Classifying VoIP µ-law Packets in Real-Time

Christian Hoene[1], Till Wimmer[2]
[1]RI, Universität Tübingen, D-72076 Tübingen
[2]TKN, Technische Universität Berlin, D-10587 Berlin
hoene@uni-tuebingen.de|wimmer@tkn.tu-berlin.de

Abstract

We present an algorithm, which classifies µ-law coded VoIP packets in real time. It is based on our algorithm calculating frame importance off-line, but uses shorter speech segments, an analysis-by-synthesis approach, and a newly developed perceptual evaluation algorithm called PESQlight. Altogether, we are able to reduce the complexity and algorithm delay significantly. The real-time classification has a correlation of up to R=0.63 compared to the reference, off-line algorithm.

1. Introduction

Packet loss significantly decreases the quality of narrow-band voice communications. If a speech frame is lost, the receiver needs to extrapolate the last successful received frame to limit the impact of the lost frame. Such algorithms are known as packet loss concealment (PLC). Nowadays, they are often standardized and part of the decoder. A lost frame causes the current speech period to become distorted as the receiver's packet loss concealment cannot fully reconstruct the lost frame. Thus, the concealed frame differs from the sent frame and hence introduces a so called *loss distortion*.

In a previous publication we presented an off-line measurement procedure, which measured distortion after losing a VoIP packet [1]. We have validated this procedure with formal listening-only tests in [2]. Applying the knowledge about frame importance, both simulations and informal listening-only tests show that only a fraction of all active speech frames need to be transmitted if (at least) speech intelligibility is to be maintained [3]. Thus, significant transmission gains can be achieved by transmitting only the packets having importance and not all active speech frames.

In this publication we present a packet classification algorithm, which measures the importance of µ-law coded speech frames in real-time. It applies a technique called analysis-by-synthesis [4] to predict the loss distortion. The algorithm, *PESQlight*, is based on a new, complexity reduced version of the ITU P.862 PESQ algorithm. Our packet classification outperforms the previously presented approaches in terms of prediction accuracy, at the cost of higher computational complexity.

2. Background and Previous Work

The *Perceptual Evaluation of Speech Quality (PESQ)* algorithm predicts human rating behaviour for narrow band speech transmission. It compares an original speech fragment with its transmitted and thus degraded version to determine an estimated MOS value, which ranges from 1 (bad) to 5 (excellent).

The ITU G.711 codec can be applied for compressing a narrow-band telephone audio signal to a rate of 64 kbps with a sample rate of 8 kHz and 8 bits per sample. Its µ-law mode is used in this work. The ITU standardized a packet loss concealment (ITU G.711 Appendix I), which limits the impact

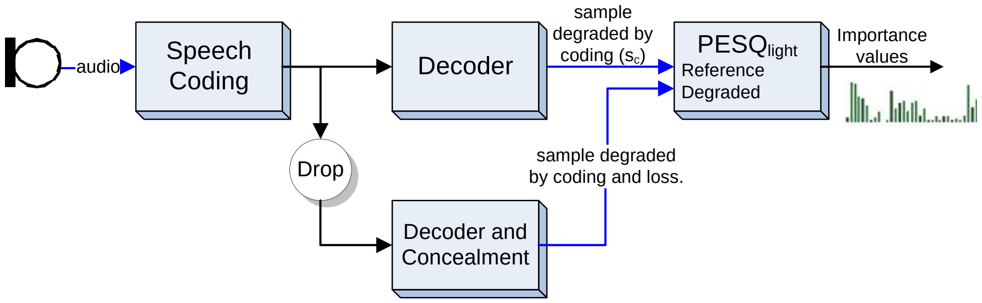

Figure 1: Schematic of a sender-side calculation of the importance values.

of transmission losses. The PLC algorithm works on frame sizes of 10 ms: If a frame is lost, the last successfully received pitch period is repeated.

A classic example of classifying speech is the voice activity detection. Periods of active speech alternate with periods of background noise. Later periods are less important for the perceptual quality of speech transmission. For active frames, Petr et al. [5] suggested a method to classify speech frames depending whether they are non-initial voiced, non-initial fricative speech, and all others. De Martin [6] has proposed an approach to mark G.729 codec packets using an analysis-by-synthesis technique. Sanneck [7] analyzed the temporal sensitivity of VoIP flows if they are encoded with µ-law PCM and G.729 and presented multiple, different packet marking strategies.

