OPTIMIZATION OF SOURCE AND CHANNEL CODING FOR VOICE OVER IP

Yicheng Huang, Jari Korhonen, and Ye Wang

School of Computing, National University of Singapore

ABSTRACT

Voice over Internet Protocol (VoIP) applications must typically choose a tradeoff between the bits allocated for Forward Error Correcting (FEC) and that for the source coding to achieve the best speech quality at a given packet loss rate. In this paper, we present a new scheme to optimize the speech quality subject to the bandwidth constraints and the packet loss rate. The scheme adopts Adaptive Multi-Rate (AMR) speech codec along with a FEC scheme based on Exclusive OR (XOR) operations. Retransmission is also taken into account if the Round Trip Time (RTT) is within a certain limit. We use a simplified E-Model as objective metric. Subjective listening tests show that our scheme improves the perceptual speech quality significantly compared to the non-adaptive baseline speech transmission system.

1. INTRODUCTION

VoIP has already shown a revolutionary impact on the telecom industry in recent years. Nevertheless, due to its inherent deficiency in QoS, it is still an interesting problem on how to allocate the limited bandwidth adaptively to achieve the best speech quality at a given channel condition. In general, for a given transmission bitrate, more bits allocated for source coding leads to higher error-free speech quality but lower robustness against packet loss. In contrast, more bits for channel coding squeezes bits for source coding with improved error robustness. Therefore, an adaptation scheme is needed to optimize the bit allocation.

Retransmissions are often considered inappropriate in interactive real-time applications due to the latency. However, some streaming systems have successfully included error recovery schemes based on a limited number of retransmission attempts. These schemes are beneficial especially when the Round Trip Time (RTT) is short. In order to cope with the special nature of interactive applications, the retransmission decision should be made dynamically during streaming.

The methods for choosing the optimal encoder mode to maintain the best quality under restricted network bandwidth has been studied extensively. Fast algorithms have been proposed for video streaming in [3, 4]. In [5, 7, 8] the impact of the source behavior, the path characteristics and the receiver behavior on video streaming has been addressed.

Optimization problem is even harder for VoIP applications than video streaming due to lack of commonly accepted computational metrics to evaluate the quality distortion in speech in a similar manner as Peak Signal-to-Noise Ratio (PSNR) used for video. This problem was first addressed at a general level in [6].

E-Model [11] is an objective method for measuring speech quality, standardized by ITU-T. The latest version has been updated in 2003. It is especially suitable for VoIP, since the model analyses network-related factors, such as frame loss rate and delay. In [9] the authors proposed an optimization scheme based on E-model. Their system focuses on the delay characteristics and enables switching between different speech codecs. Some improved methods have been proposed in [2, 10].

In this paper, we present a new scheme to optimize the speech quality in a VoIP application, using codec mode switching, FEC and retransmissions. Conceptually, our work is similar to [2]. However, we have simplified E-Model for the optimization process. Furthermore, we have included the selective retransmission scheme in our framework. The proposed framework has been evaluated using network simulations and subjective listening tests.

2. SYSTEM DESCRIPTION

2.1 System Outline

The system is outlined in Figure 1. The system mainly includes source coding, channel coding and adaptation modules, which are explained in this section.

AMR speech codec has been developed by ETSI and adopted by the 3rd Generation Partnership Project (3GPP). The codec is based on the Algebraic Code-Excited Linear Prediction algorithm (ACELP) with frame length of 20 ms and sampling rate of 8 kHz. It provides good quality for encoding narrow band speech signals at different bitrates (modes) from 4.75 kbps up to 12.2 kbps. Higher bit-rate associates with higher quality and the mode can be selected for every frame separately.

Optimization module decides the optimal AMR and FEC modes for a set of four AMR speech frames (80ms). In addition, the module is responsible for triggering

0-7803-9332-5/05/$20.00 ©2005 IEEE

selective retransmissions for lost packets. For simplicity, each packet contains only one media frame or FEC frame.

Since there are various AMR and FEC modes available, optimal combination is selected according to the current network conditions. We assume that the bandwidth constraint is known a priori and the packet loss rate is independent on the transmission rate. According to the decision, *AMR Encoder* and *XOR FEC* modules encode the next four speech frames and the related FEC frames, and then transmit them to the network along with the retransmitted frames chosen by *Retransmission Selector* module.

