This full text paper was peer reviewed by subject matter experts for publication in the MULTICOMM 2006 proceedings.

Error Propagation After Concealing a Lost Speech Frame

Christian Hoene
University of Tübingen
Germany
hoene@ieee.org

Ian Marsh
KTH Stockholm
Sweden
ianm@sics.se

Günter Schäfer
University of Illmenau
Germany
guenter.schaefer@tu-illmenau.de

Adam Wolisz
Technical University of Berlin
Germany
awo@ieee.org

***Abstract*— Depending on the content of speech frames, the quality impairment after their loss differs widely. In previous publications we described an off-line measurement procedure to determine the loss impairment – *the importance* – of single speech frames. We showed that knowing the importance of frames can enhance the transmission performance of VoIP telephones significantly if only important frames are transmitted.**

Here we study to what extend the importance can be calculated at real-time: The loss impairment is due to the imperfect packet loss concealment (PLC) and also due to error propagation (EP). EP originates from the desynchronisation of the decoder's internal state and cannot be calculated at real-time. We developed a measurement method to determine the effect of the imperfect PLC and the temporal progression of the error propagation. The results show the trade-off between algorithmic delay and the accuracy of real-time importance calculation: A good frame classification needs to look ahead 20-40 ms in order to calculate the importance precisely.

I. Introduction

Packet losses significantly decrease the quality of voice communications. Usually, packet loss rate and speech quality are considered to be closely related. However, this ignores the fact that speech frames differ significantly. For example, it is well known that speech transmission can be interrupted during silence because silent speech frame have a minor impact on the quality of speech transmission. Active speech frames differ, too: Human speech generates two types of sounds: voiced and unvoiced. Voiced sounds have a regular pattern and usually high energy (e.g. "a","o", ...). Unvoiced sounds have a random nature (e.g. "h","sh", ...). Actually, one third of all active frames can be dropped while maintaining speech intelligibility [1]. However, only the right, more precise, the irrelevant frames are allowed to be dropped. Identifying irrelevant or important frames is a non-trivial task. Parts of this problem are addressed in this paper.

If a speech frame is lost, the receiver tries to extrapolate the last successful received frame to limit the impact of the lost frame. Such algorithms are known as packet loss concealment. 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 PLC cannot fully reconstruct the lost frame. Thus, the concealed frame differs from the sent frame and hence introduces a *loss distortion.*

Low-rate speech coders that transmit only signal differences suffer from an additional effect: If a frame is lost, the decoder becomes desynchronized [2]. If the internal state of the decoder does not match the encoder's state, the decoding of the following frames is affected and an additional distortion is introduced. We refer to this effect as *error propagation* (Figure 1). This effect is well known from digital, compressed TV and video transmissions. A transmission error causes the video signal to be distorted for a long period that can even last multiple video frames.

In [3] we presented a off-line method of how to determine the impact of an individual frame's loss – called the *importance of a frame*. It considers both loss distortion and error propagation. Here we extend this method to quantify the impact of the loss distortion and temporal progression of error propagation by studying common narrow-band speech codecs. Our results show that the frame following the loss contains the larger amount of the error propagation.

This results are important for the development of a real-time algorithm to classify speech frames: We can measure the amount of loss distortion at real-time [4]. But we cannot foresee the amount of error propagation because it depends on the following speech, which has not be spoken yet. Thus, a perfect frame classification must know the future. Any algorithm which does not know the future or cannot predict the amount of error propagation is less precise. So to say, this work shows the maximal achievable accuracy of real-time classification algorithms.

This paper is structured as follows: We start with a background and related work section. Then, we describe error propagation in narrow-band speech codings. Next, we present our measurements on quantifying the amount of error propagation. Finally, we conclude.