In [3] we developed an off-line approach to quantify the importance of a VoIP packet. It uses the following definition to quantify packet importance: For a sample *s*, which is coded by an encoding and decoding implementation c, the quality of transmission is MOS(*s*,*c*). The sample *s* has a length of *t(s)* seconds. If a packet loss l_x occurs, the resulting quality is described by MOS(s,c,l_x). The importance of a loss is then calculated with the following equation:

$$\mathrm{Imp}(s,c,l_x) = \Big((4.5 - \mathrm{MOS}(s,c,l_x))^2 - (4.5 - \mathrm{MOS}(s,c,l_x))^2 \Big) \cdot t(s) \qquad (1)$$

3. Real-time classification

To present an algorithm that classifies speech frames in real-time; we simulate the impact of packet loss at the sender side in an analysis-by-synthesis approach [4]. Our algorithm is based on the off-line measurement algorithm for frame importances. The off-line approach is problematic, because it requires an entire sample, covering the preceding and following periods of speech to measure the importance of a packet. Thus, its algorithmic delay is very high and it is not suitable for real-time applications. Also, PESQ has a high computational complexity. To overcome these limits, the following three ideas are applied (Figure 1):

1. The off-line classification uses two PESQ MOS predictions to compare the original with the degraded and the original with the degraded plus concealed. In our real-time algorithm, we only compare the degraded (s_c) with the degraded plus concealed samples, as given in Equation 2.

$$\mathrm{Imp}(s,c,l_x) = (4.5 - \mathrm{MOS}(s_c,c,l_x))^2 \cdot t(s) \qquad (2)$$

2. We reduce the context information, including the preceding and following segments of speech.
3. Because we know the behavior of the codec and the packet loss concealment, we can simplify the PESQ algorithm and remove unnecessary parts from its processing steps.

The packet classification works as follows: For each packet, the algorithm encodes and decodes the speech segment, contained in each packet. It also considers the speech segments before the packet and optionally after the packet. The resulting degraded speech segment is called s_c. In addition it simulates the impact of a potential packet

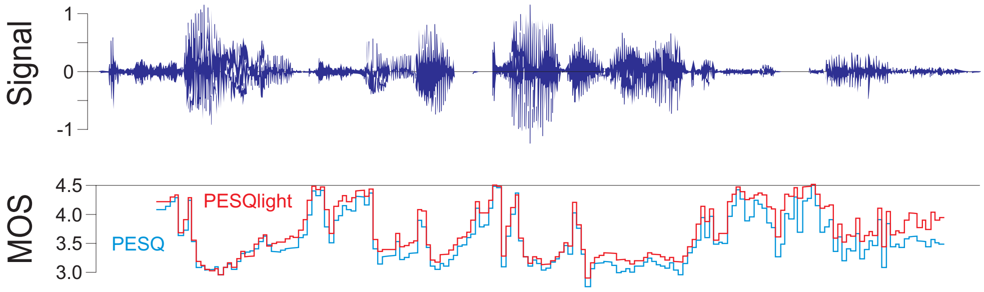

Figure 2: Visualizing PESQ and PESQlight.

loss and does not decode the packet but conceals it. The preceding and following segment are again just encoded and decoded. The two resulting speech samples (with and without loss) are then compared with PESQlight. The importance quantifier for the observed packet was calculated with Equation 2. For the next packet, these steps are repeated with a shifted speech segment to consider the next, new speech segment.

4. PESQlight

PESQlight is highly optimized for the distortions caused by G.711, its PLC, and the task of packet classification. As compared to PESQ we have done the following changes:

All samples have the constant length and are not affected by delay jitter. Thus, the dejittering parts of PESQ, which aligns references and degraded samples in time, have been removed. However, due to the PLC delaying the output by 3.75 ms, we have to remove this constant time offset.

The unmodified PESQ algorithm adapts the signal power to match the levels of original and degraded versions. In addition, both signals are raised or lowered to a normalized power level. To achieve this, the mean loudness of both signals is calculated before hand. In the packet classification task, we know that both versions have the same loudness. Also, if one assumes the presence of an adaptive gain control, which optimizes the loudness of the input signal, PESQlight does not need to determine the loudness of the signal. Instead, the signal has to be multiplied with a constant factor to achieve the normalized power level.

The unmodified PESQ determines the overall frequency distortion present in the degraded version. Then, an asymmetric signal derived from the original is calculated, which is lacking the missing frequency components of a given coding scheme or line filters. When the codec is known, its frequency filtering effect can be determined in advance. For example, the µ-law coding does not introduce any asymmetric frequency distortion. Thus, there is no need to calculate it or to consider frequency response compensation.

Furthermore, PESQ voice activity detection has been removed, usually being part of an automatic gain control already. Also, if a sample contains multiple utterances, PESQ analyses each utterance separately. For the packet classification, which deal only with very short samples (no more than 1 s), we assume the occurrence of only one utterance.