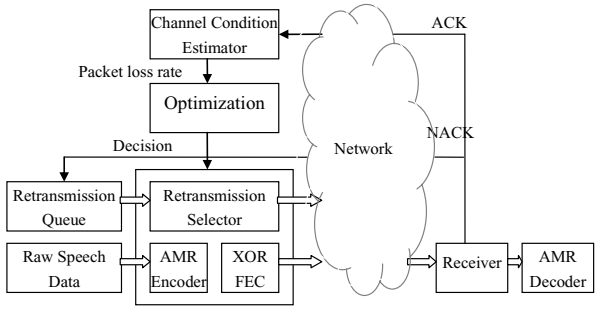

Figure 1. The proposed system outlined.

The receiver end has a playback buffer to compensate the network jitter and to reorder packets arriving in wrong order. The module is also responsible to sending ACKs/NACKs messages to the sender. Feedback information is used to update the both the network status estimated by the *Channel Condition Estimator* and the lost packets in the *Retransmission Queue*. *Channel Condition Estimator* employs Gilbert model to estimate the current packet loss rate and informs the *Optimization* module. ACK messages contain timing information that is used to estimate the RTT, which is needed to make the decision of retransmission for the lost frames.

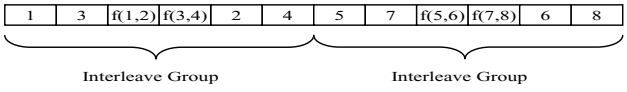

Figure2: An example packet sequence. Media frames are denoted by an ordinal number, FEC frames are denoted as f(i,j), where XOR is applied over frames i and j.

FEC is needed to include redundant information to recover missing media packets. Several methods exist for implementing FEC. Considering simplicity and efficiency we have adopted XOR operation for generating the FEC data. Our system uses four different FEC modes. Mode 0 does not provide any protection. Modes 1 and 2 apply one FEC frame to every four or two media frames, respectively. Mode 3 uses simple duplication of the original speech frames. The redundancy overhead and residual packet loss rate are shown in table 1. Interleaving is adopted to mitigate the harmful effects of burst packet losses. An example of packet arrangement in FEC Mode 2 is shown in Figure 2.

Table I: Redundancy overhead and residual packet loss rate (PLR) in different FEC modes

FEC mode	Redundancy	Residual PLR
0	0%	p
1	25%	$p(1-(1-p)^4)$
2	50%	$p(1-(1-p)^2)$
3	100%	p^2

2.2 Optimization Algorithm

E-model is a computational model used for estimating the speech quality. The model takes network related factors, such as delay and packet loss rate, as input and gives a quality estimate R as output. R ranges from 0 to 100, indicating speech quality from the worst to the best. Mean Opinion Score (MOS) is commonly used to evaluate audio quality in subjective listening tests, and the value of R can be translated into MOS using the equation (1):

$$MOS = 1 + 0.035R + 7 \cdot 10^{6} R(R-60)(100-R) \quad (1)$$

In the E-Model, R is calculated by equation (2):

$$R = 93.2 - I_d - I_e \quad (2)$$

I_d is the delay impairment score. In [1], the mapping from delay to I_d is determined by the type of the voice streaming and echo loss factors. Generally speaking, the longer the delay, the larger I_d is. However, in practical applications delay is typically constant, equal to the pre-buffering time in the jitter buffer. In addition, the perceptual quality degradation caused by delay depends on factors that are difficult to estimate computationally, such as the pace of interaction. This is why we have considered I_d as constant and omitted it in computations to simplify the E-model.

I_e is the loss impairment score. In the original E-Model there was no direct mapping from the packet loss rate to I_e available for the AMR codec. However, the problem has been addressed in [2] and an approach to estimate I_e has been proposed. Following this description, we measured MOS for eight AMR modes under various packet loss rates from 0 to 30%. The resulting MOS scores were converted into R using equation (1). Since the test was performed locally in a desktop PC, we may assume that I_d is zero. Therefore, the equation (2) can be applied to achieve values of I_e versus the packet loss rate. The test was repeated for each eight mode of AMR codec and the

results are shown in Figure 3. The results are similar to the results given in [2].

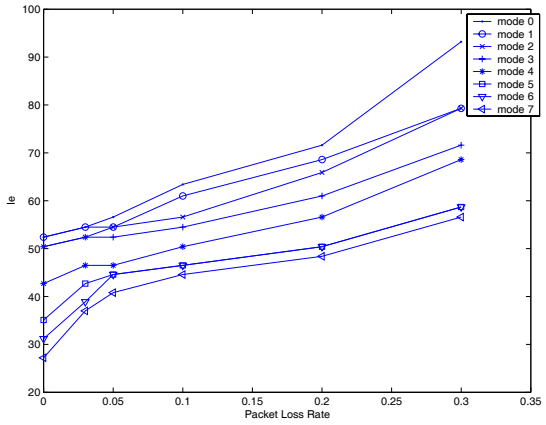

Figure 3. I_e versus packet loss rate, computed from MOS test results

In the proposed optimization scheme the constraints are the limited bandwidth and observed RTT. The objective is to optimize the expected speech quality using the E-Model. The problem can be formulated as follows.