II. Related work and Background

A. Speech quality

The perceived quality of a telephony call can be measured with subjective tests. Humans evaluate the quality of service according to a standardized quality assessment process [5]. Often the quality is described by a *mean opinion score (MOS)* value, which ranges from 1 (bad) to 5 (excellent). More precisely, values which origin from passive human test results are called MOS-Listening Quality Subjective (MOS-LQS). In listening-only tests usually speech samples are used, which have a lengths ranging from 6 to 12 s. Listening-only tests

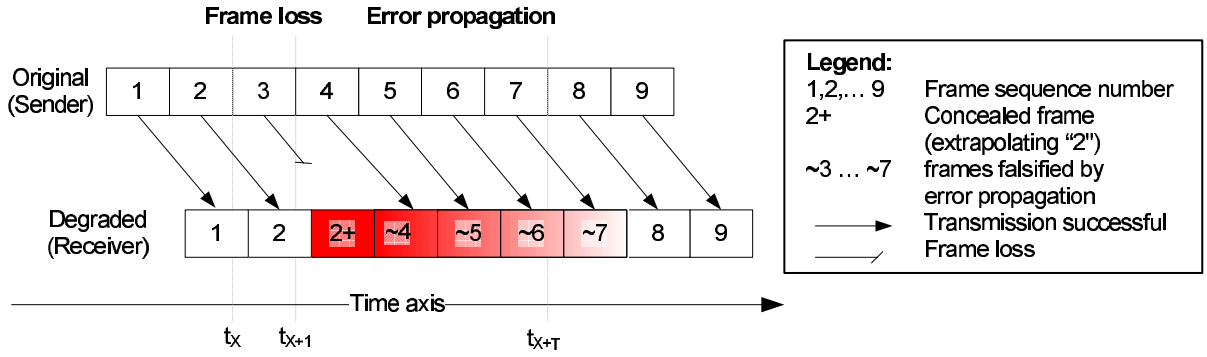

Fig. 1. Consequences of losing a frame.

are time consuming because many subjects have to be asked. Thus, in the last few years considerable effort has been made to develop instrumental measurement tools, which predict human rating behaviour.

The *Perceptual Evaluation of Speech Quality (PESQ)* algorithm predicts human rating behaviour for narrow band speech transmission [6]. It compares an original speech fragment with its transmitted and thus degraded version to determine an estimated MOS-LQO (Listening Quality Objective) value. Benchmark tests of PESQ have yielded an average correlation of R=0.935 with human based, subjected tests.

The correlation coefficient is often used to compare speech quality scores of human and instrumental predictions (e.g. PESQ). A value of R=1 would be a perfect match between both score sets, whereas R=0 means no correlation at all. A positive behaviour of correlation means that it is not influenced by linear scaling or adding an offset: Any linear regression applied to sets of measurement data does not change the value of R at all.

B. Real-time classification of speech frames

Petr et al. [7] suggested a method to mark speech frames containing background noise with the lowest priority. The next higher priority is assigned to voiced speech segments, which are not at the beginning of the voiced sounds. The next higher priority is assigned to non-initial fricative (e.g. the "ch" in the German word Bach). All other frames including the initial voiced and fricative speech segments are marked with the highest priority.

De Martin [8] has proposed an approach called Source-Driven Packet Marking, which controls the priority marking of speech packets in a DiffServ network. If packets are assumed to be perceptually critical, they are transmitted in a premium traffic class. All other packets are sent using the best-effort traffic class. The author describes a packet-marking algorithm for the ITU G.729 codec. For each frame, it computes the expected perceptual distortion, as if the speech frame were lost, under the assumption that no previous speech frames were lost. First, only speech frames with at least a minimal level of energy are considered to be marked as premium. Next, the marking algorithm takes the coding parameters (e.g. the gain, linear prediction filter, codebook indexes) and computes the parameters that would be computed by the concealment algorithm if the packet was lost. It then compares both parameter sets – the original and the concealed – in order to compute the perceptual quality degradation in case of loss.

Petracca and De Martin [9] presented a classification of AMR frames. Their analysis-by-synthesis distortion evaluation algorithm calculates the spectral distortion in dB for the LP coefficients, the percentage difference for the long-term prediction coefficients and the difference in dB for the codebook gains. If any of these values is above a given threshold, an AMR frame is marked as premium. De Martin's frame classifications do not consider any error propagation effects.

