SOURCE-DRIVEN PACKET MARKING FOR SPEECH TRANSMISSION OVER DIFFERENTIATED-SERVICES NETWORKS

Juan Carlos De Martin

IRITI-CNR, Politecnico di Torino
C.so Duca degli Abruzzi 24, I-10129 Torino, Italy
E-mail: demartin@polito.it

ABSTRACT

We present a source-driven approach to packet marking for speech transmission over packet networks implementing the Differentiated Services model. Packets generated by the speech coder are examined: if deemed perceptually critical, they are marked as *premium* and sent on a "virtual wire;" otherwise, they are sent as regular best-effort traffic. Applied to speech coded with the ITU-T 8 kb/s speech coding standard G.729, the proposed source-driven packet marking scheme outperforms source-transparent techniques and provides clearly better perceptual quality than the unprotected case sending as little as 1/5 of the coded bitstream as premium traffic. Audio samples are available at http://demartin.polito.it/icassp2001/.

1. INTRODUCTION

The Differentiated Services (DiffServ) architecture [1] is one of most promising proposals that have recently been made to introduce Quality-of-Service guarantees in IP networks. In this architecture, packets are classified and marked to receive a specific forwarding behavior on nodes along their path. In the simplest case, a 1-bit marking scheme [2] defines two classes: a *premium* class, with, typically, low-delay and low or no losses, and a regular best-effort class, as in the current Internet. In this scenario, delay- and losses-sensitive traffic, such as interactive speech, would be transmitted over the premium bandwidth; less critical data as best-effort.

In current carrier-grade networks, speech traffic is usually marked and transmitted as premium in its entirety. When properly engineered, this approach delivers toll, or nearly-toll, speech quality to end users.

Premium bandwidth, however, is a limited resource. The growth of speech traffic over data networks threatens to saturate its availability in corporate as well as in carrier networks rather quickly.

If only a fraction of each speech flow could be marked as premium and the rest were sent as best-effort, the load on the premium bandwidth would be reduced. To maximize perceptual end-user quality, the packets marked as premium should be the most perceptually relevant. Current approaches to packet marking, however, are usually source transparent. In [3], for instance, adaptive packet marking delivers soft bandwidth guarantees by randomly marking a certain share of the packets of a flow. Although simple, this approach does not exploit the fact that in speech transmission certain packets are more perceptually important than others.

We propose a source-driven approach to the marking of packets containing compressed speech. Speech packets are marked depending on the estimated distortion that their loss would introduce at the decoder and the desired level of perceptual quality of service.

The paper is organized as follows. In Section 2, the source-driven approach to packet marking is explained. In Section 3, the approach is presented for the specific case of the ITU-T 8 kb/s speech coding standard G.729 [4], one of the most widely used speech coders in Voice over IP applications. Results of formal A-B listening tests comparing the proposed method to current techniques as well as to other test conditions are presented in Section 4. Finally, conclusions are presented in Section 5.

2. SOURCE-DRIVEN PACKET MARKING

2.1. Overview

Let us assume that a 1-bit Premium Service DiffServ architecture is adopted (it is straightforward to generalize this example to the case of more than two classes): speech packets are transmitted either on a low-delay, no-loss "virtual wire" (a concept recently proposed by Jacobson et al. [5]) or on a regular best-effort network subject to potentially unbounded delays and packet losses. Figure 1 shows packet classification and marking for such kind of architecture. The packet classifier can, in principle, accept input from the network. The network, for instance, may periodically report current loss and delay statistics to the classifier, which may then

adapt its thresholds to optimize perceptual QoS, network usage or the desired trade-off between the two.

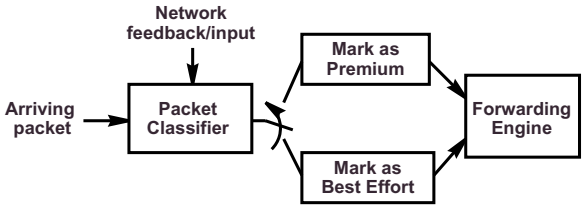

Fig. 1. 1-bit packet marker.

Packet marking for speech (or, in general, multimedia) transmission over DiffServ networks is usually accomplished by marking as premium the entire flow. Premium bandwidth is, therefore, devoted to real-time transmission, and when no more bandwidth is available, service is denied or degrades without control.

