally, speech decoded from either MD-G.729 description is assumed to be of similar quality, we only 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 $S_k$ denote the scenario that $k$ descriptions are received before the playout time. Conditioned on the event that the packet can be restored, we let $q_k$ be the probability to play out the packet using $k$ descriptions. Formally, it is given by $q_k = P(S_k)/(P(S_1) + P(S_2))$. It is improtant to notice that quality degradation resulting from $S_1$ and $S_2$ are different perceptual experiences. For scenario $S_2$, the standard G.729 decoding process is carried out after combining the two descriptions into one bitstream. Let $I_{e,k}$ denote the equipment impairment as a result of playing out $k$ received descriptions. 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 follows:

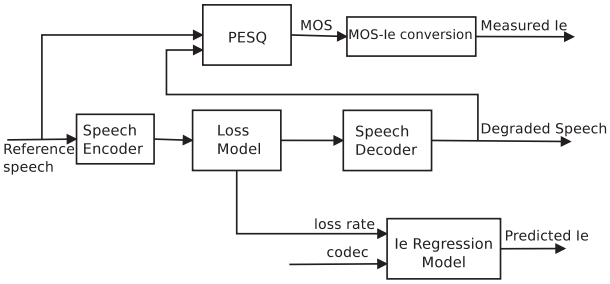


**Fig. 3** Schematic diagram for prediction of $I_e$ model.

$$I_e(e) = q_1 I_{e,1}(e) + q_2 I_{e,2}(e). \quad (2)$$

The next issue to be addressed is how to derive an equipment impairment $I_{e,k}$ corresponding to each playout scenario $S_k$. We followed the work of [16], which describes an objective method for prediction of $I_{e,k}$ regression model using the PESQ algorithm [17]. As shown in Fig. 3, each single measurement consists of three steps and is repeated several times with different transmission configurations. First, a speech sample is selected from an English speech database that contains 16 sentential utterances spoken by eight males and eight females. Each sample has a duration of 8 seconds and sampled at 8 kHz. Second, the speech sample is encoded using MD-G.729 codec and then processed in accordance with the simulated loss model to generate the degraded speech. In our experiments, the decoder deals with packet erasure by using the error concealment algorithm of G.729 [13] to conceal erased packets, while in other scenarios speech packets are reconstructed depending on how many descriptions are received by the playout deadline. Third, the reference speech and degraded speech are processed by the PESQ to obtain a mean opinion score (MOS). For each speech sample, a MOS value for one packet-erasure rate is obtained by averaging over 30 different erasure locations in order to remove the influence of erasure location. Further, these MOS values are averaged over all speech samples and then converted to a rating $R$ to give an equipment impairment value $I_{e,k} = 94.2 - R$. The R-factor can be obtained from the average MOS with a conversion formula as follows:

$$R = 3.026MOS^3 - 25.314MOS^2 + 87.06MOS - 57.336. \quad (3)$$

Figure 4 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. The $I_{e,k}$ value for zero packet-erasure rate represents the codec impairment itself. It is obvious that the speech playout resulting from $S_2$ has a lower codec impairment and has a high robustness to packet loss. By inspecting Fig. 4, we observe that our measured $I_{e,2}$ value for zero packet erasure, 21.96, is inconsistent with the ITU-published $I_e$ value, 10, for codec G.729 [9]. One possible reason for this discrepancy may lie in the codec algorithm. As the G.729 is a CELP-based codec, the use of linear predictive model of speech production can lead

