In Proc. of Thirteenth International Workshop on Quality of Service (IWQoS 2005), Passau, Germany, June 2005

# Calculation of Speech Quality by Aggregating the Impacts of Individual Frame Losses

Christian Hoene, Sven Wiethölter, Adam Wolisz

Technical University of Berlin, Germany
hoene|wiethoel|wolisz@tkn.tu-berlin.de

**Abstract.** Losing VoIP packets or speech frames decreases the perceptual speech quality. The statistical relation between randomly lost speech frames and speech quality is well known. In cases of bursty and rate-distortion optimized losses, a precise quality model is required to relate losses to quality. In the present paper, we present a model that is based on the loss impact - or the *importance* - of single speech frames. We present a novel metric to calculate the impact of the loss of multiple frames by adding the importance of the respective single frames. This metric shows a high prediction accuracy for distant losses. For losses following each other closely, we present an aggregation function which models the psychoacoustic post-masking effect. Our model helps to develop networking algorithms that control the packet dropping process in networks. For example, we show that a proper packet dropping strategy can significantly increase the drop rate while maintaining the same level of speech quality.

## 1 Introduction

In packet-based communication networks, such as the Internet, packet losses are a major source of quality degradation. This is true especially for real-time multimedia services over wireless links such as Wifi-VoIP. One would expect that the impact of VoIP packet loss[1] on speech quality is well understood. However, this is not the case because it is a highly interdisciplinary problem. Multiple "layers" have to be considered covering the loss process of IP-based networks, the behavior of speech codecs and frame loss concealment, the psychoacoustics of the human hearing, and even the cognitive aspects of speech recognition.

State of the art algorithms look up speech quality scores in tables, depending on the measured loss rate and the speech coding. Alternatively, these tables are modeled as linear equations or with neural networks [1]. However, the relation between mean packet loss rate and speech quality is only a statistical description because the deviation for specific loss pattern can be high and depends on the content of the lost frames. Also, these relations are only valid for a specific loss

[1] Speech frames are compressed segments of speech which are generated by an encoder. VoIP packets carry one or multiple speech frames. Usually, a speech frames or VoIP packet carry a segment of speech, which has a length of 10, 20, or 30 ms.

pattern. For example, bursty losses (multiple VoIP packet or speech frame losses in a row) can have a different impact [2,3] depending on the codec and the duration of the burst. Lately, rate-distortion optimized multimedia streaming or selective prioritization algorithms have been introduced [4,5,6], which control the loss process and select which media frames to drop. If losses cannot be avoided, they try to drop negligible instead of important media frames. Thus, they increase the service quality for a given loss rate. The loss rate can be rather high if only unimportant losses occur (refer to [7]). On the other side, losses of important frames degrade the speech quality highly.

In the same research context, a method has been developed and validated, which measures the *importance* of a single speech frame [8]. The importance of a speech frame is defined as the impact on the speech quality caused by a frame loss. In this paper we assume that we can use this method to determine the impact of one frame loss. Then, the question arises how the impact of multiple losses can be determined using the importance of single frame losses. The development of a novel metric or dimension of frame importance, which simply can be summed up to get an overall impact of multiple frame losses, is presented here.

The ITU-T P.862 PESQ algorithm [9,10,11] can assess the impact of one or multiple frame losses but works only for audio files and not on a speech frame level. It is an instrumental speech quality assessment tool, which simulates the human rating behavior of speech quality. It compares two speech samples – the original sample and the degraded versions, which might include coding and frame loss distortions – to calculate the mean option score (MOS) value. PESQ by itself cannot be directly applied on VoIP packets [12] and has a high computational delay and complexity, which inhibits its on-line and real-time application.

Thus, we remodel the internal behavior of PESQ algorithms using it for frame losses: We apply the algorithm which PESQ uses to aggregate signal distortions over time in order to accumulate frame loss distortions over time. This aggregation algorithm is also the basis for the novel importance metric, which allows adding the frames' importance linearly and thus has a low complexity. It shows a high prediction performance, if the losses are distant. If frame losses occur shortly one after the other, temporal auditory masking have to be considered. We develop a heuristic equation to model these effects. Overall, our approach shows a high correlation with instrumental speech quality measurements for many loss patterns.

The following paper first describes the required technical background. We also give an example showing the impact of the packet dropping strategy on the speech quality. Then, we present the approach on how to assess multiple speech frame losses. In the fifth chapter we compare our approach with the PESQ's speech quality predictions. Finally, we summarize this work and give an outlook to further research.