Sanneck et al. [10] analyzed the temporal sensitivity of VoIP flows if they are encoded with μ-law PCM and G.729: Single losses in PCM flows have a small sensitivity to the current speech properties. Multiple consecutive losses have a higher impact on the quality degradation than single, isolated losses. The concealment performance of G.729, on the other hand, largely depends on the change of speech properties. If a frame is lost shortly after an unvoiced/voiced transition, the internal state of the decoder might be de-synchronized for up to the next 20 following frames.

Rosenberg et al. [2] measured the length of desynchronisation after losing a G.729 frame. Our work extends these initial results, includes other codecs and enhances the accuracy of the measurement procedure.

C. The Importance of Individual Speech Frames

In [3] Hoene et al. describe an off-line measurement procedure, which measures the impact of loss on speech quality and quantifies the importance of frames. They used this method in an extensive experiment effort evaluating more than two million different, deliberately simulated packet and frame losses. Hereby they considered the most common standardized, narrow-band speech codecs and concealment algorithms, which are Adaptive Multi-Rate (AMR), G.711 plus Annex I[1], and G.729. Also, they validated their method with formal listening-only tests [11].

[1]Its "frame" length is set to 10 ms.

In [1] Hoene et al. developed an quality metric to describe the importance of speech frame or VoIP packets. Under many conditions this metric shows an additive property of equality. Thus, is it possible to to give a statement like “frame A and frame B are as important as frame C” or “frame A is three times more important than frame B”. The metric’s definition is given as: *The importance of frame losses is the difference between the quality due to coding loss and the quality due to coding loss plus frame losses, multiplied by the length of the sample.* The following equation describes how to calculate the importance. For a given sample s that has a length of $t(s)$, a given codec implementation c, and a loss event described with e $\mathrm{MOS}(s,c)$ describes the speech quality due to coding loss, and $\mathrm{MOS}(s,c,e)$ describes the speech quality due to coding as well as frame loss.

$$\begin{gathered} Imp(s,c,e) = (cl - c) \cdot t(s) \\ \text{with } cl = (4.5 - MOS(s,c,e))^2 \\ \text{and } c = (4.5 - MOS(s,c))^2 \end{gathered} \tag{1}$$

In this paper, we extend the off-line measurement procedure and use (1) to quantify the importance.

III. Real-time packet classification

To control the transmission of speech frames, their importance should be known at transmission time. For example, in addition to the encoding of speech, the sender could calculate the importance of each speech frame. This leads to the question, is it possible to predict the importance of speech frames at transmission time? In general, the consequences of packet loss can be split into two effects (Figure 1):

First, the lost frame is concealed at the receiver, which causes a distortion if the concealment does not perfectly predict the frames content. In the illustration this refers to frame 3 (transmitted) and frame 2+ (concealed). The encoder knows the original and degraded speech segment. It can also predict the behaviour of the decoder in case of loss, as the decoder’s concealment algorithm is known (since it is standardized). In principle, the encoder can therefore calculate the impact of imperfect concealment.

The second effect of packet loss is due to error propagation. After a frame loss the internal state of the concealment algorithm is desynchronized. The impact of error propagation cannot be known at the time of transmission because the length of error propagation depends on the following speech content. In case of interactive telephony the following speech has not yet been spoken. Thus, predicting the importance of a speech frame at run-time will always be falsified by the effect of error propagation.

In Figure 2, we display how long it takes until synchronisation of the decoder is achieved. We measure desynchronisation lengths for the ITU G.729 coding, which last up to 650 ms.

To demonstrate the impact of imperfect packet loss concealment and error propagation we plotted the speech signals of a sample segment in Figure 3 for different encoding schemes. Beside the original sample, the figures also contain the encoded/decoded (=degraded) signal, the encoded/lost/decoded/concealed signal, and the difference between those signals. Also, the figures contain the PESQ MOS values to quantify the perceptual impact of coding and concealment degradation.