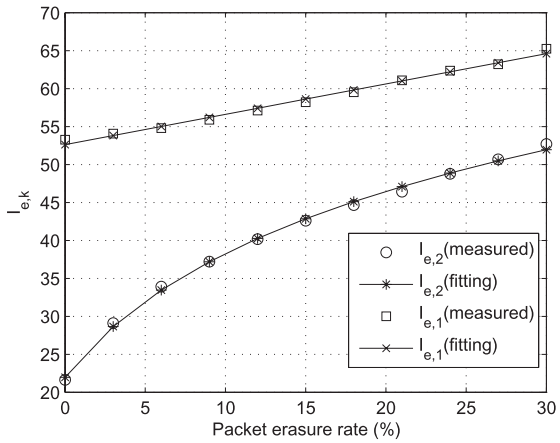


**Fig. 4** $I_{e,k}$ vs. packet erasure rate $e$.

to variations in codec performance with different talkers or languages [18]. Support for such a speculation can be found in at least two studies using the same codec [6], [19], which, in case of zero packet loss and using different speech samples from the ITU-T data set [20], rendered measured $I_e$ values of 21.14 and 17.128, respectively, similar to the value obtained for this study. From the curves, a nonlinear regression model can be derived for each $I_{e,k}$ by the least-squares data fitting method. The fitting curves are also shown in Fig. 4. The derived $I_{e,k}$ model for scenario $S_k$ has the following form: $I_{e,k}(e) = \gamma_{1,k} + \gamma_{2,k} \ln(1 + \gamma_{3,k} e)$, where $e$ is the packet-erasure rate in percentage. Our findings indicate that the regression model parameters $(\gamma_1, \gamma_2, \gamma_3)$ for $S_1$ are $(52.61, 7.52, 10)$ and $(21.96, 17.02, 16.09)$ for $S_2$.

## 4. Playout Scheduling of Multiple Voice Streams

The main attraction of multi-stream transmission arises from its flexibility to trade off different sources of impairments against each other. Waiting for the arrival of both descriptions results in lower equipment impairment $I_e$, but at the cost of higher delay impairment $I_d$. On the other hand, playing out the voice description with lower delay avoids latency, but increases the equipment impairment. Since playout scheduling aims to improve the overall conversational speech quality, which hangs on the balance between delay and packet loss, full reconstruction of both descriptions may not always be the priority if the overall impairment does not justify the extra delay from waiting. Given that, the design of a playout buffer must play around with switching between different playout scenarios in order to maximize the benefits of packet path diversity. To accomplish this goal, the proposed voice quality prediction model is applied on adaptive control of the multi-stream playout buffer. Prior to the arrival of each packet $i$, the playout delay for that packet is determined according to the past recorded delays. The playout delay of packet $i$ is denoted by $d_{play,i}$, which is defined as the time from the moment that packet is delivered to the network until it has to be played out. A packet may get lost due to its late arrival, if its network delay is larger than the playout delay.

The basic adaptive playout algorithm estimates two statistics characterizing the network delay, and uses them to calculate the playout delay as follows:

$$d_{play,i} = \hat{d}_i + \beta \hat{v}_i. \tag{4}$$

where $\hat{d}_i$ and $\hat{v}_i$ are running estimates of the mean and variation of network delay seen up to the arrival of the $i$th packet. The safety factor $\beta$ has a critical impact on the tradeoff between delay and late packet loss, which in turn influences the conversational speech quality. From (4) it can be deduced that increasing $\beta$ leads to lower late loss rate as more packets arrive in time, however the end-to-end delay increases. All of the algorithms [5]–[8] used a fixed value of $\beta$, e.g., $\beta = 4$, to set the buffer size, so that only a small fraction of the arriving packets should be lost due to late arrival. In this work, a $\beta$-adaptive algorithm is instead used to control the playout buffer so that the reconstructed voice quality is maximized in terms of delay and loss. The idea behind our algorithm is to adaptively adjust the value of $\beta$ with each incoming talkspurt, depending on the variation in the network delays.