Given B as the maximum bandwidth for the application, we assume that there are n_c new frames to be transmitted and n_r old frames needed to be retransmitted in the sender queue. The AMR mode is denoted as m_i for the new frames (i=0..n_c) and mr_i for the old frames in the retransmission queue (i=0.. n_r). In addition to the eight AMR modes (0..7), the mode 8 indicates no retransmission. The bandwidth requirement for each frame is denoted as b_i (new frames) and br_i (old frames) and the FEC mode as f_i and fr_i, respectively. Based on the prevailing packet loss rate and the FEC mode chosen (see Table 1), we may also estimate the packet loss probability for each frame and mark it with p_i (new frames) and pr_i (old frames). In addition, we denote the expected quality with Q_i and Qr_i. The quality estimate can be determined using the equation (2), where I_d can be solved when the end-to-end delay is known, and I_e, can be solved when the frame loss probability p and AMR mode m are known. If the AMR mode 8 is chosen for a retransmitted frame, quality Q is zero. Because the impact of delay is considered constant, the function for resolving the quality Q can be simplified in the form of equation (3) by removing I_d from equation (2):

$$Q(m, f) = 93.2 - I_e \tag{3}$$

$$\max Q = \sum_{i=0}^{n_c} Q(m_i, f_i) + \sum_{i=0}^{n_r} Q(mr_i, fr_i) \tag{4}$$

$$\sum_{i=0}^{n_c} b_i + \sum_{i=0}^{n_r} br_i < B \tag{5}$$

The constraint problem can be formulated in the form of equation (4) that is a subject to the condition (5).

The constrained optimization problem can be converted to an unconstrained optimization problem by the Lagrange multiplier technique. The Lagrangian cost function can be given by equation (6):

$$J = \sum_{i=0}^{n_c} (D_i + \lambda \cdot b_i) + \sum_{i=0}^{n_r} (Dr_i + \lambda \cdot br_i), \tag{6}$$

where λ is the Lagrange multiple and D_i and Dr_i show the distortion of the speech quality (D=100-Q). The formula can be solved by following the techniques described in [3, 4].

In practice, this problem can still be simplified, because the AMR and FEC modes do not need to be calculated for frames in the retransmission queue. No FEC is applied and only two AMR modes are available for the retransmitted frames, namely the mode chosen for the first transmission and the special mode 8 (i.e., no retransmission). Because the optimization algorithm is applied for four speech packets at a time only, the search space is reasonable and the algorithm can be used in real-time. The optimization algorithm can be summarized as follows:

1) Use *Channel Condition Estimator* to estimate the packet loss rate p according to the feedback.

2) If there are packets in the *Retransmission Queue*, use the estimated RTT and known pre-buffering delay to evaluate whether these packets can meet the deadline at the receiver side. If the packet is out-of-date, delete it from the *Retransmission Queue* directly.

3) Exhaust all the possible combinations of m_i, f_i and mr_i. Use f_i to compute p_i as shown in Table 1. Solve I_e for each packet using the parameters m_i, f_i, mr_i and packet loss rate p.

4) Calculate Q_i and Qr_i using Equation (3).

5) Sum all of Q_i and Qr_i, resulting in total quality Q. Choose the combination with the maximum value for Q.

6) Delete the frames that has been retransmitted and the packets with I_d higher than 80 from the *Retransmission Queue*.

3. SIMULATION RESULTS AND DISCUSSION

We have used a peer-to-peer network configuration for the experiments. Four different error pattern files have been used to simulate packet losses in the network, with packet loss rates from 3% up to 20%. The pre-buffering delay is set to 300ms and three different RTTs have been

simulated, 200ms, 400ms and 600ms. In the simulations the bandwidth ranges from 5bps to 29bps. The results are evaluated by five listeners, in a testing setup similar to that in [12].

In the baseline system, we have used constant AMR and FEC modes during the conversations. The AMR mode with largest bitrate has been selected, considering the predefined bandwidth limitation and the FEC mode. We have simulated all the suitable FEC modes for each bitrate. The baseline system does not use retransmissions.