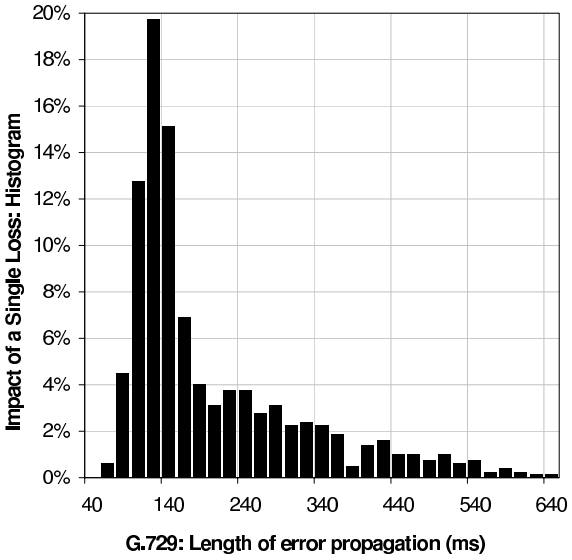

Fig. 2. Histogram of error propagation lengths in case of loss of one G.729 frame. We measured the time until the internal state of the G.729 decoder matches the non-loss state again. The decoders’ post-filter is ignored as it does not synchronise again.

IV. Quantifying error propagation

A. Method

The aim of this paper is to quantify the imperfect concealment and error propagation caused by a single frame loss. The question arises how should the effects be measured? The speech sample could be split into two parts. The first part contains the content until the end of the concealed frame (e.g. frame 1, 2, 2+). The next part contains the remaining content (e.g. frame ~4 to 8). The position of the split is exactly after concealing and decoding the lost frames. Thus, the effect of concealment and the effect of error propagation are separated into two samples. For both samples the degradation can be measured with PESQ and compared to the corresponding samples that do not contain any frame loss. This method is problematic due to two reasons.

First, PESQ judges the speech quality largely different if the sample content differs. Thus, splitting the sample and thus changing the sample’s length introduces a source of error. Instead, the sample content must not be changed.

Second, a hard split between two samples introduces an additional clicking sound, which falsifies the results.

Therefore, we developed the following measurement procedure (see Figure 4). We generate two samples containing first the degraded sample without loss and second, the degraded sample with one frame loss. Then, we mix both samples to