We formulated the parameter adjustment as a perceptually motivated optimization problem and the adopted criterion relies on the use of the proposed multi-stream voice quality prediction model. Let $d_i$ be the end-to-end delay experienced by the $i$th packet, which consists of encoding delay $d_c$ and playout delay $d_{play,i}$. $e_i$ is the packet-erasure probability to lose two descriptions, no matter if the description is dropped by the network or discarded due to its late arrival, and is given by

$$\begin{aligned} e_i = {} & e_n^{(1)}(1 - e_n^{(2)})e_{b,i}^{(2)} + e_n^{(2)}(1 - e_n^{(1)})e_{b,i}^{(1)} + \\ & (1 - e_n^{(1)})(1 - e_n^{(2)})e_{b,i}^{(1)}e_{b,i}^{(2)} + e_n^{(1)}e_n^{(2)} \end{aligned} \tag{5}$$

where $e_n^{(l)}$ and $e_{b,i}^{(l)}$ represent the link loss probability and estimated late loss probability of packet $i$ in stream $l$, respectively. Now, we define an overall impairment function $I_m$ which is a function of both $d_i$ and $e_i$, with $I_m(d_i, e_i) = I_d(d_i) + I_e(e_i)$. Using (1) and (2), $I_m$ can be expressed as

$$\begin{aligned} I_m(d_i, e_i) = {} & 0.024 d_i + 0.11(d_i - 177.3)H(d_i - 177.3) \\ & + \textstyle\sum_{k=1,2} q_k I_{e,k}(e_i). \end{aligned} \tag{6}$$

where $q_1 + q_2 = 1$ and the probability to receive both descriptions is given by

$$q_2 = \frac{1}{1 - e_i}(1 - e_n^{(1)})(1 - e_n^{(2)})(1 - e_{b,i}^{(1)})(1 - e_{b,i}^{(2)}). \tag{7}$$

Our optimization framework requires an analytic expression for the packet erasure probability $e_i$ as a function of the single parameter $\beta_i$. Notice that $e_{b,i}^{(l)}$ and the playout delay $d_{play,i}$ 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 $F_l(x) = 1 - (g_l/x)^{\alpha_l}$.

The parameters of Pareto distribution $\alpha_l$ and $g_l$ can be estimated from past recorded delays using the maximum likelihood estimation method [8]. Then, the late loss probability of packet $i$ in stream $l$ can be computed as follows:

$$e_{b,i}^{(l)} = 1 - F_l(d_{play,i}) = (g_l/d_{play,i})^{\alpha_l} \tag{8}$$

This reduces the expression of the packet-erasure probability $e_i$ to be a function of the playout delay $d_{play,i}$, which in turn is a function of the parameter $\beta_i$. Its gradient with respect to $\beta_i$ is given by

$$\begin{aligned}\frac{de_i}{d\beta_i} = & \\ & \tfrac{-\hat{v}_i}{d_{play,i}}\{(1-e_n^{(1)})(1-e_n^{(2)})e_{b,i}^{(1)}e_{b,i}^{(2)}(\alpha_1+\alpha_2) \\ & +e_n^{(1)}(1-e_n^{(2)})e_{b,i}^{(2)}\alpha_2 + e_n^{(2)}(1-e_n^{(1)})e_{b,i}^{(1)}\alpha_1\}\end{aligned} \tag{9}$$

The overall impairment function $I_m$ is a function of the playout delay $d_{play,i}$ and the probability $q_k$ as well as the packet-erasure probability $e_i$. Since these parameters are all functions of the parameter $\beta_i$, the overall impairment $I_m$ is also a function of $\beta_i$, i.e., $I_m(d_i, e_i) = I_m(\beta_i)$. By differentiating it with respect to $\beta_i$, we get the following equation for the gradient:

$$I_m'(\beta_i) = c\hat{v}_i + \textstyle\sum_{k=1,2}\left\{q_k \frac{\gamma_{2,k}\gamma_{3,k}}{1+\gamma_{3,k}e_i}\frac{de_i}{d\beta_i} + \frac{dq_k}{d\beta_i}I_{e,k}(e_i)\right\}. \tag{10}$$

where

$$c = \begin{cases} 0.024, & \beta_i < (177.3 - d_c - \hat{d}_i)/\hat{v}_i; \\ 0.134, & \beta_i > (177.3 - d_c - \hat{d}_i)/\hat{v}_i. \end{cases} \tag{11}$$