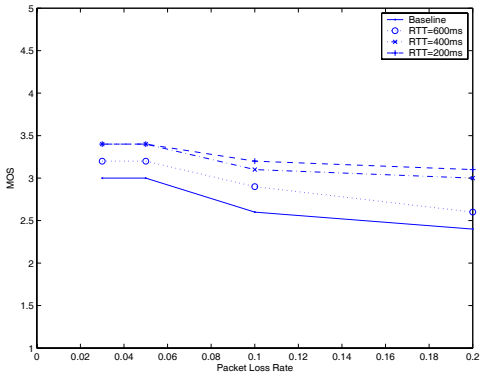

Figure 4: Comparison results at the bandwidth of 15 kbit/s.

Then we compared the proposed scheme with the baseline. In each case we chose the best performer out of the four baseline alternatives for the comparison. Figure 4 shows the results when the bandwidth is 15 kbit/s. The proposed scheme performs always equally or better than the baseline, even when there are no retransmissions used (RTT=600ms). This shows clearly the advantage of the adaptive system. When the bandwidth is sufficient, the baseline system and the proposed system without retransmissions perform equally well, because in this case both systems use AMR mode 8 and FEC mode 3. When retransmissions are allowed (RTT=400 ms and RTT=200 ms), the proposed scheme outperforms the baseline counterpart clearly.

Retransmission mechanism reduces the residual packet loss rate significantly, especially when the packet loss rate is high. Our simulation results show that single retransmission attempt together with FEC is sufficient in most cases. Multiple retransmission attempts are therefore not necessarily needed.

4. CONCLUSIONS

In this paper a scheme to adaptively optimize the VoIP speech quality under constraint bandwidth and varying packet loss rate is proposed. Simplified E-model is used as an objective metric for predicting the speech quality. The results from practical experiments show that our scheme achieves better speech quality compared to a non-adaptive streaming system which uses the best-performing combination of speech coding and FEC parameters.

When the RTT is within certain limit, it is reasonable to employ a selective retransmission mechanism to recover lost packets. Although our system uses AMR speech codec and XOR operation, the proposed optimization scheme can be extended to other speech codecs and FEC paradigms as well.

11. REFERENCES

[1] N. Kitawaki, and K. Itoh, “Pure Delay Effects on Speech Quality in Telecommunications,” IEEE Journal on Selected Areas in Communications, vol. 9, pp. 586-593, May 1991.

[2] J. Matta, C. Pepin, K. Lashkari, and R. Jain, “A Source and Channel Rate Adaptation Algorithm for AMR in VoIP using the Emodel,” NOSSDAV, pp. 92-99, Monterey, USA, Jun 2003.

[3] J. Lee, and B. W. Dickinson, “Rate distortion optimized frame type selection for MPEG encoding,” IEEE Trans. Circuits Syst. Video Technol, vol. 7, pp. 501-510, Jun 1997.

[4] Y. Shoham, and A. Gersho, ”Efficient bit allocation for an arbitrary set of quantizers,” IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 36, pp. 1445-1453, Sep 1988.

[5] D. Wu, Y.T. Hou, B. Li, W. Zhu, Y,Q. Zhang, and H. J. Chao, “An End-to-End Approach for Optimal Mode Selection in Internet Video Communication: Theory and Application,” IEEE Journal on Selected Areas in Communications, vol. 18, pp. 977-995, Jun 2000.

[6] J. Bolot, and D. Towsley, “Adaptive FEC-Based Error Control for Internet telephony,” IEEE Infocom, vol. 3, pp. 1453-1460, New York, USA, Mar 1999.

[7] J. Lee, and B. W.Dickinson, “Rate-Distortion Optimized Frame Type Selection for MPEG Encoding,” IEEE Trans. on Circuits and System for Video Technology, vol. 7, pp. 501-507, Jun 1997.

[8] M. Gallant F. Kossentini, “Efficient Scalable DCT-Based video Coding at Low Bit Rates,” International Conference on Image Processing, vol. 3, pp. 782-786, Kobe, Japan, Oct 1999.

[9] M. T. Gardner, V.,S., Frost, and D. W. Petr, “Using Optimization to Achieve Efficient Quality of Service in Voice over IP Networks,” Performance, Computing, and Communications Conference, vol. 3, pp. 475-480, Phoenix, USA, Apr 2003.

[10] W. Jiang, H. Schulzrinne, “Comparison and Optimization of Packet Loss Repair Methods on VoIP Perceived Quality under Bursty Loss,” NOSSDAV, Miami, USA, pp. 73-81, May 2002.

[11] ITU-T Recommendation G.107, “The E-model, a computational model for use in transmission planning,” 2003.

[12] Y. Wang, W. Huang, J. Korhonen, “A Framework for Robust and Scalable Audio Streaming,” ACM Multimedia, pp. 144-151, New York, USA, Oct. 10-16, 2004