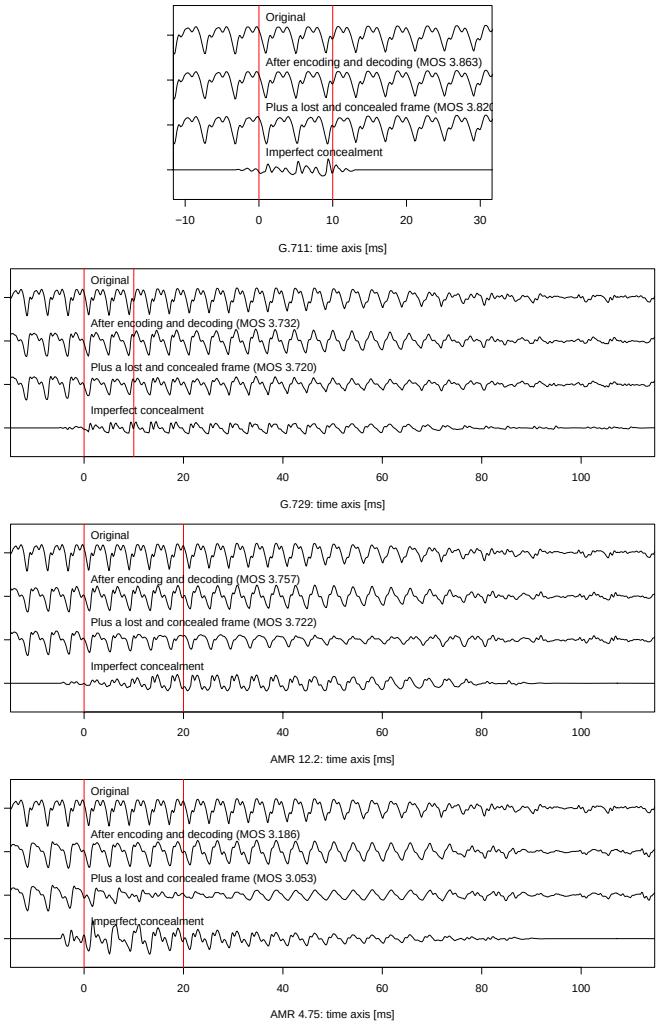

Fig. 3. Speech signals before and after decoding, after loss concealment, and the difference between the decoded and concealed signals. The two vertical lines define the length of the frame loss.

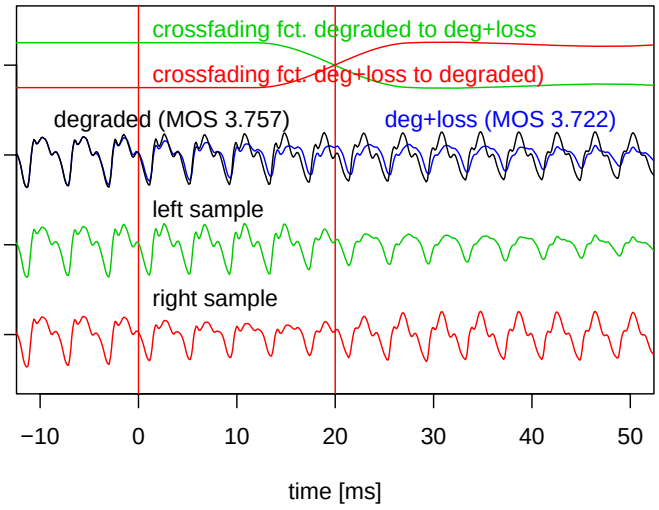

Fig. 4. Splitting the imperfect concealment and error propagation into two different speech samples. The position of the lost frame is marked with two vertical, red lines. Two degraded samples are generated, with loss (blue) and with-out loss (black). Then, to get the impact of PLC we crossfade from the loss to the no-loss sample to produce a new speech sample called “left”. Similar, to get the impact of EP we crossfade from the no-loss to loss sample to produce a new speech sample called “right”.

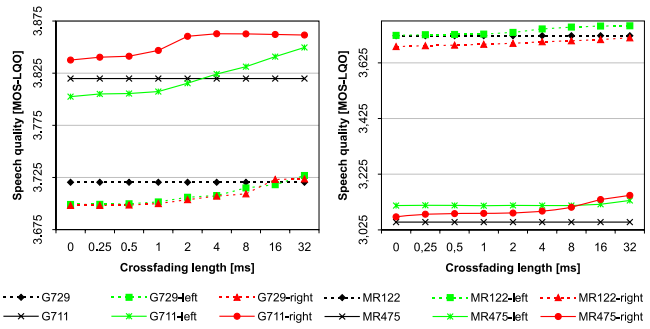

Fig. 5. Impact of crossfading length on speech quality.

produce new samples: We crossfade just after the lost frame (right vertical line in Figure 4). The crossfading function is a cosine curve. Then, two new samples are produced. The first called “left” contains the concealment frame and the second called “right” contains the error propagation. The speech quality of those samples is then measured with PESQ.

This algorithm leads to another question: How long should this crossfading period be? For one test condition containing one frame loss we conducted measurements with varying crossfading lengths (see Figure 5).

The black lines represent the speech quality considering imperfect concealment and error propagation. If the crossfading is done in less than 4 ms, it introduces an addition distortion that lowers the speech quality. However, if the crossfading is too slow, the short effect of a single frame loss is smeared over the left and right samples. Thus, we will use a crossfading length of 4 ms in the following work.