$$\begin{aligned}\tfrac{dq_2}{d\beta_i} = & \tfrac{\hat{v}_i}{d_{play,i}(1-e_i)}(1-e_n^{(1)})(1-e_n^{(2)})[\alpha_1 e_{b,i}^{(1)}(1-e_{b,i}^{(2)}) \\ & + \alpha_2 e_{b,i}^{(2)}(1-e_{b,i}^{(1)})] \\ & + \tfrac{1}{(1-e_i)^2}\tfrac{de_i}{d\beta_i}(1-e_n^{(1)})(1-e_n^{(2)})(1-e_{b,i}^{(1)})(1-e_{b,i}^{(2)})\end{aligned} \tag{12}$$

Then, our general problem can be stated as follows: Given estimates of the parameters characterizing the delay distribution and $I_e$ regression model, find the optimal value of $\beta_i$ so as to minimize the overall impairment function $I_m(\beta_i)$. This task belongs to the class of set-constrained optimization problems, which can be solved efficiently by means of one-dimensional search methods [21]. For computational purposes, we applied the secant method [21] to search for the minimizer $\hat{\beta}_i$ of $I_m$ over the constraint set $\{\beta_i \in R, \beta_i > 0\}$. Starting with two initial values $\beta_i(-1)$ and $\beta_i(0)$, the iterative formula for the secant algorithm at the $j$-th iteration has the form

$$\beta_i(j+1) = \beta_i(j) - \frac{\beta_i(j) - \beta_i(j-1)}{I_m'(\beta_i(j)) - I_m'(\beta_i(j-1))} I_m'(\beta_i(j)). \tag{13}$$

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

1. Apply an autoregressive algorithm [5] to estimate the delay mean $\hat{d}_i^{(l)}$ and variance $\hat{v}_i^{(l)}$ for individual stream $l$ ($l = 1, 2$) as follows:

$$\hat{d}_i^{(l)} = \mu\hat{d}_{i-1}^{(l)} + (1-\mu)n_i^{(l)}. \tag{14}$$

$$\hat{v}_i^{(l)} = \mu\hat{v}_{i-1}^{(l)} + (1-\mu)|n_i^{(l)} - \hat{d}_i^{(l)}|. \tag{15}$$

   where $n_i^{(l)}$ is the network delay of packet $i$ in stream $l$ and $\mu = 0.998002$ is a weighting factor for convergence control.
2. At the beginning of each talkspurt, update network delay records for the past $L = 200$ packets in every stream $l$ ($l = 1, 2$), and use them to calculate the Pareto distribution parameters $(\alpha_l, g_l)$ by the maximum likelihood estimation method. Given a set of past network delays $\{n_{i-1}^{(l)}, n_{i-2}^{(l)}, \ldots, n_{i-L}^{(l)}\}$, we compute

$$g_l = \min\{n_{i-1}^{(l)}, n_{i-2}^{(l)}, \ldots, n_{i-L}^{(l)}\} \tag{16}$$

$$\alpha_l = L/\Sigma_{j=i-1}^{i-L}\log\left(\frac{n_j^{(l)}}{g_l}\right) \tag{17}$$

3. Use the values of $(\alpha_l, g_l)$ in the secant method to determine the minimizer $\hat{\beta}_i^{(l)}$ of the utility function,

$$\begin{aligned}I_m(\beta_i^{(l)}) = & I_d(d_c + \hat{d}_i^{(l)} + \beta_i^{(l)}\hat{v}_i^{(l)}) \\ & + I_e(e_i(\beta_i^{(l)})).\end{aligned} \tag{18}$$

4. Set the playout delay to

$$\begin{aligned}& d_{play,i} = \hat{d}_i^{(l^*)} + \hat{\beta}_i^{(l^*)}\hat{v}_i^{(l^*)}, \\ & l^* = \arg\min\{I_m(\hat{\beta}_i^{(l)}), l = 1, 2\}\end{aligned} \tag{19}$$