Instead of assigning all packets of a given speech flow either to the premium class, as done in most carrier-grade Telephony over IP systems, or to the best-effort class, as is currently the case for most Internet-based Voice over IP services, packet classification and marking can be performed on a packet-by-packet basis. Specifically, each speech packet can be analyzed and assigned to one class or the other depending on its perceptual importance. To do so, the packet marker needs to interpret the semantics of the payload and estimate the perceptual impact of the packet at the decoder.

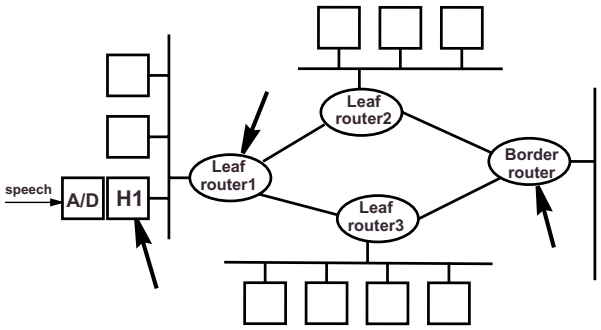

Fig. 2. Possible placements of a source-driven packet marker.

The packet marker may also act as a function of the input speech signal itself; in that case, however, packet marking can be accomplished only in the network node originating the speech flow (in Figure 2, host H1). Classification based on compressed speech alone, instead, may be accomplished at different points in the network (leaf router, border router, elsewhere), permitting a more flexible system architecture.

From a complexity point of view, source-driven marking is best done at the speech encoder. Packet classification, in fact, can be easily generated as a by-product of the encoding operation at little or no extra cost in terms of computation.

2.2. Distortion-Based Marking

The perceptual importance of a packet can be expressed in terms of the distortion that would be introduced by its loss. The optimal measure of distortion would be to compare speech decoded using the correct parameters and speech decoded using the parameters estimated by the frame erasure concealment technique. However, absent an undisputed objective distortion measure, we will look at distortion in the parameters space, an approach which has also the advantage of being less complex. To do so, the packet marker needs to:

1. decode the speech parameters, $\mathbf{P}$;
2. replicate the behavior of the decoder in presence of a frame erasure and generate estimates of the erased parameters, $\mathbf{P}'$;
3. compute distortion measures, $\mathbf{D}$, between original parameters and corresponding estimates.

The marking decision depends on this set of parametric distortions. Ideally, a function directly mapping distortions to subjective quality should be used. Absent that, the algorithm will be based on psychoacoustics knowledge and listening tests. To a certain extent, detection of perceptually critical frames is similar to the identification of anchor frames in variable frame rate speech transmission [6].

Figure 3 shows the block diagram of the proposed distortion-based marking scheme.

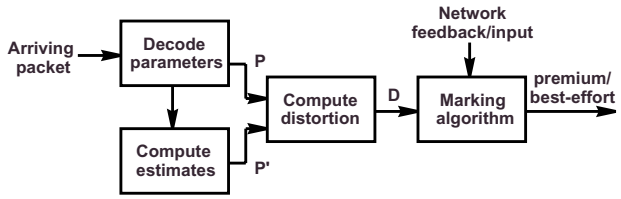

Fig. 3. Block diagram of source-driven packet marking.

Regarding step 2, the generation of the estimates, assumptions about the current state of the decoder need to be made. For low levels of frame erasures it probably suffices to assume that the previous frame has been correctly received. More complex models, however, will take into account the probability that one or more frames in the past have been lost. This can be accomplished with, for instance, a Markov model of memory up to 6–8 frames. In this case,

the computation of the distortion would generate a range of values, one for each possible state of the model. The decision will then be made on the product $p_i D_i$, where p_i and D_i are the probability of being in state i and its associated distortion, respectively.

3. PACKET MARKING OF ITU-T G.729 SPEECH

We chose to test source-driven packet marking using speech coded with the ITU-T 8 kb/s speech coding algorithm – ITU-T G.729. The G.729 decoder includes a standard frame erasures concealment technique, which we chose not to modify.