If in addition the split is conducted not only at the end of lost frame (refer to as position 0 ms) but also at positions shortly after the lost frame, we can observe the temporal progression of the error propagation.

B. *Experimental set-up*

In order to study the impact of frame losses on the speech quality, we conducted experiments as depicted in Figure 6 and described in [1], [3]. We used speech recordings, taken from an ITU coded speech database [12] that consists of 832 files, each 8 seconds long, with 16 different speakers, 8 female and 8 male, spoken in four different languages, without any background noise. We chose this database to limit the influence of specific languages [13], speakers, or samples. We chose three common narrow-band-speech-coding algorithms: ITU’s G.711 and G.729, and ETSI’s Adaptive-Multirate (AMR).

We simulated packet losses at different positions within the sample. We varied the coding scheme, the packet loss positions and the sample content, and generated for each test case a

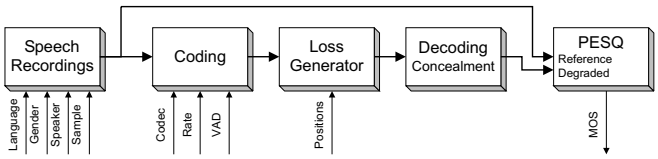

Fig. 6. Measuring one frame loss.

degraded audio. In addition, we split the sample as described above. To assess the speech quality we applied the ITU's PESQ algorithm [6] to calculate a MOS value. Some million PESQ rating results were gathered to achieve a high accuracy for statistical analysis.

C. Results

For each test condition we calculate the distribution of importance values. In Figure 7, we display the importance value of the left and right parts of the loss distortion. Actually, we choose to display the 75% percentile as it is close to the median importance of all active speech frames. In addition, the sum of the left and right importance values are displayed since the importance metric is to some extent additive.

The first graph containing G.729 values shows that this codec has a high amount of error propagation and it takes approximately 80 ms until this effect disappears. The next graph using G.711 is to demonstrate the quality of our measurement procedure as in case of G.711 the error propagation is fixed to a maximal length of at most 3.75 ms [14]. It shows that our measurement procedure does not split perfectly both distortion effects, but has an inaccuracy of 0–10 ms. Last, the values for AMR coding are shown. The amount of error propagation is small and disappears after 20–40 ms.

D. Analysis

Coming back to the main question of this paper: How well can the importance of a speech frame can be predicted in real-time? As a performance metric we will calculate the Pearson's correlation coefficient of the offline, reference importance values and the left, right, and both (left+right) importance values.

In Figure 8 we display the correlation to compare the importance value sets of the reference (offline), left, right, and both measurements. If only the imperfect concealment is considered to calculate importance values, the performance for G.729 coding is R=0.72, for G.711: R=0.91, and for AMR: R=0.73. If in addition the next frame after the concealed frame is considered, the performance increases to G.729: R=0.92, G.711: R=1.00, and AMR: R=0.97. Given the precision of our measurement procedure (R=0.94, [11]), the later results are almost perfect.

V. Conclusions

Given the knowledge of packet importance, we showed that significant performance gains can be achieved if only packets are transmitted with priority that are important. However, the importance of speech frames has to be known precisely, otherwise this performance gains are lost [15]. The importance of a packet can be measured both off-line and in real-time. A measurement procedure that identifies the impact of a single frame loss offline has already been developed and has been verified with formal listening-only tests in previous publications.

In this paper, we studied how the importance can be measured in real-time. This is difficult, as the importance values partially depend on the amount of error propagation which is not known at the time of transmission. Waiting for the next frame before calculating the importance value significantly increases the accuracy of the importance predictions. The enhancement comes at the cost of an increased algorithmic delay. A good compromise is a look ahead of 20 to 40 ms to minimize error propagation effects.

References

[1] C. Hoene, S. Wiethoelter, 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.

[2] J. Rosenberg, "G.729 error recovery for internet telephony," Columbia University Computer Science, Prof. H. Schulzrinne, New York, NY, Tech. Rep. CUCS-016-01, Dec. 2001.