## 5. Experimental Results

A set of experimental conditions was designed for the use of artificially degraded speech samples to verify the detrimental effects estimated by the proposed $I_e$ regression model in relation to the traditional E-model. The two models, despite their agreement in including packet loss as a main impairment factor, differ in how reconstruction in conditions with partial packet losses is treated. The proposed model differentiates partial reconstruction with one description from full reconstructions with two descriptions. The three states of frame reconstruction dictated by the model are 1) the fully restored, when both descriptions are available and thus played with high quality, 2) the partially restored, when only one description is available and thus played with less than optimal quality, and 3) not restored, when both descriptions are lost and thus not playable. In contrast, the traditional model treats the full and the partial reconstruction states uniformly as the no-loss state, leaving out any differentiation of the processes involved that lead to the no-loss at the receiver end. It is thus reasonable to hypothesize that the traditional model fails to register any quality impairment due to partial reconstruction. As such, if the $I_e$'s estimated with

the two models show significant differences in their closeness to the $I_e$'s measured, then adding such a differentiation scheme into the modelling process should prove a valid approach. The speech samples considered here were one male and one female utterance. The G.729 speech codec and the proposed MD coding scheme were used sequentially, which turned each utterance into a bitstream of frames with two identical descriptions to be transmitted along separate dynamically-changing paths. At the receiver end, each utterance was artificially degraded to render two tokens, each with its own composition of frames of the three reconstruction states. Since the proposed model diverges from the traditional model by treating the loss of one packet as a separate state from either total loss or no loss, the underlying variable being manipulated in the frame composition was the rate $q_1$ of partial loss. Thus, there was a total of four (2 voices x 2 rates of one description loss) test conditions.

Table 1 lists for each condition the percentages of frames that are not restored and restored with only one description, followed by the three corresponding $I_e$'s as estimated by the traditional model, by the proposed model, and as measured then converted with PESQ. The results showed that, unlike the traditional model that yielded poorer estimations for samples containing higher percentages of one description loss, the proposed model gave estimations that are quite robust regardless of the sample frame composition. For example, given the same percentage increases from 6.84% to 22% and from 14% to 31% in the female and the male utterance respectively, the traditional model showed deviations from the measured $I_e$'s that were increased from 1.38 to 6.43 and from 4.79 to 5.9, respectively, while the proposed model yielded across conditions more stable and smaller deviations that ranged from 0.6 to 1.6. Taken together, these comparison data suggest that independent evaluation of impairments due to loss of one vs. both descriptions adds to the robustness of the proposed model.

Computer simulations were carried out to evaluate the performances given by three examples, MD1-3, of the MD voice transmission scheme, which used the 9.2 kbps MD-G.729 codec for the generation of two balanced descriptions. An FEC-protected single description (SD) transmission scheme was also tested for its comparative strength. The SD scheme applied the 8 kbps G.729 codec and performed packet-level (9,8) Reed-Soloman channel code, a condition in which an FEC packet was generated for every 8 packets and whenever any 8 of the 9 packets (8 + the resulting FEC packet) had been received over a period of time, the 8 packets were fully recovered at the receiver end. It was hypothesized that the performances of these four schemes being tested would be set apart mainly by the value of $\beta$ (fixed or dynamically changing) they each assumed during the test period, and that the best performance should come with $\beta$ values whose calculation was based on link loss, packet-erasure loss and various transmission scenarios. MD1 had a fixed $\beta = 4$, and MD2 took values of $\beta$ that were dynamically adjusted by the playout buffer according to the proposed voice quality prediction model. MD3 differed from the previous two by having its $\beta$ set following the play-first strategy proposed by Liang et al. [3]. The SD scheme, with the FEC feature, assumed a dynamic $\beta$ value as determined by the E-model [22]. 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. Both samples were encoded and then processed in accordance with the delay and loss characteristics of the trace data to degrade the speech. Figure 5 plots the perceived speech quality for the SD and the 3 MD schemes as a function of the link loss rate. As described in Sect. 2, the perceived quality was gauged by calculating the predicted average *R*-factor according to the E-model, and the link loss rate was varied from 0-30%. It can be seen that, although the quality deteriorated for all four schemes as the link loss rate increased, the three MD schemes yielded better speech quality than the SD scheme, especially at increased link loss rate. At rates slightly beyond the minimum (eg., 5%), the SD scheme, despite its FEC feature, started showing incapability of recovering the lost packets in facing link losses. Among the three MD schemes that showed three levels of dynamics in making decisions about delay, MD1, with its fixed $\beta$ value, yielded the worst quality at 0% link loss rate, yet showed better results at rates above 20% than MD3, suggesting a limitation of the Lagrangian cost function in predicting the actual perceived speech quality. The best results, as hypothesized, were obtained with the currently proposed scheme MD2, which can be attributed to its all encompassing algorithm and thus the overall function that takes into account the impairment impacts as a result of delay, packet-erasure loss and various transmission