5. Validation

Runtime tests of PESQlight and PESQ on a Pentium M Centrino 1.5GHz notebook have shown that PESQlight outperforms PESQ by a factor of about 3.

The execution times depend strongly on the overall sample size, including the speech segment before and after the analyzed frame. For a segment of ¼, ½, and 1 s PESQ’s execution times where 26, 31, and 63 ms. In comparison, PESQlight performed the packet classification in 9, 11, and 22 ms respectively.

Figure 2 displays both MOS ratings after packet loss of PESQ and PESQlight over time. It also displays the original sound signal. One can see that PESQlight and PESQ predict speech quality similarly.

As a performance metric of the real-time packet classification algorithms, we calculated the correlation between the values as measured with the reference, off-line algorithm and the values generated with the algorithm under study. The test cases cover 832 samples, 4 languages, 16 speakers, and different loss positions within each sample. Altogether, about a million values were compared to achieve a high statistical accuracy. With varying overall sample length and varying position of the lost frame, the correlation between PESQ and PESQlight measurements are displayed in Table 1.

Position of lost frame	Correlation between reference, off-line and real-time algorithm (R)		
Sample length	¼ s	½ s	1 s
Last	0.32	0.28	0.35
Next to last (+10 ms delay)	0.60	0.63	0.58

Table 1: Performance of the real-time classification scheme for µ-law frames considering only active, non silent speech frames.

The longer the speech sample, the better the correlation. Also, it is beneficial, if not the last frame is dropped but the next to last. However, the algorithmic delay then increases by 10ms.

6. Conclusions

Given the knowledge of packet importance, we showed that significant performance gains can be achieved if only packets are transmitted that are important. However, the importance of speech frames has to be known precisely and at run-time, otherwise these performance gains are lost [8].

The importance of a packet can be measured both off-line and in real-time. In this publication we studied how the importance can be measured in real-time. We presented a novel algorithm that predicts frame importance more precisely than the previously published algorithms as it shows a higher correlation with the reference off-line values [8]. A modified PESQ algorithm called PESQlight was developed [9] that is fast enough to calculate the importance values in real time.

Further enhancements can be expected if speech frame importance is calculated using the inherent features of the encoding process. Alternative PESQlight can be further optimized (e.g. as in [10]). These optimizations have not been addressed in this publication and we suggest further research with the aim of reducing the computational complexity and increasing the prediction accuracy.

Bibliography

[1] C. Hoene, B. Rathke, and A. Wolisz, “On the importance of a VoIP packet”, In *ISCA Tutorial and Research Workshop on the Auditory Quality of Systems*, Mont-Cenis, Germany, April 2003.

[2] C. Hoene and E. Dulamsuren-Lalla, “Predicting performance of PESQ in case of single frame losses”, In *Measurement of Speech and Audio Quality in Networks Workshop (MESAQIN)*, Prague, CZ, June 2004.

[3] C. Hoene, S. Wiethölter, and A. Wolisz, “Calculation of speech quality by aggregating the impacts of individual frame losses´”, In *Thirteenth International Workshop on Quality of Service (IWQoS 2005)*, Passau, Germany, June 2005.

[4] F. D'Agostino, E. Masala, L. Farinetti, and J.C. De Martin, “A simulative study of analysis-by-synthesis perceptual video classification and transmission over diffserv IP networks”, In *IEEE International Conference on Communications (ICC '03)*, volume 1, pages 572-576, 2003.

[5] D.W. Petr, Jr. DaSilva, L.A., and V.S. Frost, “Priority discarding of speech in integrated packet networks”, *IEEE Journal on Selected Areas in Communications*, 7(5):644-656, 1989.

[6] J. C. De Martin, “Source-driven packet marking for speech transmission over differentiated-services networks”, In *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing*, pages 753-756, Salt Lake City, UT, May 2001.

[7] H. Sanneck, N. Tuong, L. Le, A. Wolisz, and G. Carle, “Intra-flow loss recovery and control for VoIP”, In *Ninth ACM international conference on Multimedia (MULTIMEDIA '01)*, pages 441-454, New York, NY, 2001. ACM Press.

[8] C. Hoene, “*Internet Telephony over Wireless Links*”, PhD thesis, Technical University of Berlin, 2005.

[9] T. Wimmer, “Developing a quality model to predict the importance of a VoIP packet”, Studienarbeit, Technische Universität Berlin, Berlin, Germany, March 2005.

[10] J. Holub, R. Smid, and J. Ocenasek, „Processing power optimisation for PESQ,“ in Measurement of Speech and Audio Quality in Networks (MESAQIN), Prague, CZ, 2004, pp. 57-60.