[3] 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, Apr. 2003.

[4] F. D'Agostino, E. Masala, L. Farinetti, and J. 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)*, vol. 1, 2003, pp. 572–576.

[5] *Methods for subjective determination of transmission quality*, Recommendation P.800, ITU-T Std., Aug. 1996.

[6] *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-To-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, Recommendation P.862, ITU-T Std., Feb. 2001.

[7] D. Petr, J. DaSilva, L.A., and V. Frost, "Priority discarding of speech in integrated packet networks," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 5, pp. 644–656, 1989.

[8] 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*, Salt Lake City, UT, May 2001, pp. 753–756.

[9] M. Petracca, A. Servetti, and J. C. De Martin, "Voice transmission over 802.11 wireless networks using analysis-by-synthesis packet classification," in *First International Symposium on Control, Communications and Signal Processing*, Hammamet, Tunesia, Mar. 2004, pp. 587–590.

[10] 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)*. New York, NY: ACM Press, 2001, pp. 441–454. [Online]. Available: citeseer.nj.nec.com/article/sanneck01intraflow.html

[11] 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.

[12] *Coded-speech Database*, Recommendation P.Supplement 23, ITU-T Std., Feb. 1998.

[13] D. Goodman and R. Nash, "Subjective quality of the same speech transmission conditions in seven different countries," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '82)*, vol. 7, 1982, pp. 984–987.

[14] *A High Quality Low-Complexity Algorithm for Packet Loss Concealment with G.711*, Recommendation G.711 Appendix I, ITU-T Std., Sept. 1999.

[15] C. Hoene, "Internet telephony over wireless links," Ph.D. dissertation, Technical University of Berlin, TKN, 2005.