**Table 1** $I_e$ comparison for different estimation methods.

| Speech | $e$% | $q_1$% | Traditional $I_e$ | Proposed $I_e$ | Measured $I_e$ based on PESQ |
|---|---|---|---|---|---|
| Female | 9.88 | 6.48 | 43.16 | 44.35 | 44.54 |
| | 4.93 | 22 | 34.41 | 39.48 | 40.84 |
| Male | 4.84 | 14 | 31.78 | 34.97 | 36.57 |
| | 12 | 31 | 40.37 | 45.67 | 46.27 |

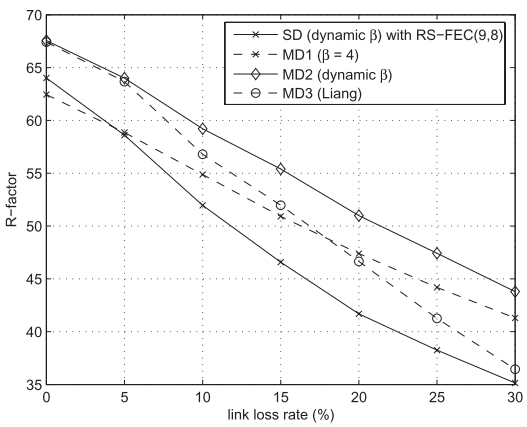


**Fig. 5** Performance comparison for different playout algorithms.

**Table 2** MOS comparison for different playout algorithms.

| Link loss rate (%) | 0 | 5 | 10 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|---|---|
| SD with RS-FEC(9,8) | 3.305 | 3.028 | 2.678 | 2.396 | 2.147 | 1.979 | 1.833 |
| MD1 | 3.226 | 3.040 | 2.832 | 2.623 | 2.439 | 2.274 | 2.128 |
| MD2 | 3.481 | 3.305 | 3.059 | 2.860 | 2.627 | 2.441 | 2.254 |
| MD3 | 3.473 | 3.288 | 2.942 | 2.677 | 2.399 | 2.126 | 1.893 |

scenarios. To elaborate further, MOS performances of various playout algorithms were examined for MD transmission with link loss rates ranging from 0% to 30%. As shown in Table 2, the results indicate that the proposed playout algorithm is preferable to other algorithms in all the tests and its performance gain tends to increase with increasing link loss rates.

## 6. Conclusions

In this paper, we proposed a perceptually motivated optimization criterion and a practically feasible new algorithm for multi-stream playout buffer design. We start by considering the perceived voice quality as a function of playout scenario, the packet erasure rate and the end-to-end delay. Adaptive scheduling of multiple streams is then formulated as an optimization problem leading to the minimum overall impairment. 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.

## Acknowledgements

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

### References

[1] J. Rosenberg, L. Qiu, and H. Schulzrinne, "Integrating pakcet FEC into adaptive voice playout buffer algorithms on the internet," Proc. IEEE INFOCOM 2000, vol.3, pp.1705–1714, Tel Aviv, Israel, March 2000.

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

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

[4] J. Balam and J.D. Gibson, "Multiple descriptions and path diversity for voice communications over wireless mesh networks," IEEE Trans. Multimed., vol.9, no.5, pp.1073–1088, Aug. 2007.