3.1. Distortion Measure

For the i-th frame, the distortion measure module extracts and decodes

1. the spectral-envelope information (LSF's), $\mathbf{L}^i$;
2. the adaptive-codebook indices, $\mathbf{P}^i$;
3. the adaptive-codebook gains, $\mathbf{AG}^i$;
4. the fixed-codebook gains, $\mathbf{FG}^i$.

It then computes the estimates that would be generated by the decoder if the frame were declared lost, assuming, for simplicity's sake, that the previous frame has been successfully received:

1. $\hat{\mathbf{L}}^i = \mathbf{L}^{i-1}$,
2. $\hat{\mathbf{P}}^i = \mathbf{P}^{i-1} + 1$,
3. if frame is voiced: $\widehat{\mathbf{AG}}^i = 0.9\mathbf{AG}^{i-1}$, $\widehat{\mathbf{FG}}^i = 0$;
4. if frame is unvoiced: $\widehat{\mathbf{FG}}^i = 0.98\mathbf{FG}^{i-1}$, $\widehat{\mathbf{AG}}^i = 0$.

For each set of parameters, the distortion between original and reconstructed parameter (vector or scalar) is computed: *spectral distortion* in dB for the spectral envelope, *percentage difference* for the adaptive codebook indices, *difference* in dB for the gains.

Figure 4 shows spectral distortion for a segment of male speech as a function of time. Note the peaks at the beginning of each talk spurts: in times of rapid transition, the concealment technique, which assumes temporal continuity of the spectral envelope, performs poorly.

The set of distortion values is then passed to the marking algorithm.

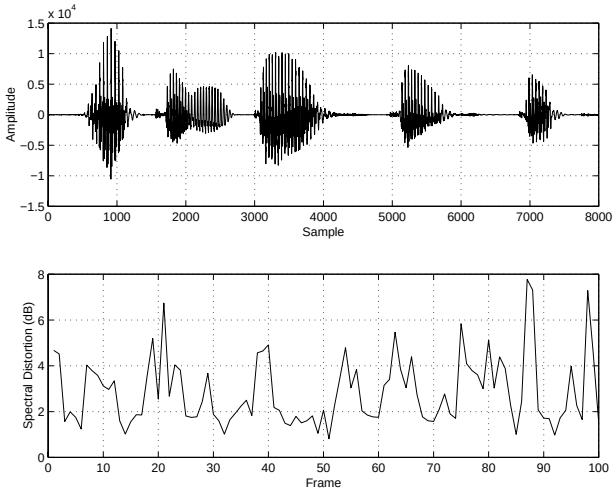

Fig. 4. Spectral distortion as a function of time.

3.2. Marking Algorithm

In general, the marking algorithm will depend on the share of traffic that we want to mark as premium and/or the desired level of perceptual quality of service generated at the decoder. Such constraints may be determined at design time, or made dependent on instantaneous network conditions, in which case packet marking would be a function of both the transmitted speech signal and the network status.

In our tests, low-energy packets were marked "best effort" without further inspection. If the overall energy is low enough, in fact, distortion in other parameters is hardly perceptible, if at all.

The rest of the marking algorithm was closely matched to the standard concealment technique of G.729. Specifically, if the missing frame is classified as voiced (voicing is inherited from the last correctly received frame), the only parameters employed by the concealment techniques are adaptive-codebook indices and gains, plus spectral envelope information. If the distortion between original and estimated parameter for any of those three sets was above a given threshold, the packet was declared premium. To achieve a share of premium packets of about 20%, the following thresholds were used:

- adaptive-codebook index difference $> 20\%$;
- adaptive-codebook gain difference > 5 dB;
- spectral distortion > 4 dB.

If the missing frame, instead, is declared unvoiced, only fixed-codebook gains and spectral envelope information are used. In this case, the thresholds were:

- fixed-codebook gain difference > 5dB;
- spectral distortion > 4 dB.

4. SUBJECTIVE TEST RESULTS

We conducted formal subjective listening tests to assess the performance of the proposed marking scheme. This evaluation consisted of A-B comparison tests with 12 sentence pairs each, uttered by both male and female speakers. The test material was flat filtered clean speech taken from the NTT Multi-lingual Speech Database. The material was encoded using the ITU-T G.729 floating-point reference software. The pairs were randomized and presented to eight different listeners between the ages of 26 and 41, all using headphones in a controlled environment. In all experiments, the packet size was 10 ms, i.e., each packet contained one G.729 output frame.