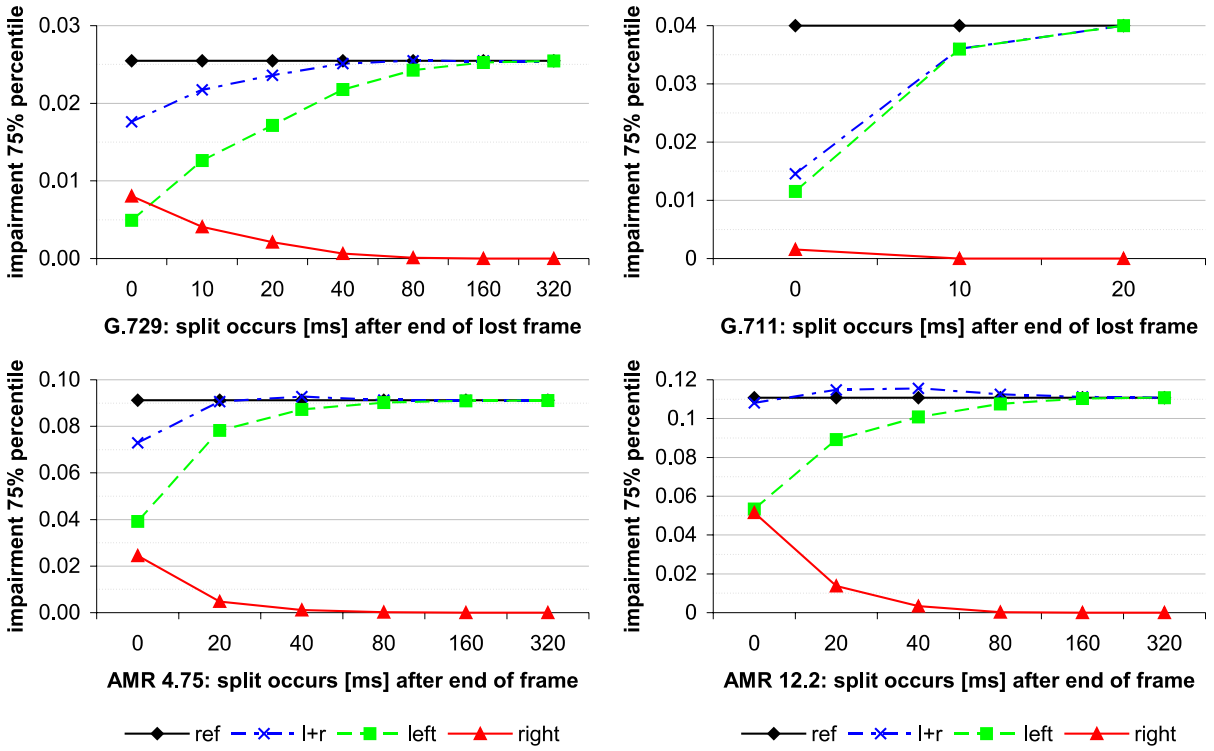

Fig. 7. Temporal impact of error propagation on the importance of frame losses.

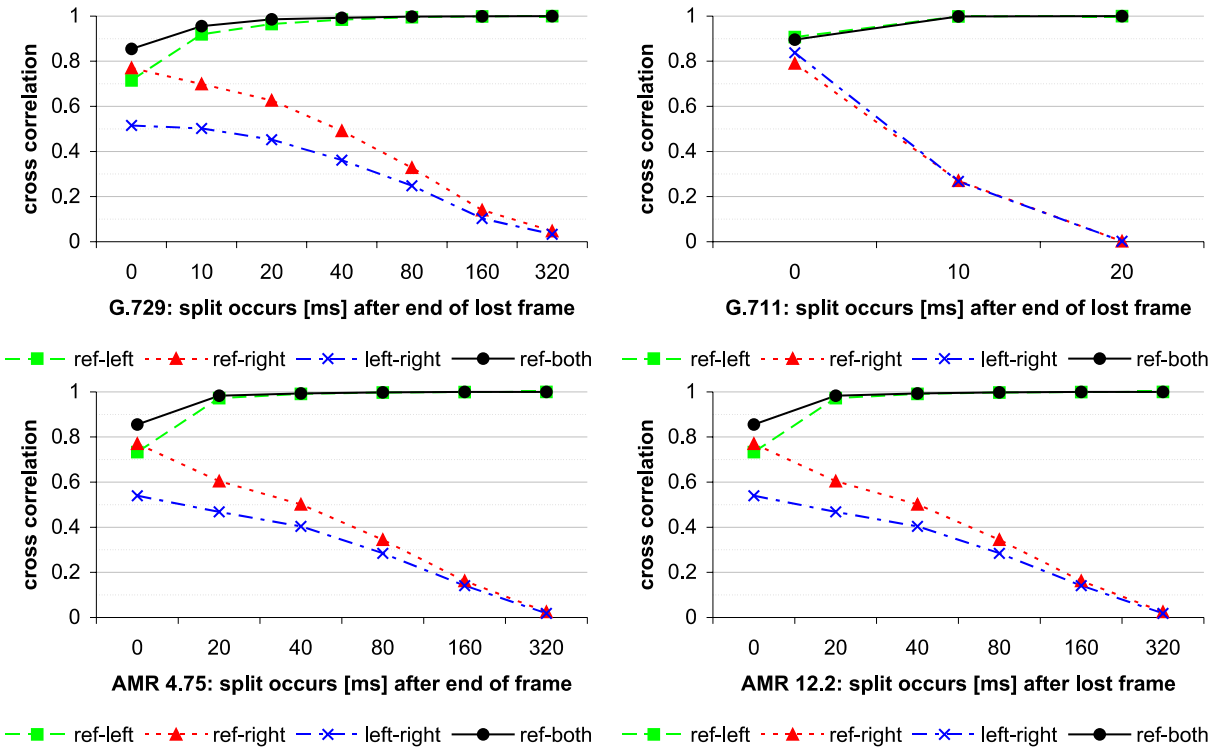

Fig. 8. How well can the frame importance be predicted if some X milliseconds of the following speech is considered in addition to concealment effect? (This result is displayed in the “ref-left” line.)