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

[6] L. Sun and E. Ifeachor, "Voice quality prediction models and their application in VoIP networks," IEEE Trans. Multimed., vol.8, no.4, pp.809–820, Aug. 2006.

[7] L. Atzori and M.L. Lobina, "Speech playout buffering based on a simplified version of the ITU-T E-Model," IEEE Signal Process. Lett., vol.11, no.3, pp.382–385, June 2004.

[8] K. Fujimoto, S. Ata, and M. Murata, "Adaptive playout buffer algorithm for enhancing perceived quality of streaming applications," Proc. IEEE Globecom, Nov. 2002.

[9] International Telecommunication Union, "The E-model, a computational model for use in transmission planning," ITU-T Recommendation G.107, March 2005.

[10] T.J. Kostas, M.S. Borella, I. Sidhu, G.M. Schuster, J. Grabiec, and J. Mahler, "Real-time voice over packet-switched networks," IEEE Netw., vol.12, no.1, pp.18–27, Jan. 1998.

[11] N. Kiatawaki and K. Itoh, "Pure delay effect on speech quality in telecommunications," IEEE J. Sel. Areas Commun., vol.9, no.4, pp.586–593, May 1991.

[12] International Telecommunication Union, "One-way transmission time," ITU-T recommendation G.114, May 2003.

[13] International Telecommunication Union, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP)," ITU-T Recommendation G.729, Nov. 2000.

[14] S. McCanne and S. Floyd, "Network simulator ns-2," http://www.isi.edu/nsnam/ns/, 1997.

[15] R. Cole and J. Rosenbluth, "Voice over IP performance monitoring," J. Computer Communication Review, vol.31, no.2, pp.9–24, April 2001.

[16] L. Sun and E. Ifeachor, "Prediction of perceived conversational speech quality and effects of playout buffer algorithms," Proc. ICC, 2003.

[17] International Telecommunication Union, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation P.862, Feb. 2001.

[18] P.A. Barrent, R.M. Voelcker, and A.V. Lewis, "Speech transmission over digital mobile radio channels," BT Technology Journal, vol.14, no.1, pp.45–56, Jan. 1996.

[19] L. Ding and R.A. Goubran, "Assessment of effects of packet loss on speech quality in VoIP," Proc. IEEE International Workshop on Haptic, Audio and Visual Environments and Their Applications, pp.49–54, Sept. 2003.

[20] International Telecommunication Union, "Objective measuring apparatus, Appendix 1: Test signals," ITU-T Recommendation P.50, Feb. 1998.

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

[22] C.-F. Wu and W.-W. Chang, "Perceptual optimization of playout buffer in VoIP appications," Proc. Chinacom, Oct. 2006.

**Chun-Feng Wu** received the B.S. degree in Mathematics from National Taiwan University and the M.S. degree in communication engineering from National Chiao-Tung University in 2003 and 2006, respectively. Currently, he is working toward the Ph.D. degree in communication engineering at National Chiao-Tung University. His research interests include multimedia communication, joint source-channel coding and wireless communication.

**Wen-Whei Chang** received the B.S. degree in communication engineering form National Chiao-Tung University, Hsinchu, Taiwan, ROC, in 1980 and the M.Eng. and Ph.D. degrees in electrical engineering from Texas A&M University, College Station, TX, in 1985 and 1989, respectively. Since August 2001, he has been a professor with the Department of Communication Engineering, National Chiao-Tung University, Hsinchu, Taiwan, ROC. His current research interests include speech processubg, joun source-channel coding and wireless communication.

**Yuan-Chuan Chiang** received her B.A. degree from National Taiwan University in 1980, in foreign languages and literature, and her M.A. and Ph.D. in 1988 and 1995, respectively, from University of Massachusetts, Amherst, both in communication disorders. She has been teaching in the department of special education in National Hsinchu University of Education, Taiwan, since 1996. Her research interests include perception and production of speech by the hearing impaired, psychoacoustics, and hearing aids.