The first experiment compared the proposed source-driven marking algorithm with random marking. In both cases, the same share (about 19%) of packets was marked as premium. Subsequently, packets were subject to 5% random frame erasure. As reported in [7], in 5% frame erasures, the perceptual quality of G.729 drops to 3.4 MOS (as opposed to about 4.0 in error-free condition). The first three bars in Figure 5 shows that source-driven marking was clearly preferred over source-transparent marking.

The second and the third experiments were conducted to obtain some indications about the absolute perceptual quality of speech source-driven marked and then subject to 5% frame erasures (as in experiment 1). The reference for the second experiment was error-free speech. The strong preference for the reference condition indicates that the concealment technique could not make the loss of several best-effort packets inaudible. The reference for the third experiment, instead, was speech in 5% frame erasures, no marking. In this case, the results show a clear preference for source-driven marked speech, as expected. The third experiment indicates that delivering as little as 1/5 of the overall bitstream over premium bandwidth generates a significant improvement of the perceptual quality.

Considering that the mean opinion score of the reference conditions of experiment 2 and 3 are usually regarded to be 4.0 and 3.4, respectively, it can be surmised that with source-driven marking (premium share of about 19%), the MOS score could be in the range of 3.7. Additional experiments, however, are needed in order to test performance for different values of the premium share, different frame erasures conditions and more advanced versions of the marking algorithm.

5. CONCLUSIONS

We have presented a source-driven approach to marking speech packets for networks offering differentiated services. A specific technique for speech coded with ITU-T 8 kb/s standard G.729 was implemented and tested. Formal A-B listening tests showed that source-driven packet marking outperforms source-transparent techniques and provides clearly better perceptual quality than the unprotected case sending as little as 1/5 of the bitstream as premium traffic.

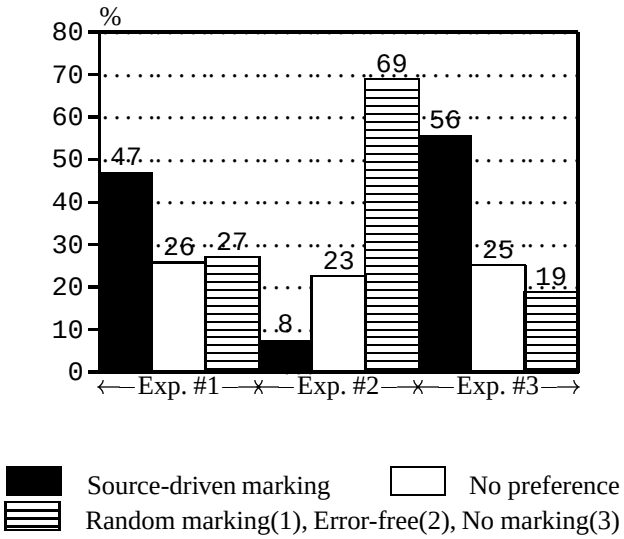

Fig. 5. Results of A-B listening tests.

6. REFERENCES

[1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," *RFC 2475*, December 1998.

[2] K. Nichols, V. Jacobson, and L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet," *RFC 2638*, July 1999.

[3] W. Feng, D.Kandlur, D. Saha, and K. Shin, "Adaptive Packet Marking for Providing Differentiated Services in the Internet," in *Proc. ICNP'98*, Austin, Texas, October 1998, pp. 108–117.

[4] R. Salami et al., "Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 116–130, March 1998.

[5] V. Jacobson, K. Nichols, and K. Poduri, "The 'Virtual Wire' Per-Domain Behavior," *Internet Draft draft-ietf-diffserv-pdb-vw-00.txt*, July 2000, work in progress.

[6] V.R. Viswanathan, J. Makhoul, R.M. Schwartz, and A.W. Huggins, "Variable Frame Rate Transmission: A Review of Methodology and Application to Narrow-Band Speech Coding," *IEEE Transactions on Communications*, , no. 4, pp. 674–686, April 1982.

[7] Mark E. Perkins, Keith Evans, Dominique Pascal, and Leigh A. Thorpe, "Characterizing the Subjective Performance of the ITU-T 8 kb/s Speech Coding Algorithm – ITU-T G.729," *IEEE Communications Magazine*, pp. 74–81, September 1997.