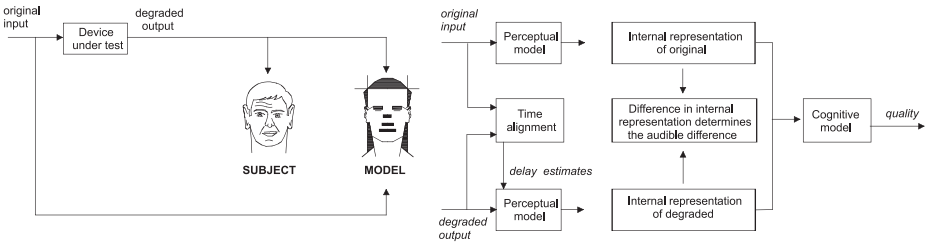



**Fig. 1.** Overview of the basic architecture of PESQ [9].

## 2 Background

### 2.1 PESQ

The Perceptual Assessment of Speech Quality algorithm predicts human rating behavior for narrow band speech transmission. It compares an original speech fragment with its transmitted and thus degraded version to determine an estimated mean option score (MOS), which scales from 1 (bad) to 5 (excellent). For multiple known sources of impairment (typical for analogue, digital and packetized voice transmission systems) it shows a high correlation (about $R = 0.93$) with human ratings. In the following we will describe some details of PESQ because they are required to understand the following sections.

*Overview:* PESQ transforms the original and the degraded signal input to internal representations of a perceptual model. If the degraded signal is not time aligned, e.g., due to jitter or delay, it is first adjusted to the original signal [10]. Next, the perceptual difference between the original signal and the degraded version is calculated [11] considering the human cognition of speech. Finally, PESQ determines perceived speech quality of the degraded signal (see Fig. 1) .

*Computation of the PESQ MOS score:* The final MOS score is simply a linear combination of so called normal and asymmetrical disturbance. In most cases, the output range will be a MOS-like score between 1.0 and 4.5, the normal range of MOS values found in human subjective experiments:

$$PESQ_{MOS} = 4.5 - 0.1 \cdot D_{indicator} - 0.0309 \cdot A_{indicator} \tag{1}$$

with $D_{indicator}$ being the normal disturbance and $A_{indicator}$ being the asymmetrical disturbance. Before this final calculation the following processing steps are conducted:

*Time-frequency decomposition:* PESQ's perceptual model performs a short term FFT on the speech samples that have been divided into 32 ms *phoneme*. The phoneme overlap each other with 50% so that each position within the sample is covered by exactly two phonemes. PESQ calculates the spectral difference between orignal and degraded to calculate distortion of at given phoneme.

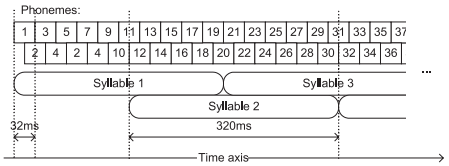



**Fig. 2.** PESQ: Structuring of the speech to phonemes (32 ms) and syllables (320 ms). The lengths presents have been chosen because they yield highest prediction performance.

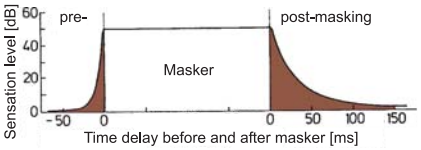



**Fig. 3.** Schematic drawing to illustrate and characterize the regions, in which pre- and post-masking occur (shaded areas) if a masker is present [14].

*Asymmetric effect:* If high correlation between PESQ and subjective listening-only ratings [13] should be achieved, an asymetric effect has to be considered: Humans do not know the quality and spectrum of the original speech because they just hear the degraded speech. Actually, they compare the degraded speech with an immagetive original, which differs to the original by lacking curtain spectrum components. It is caused by the fact that the listener adapts to constant limitations of the transmitted signal spectrum. PESQ models this behavior and calculates separately two perceptual differences for both the normal and the asymmetric signals. Both disturbances are aggregated separately over time. Finally, they are combined.

*Weighting of disturbances over time:* PESQ uses a two layer hierarchy to group phonemes to *syllables* and to aggregate syllables over the entire sample length (Fig. 2). Twenty phoneme disturbances are combined to one syllable distortion with (2). Phonemes are aggregated with an exponent of 6 to model a cognitive effect: Even if only one phoneme is distorted, it is not possible to recognise the syllable anymore [15]. The authors of PESQ argue that this is a cognitive effect which needs to be considered for high prediction performance.

$$syllable_{indicator}^{AorD}[i] = \sqrt[6]{\frac{1}{20}\sum_{m=1}^{20} phoneme_{disturbance}^{AorD}[m+10i]^6} \tag{2}$$

A syllable has the length of 320 ms. Similar to phonemes, syllables are also 50% overlapping and cover half of the previous and following syllables. The syllables are aggregated with (3) over the entire sample. Syllables are aggregated with an exponent of 2 because disturbances occurring during active speech periods are perceived stronger than those during silence [15].

$$AorD_{indicator} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} syllable_{indicator}^{AorD}[n]^2} \tag{3}$$

### 2.2 Temporal Masking

Zwicker and Fastl [14] describe temporal masking effects which characterize the human hearing: The time-domain phenomena pre- and post-masking plays an important role (Fig. 3). If faint sound follows shortly after a loud part of speech, the faint part is not heareable because it is masked. Also, if the maskee precedes the masker it vanishes.

If the temporal masking effect is applied to distortion values, the distinction between masker and maskee on the one side and between pre- and post masking the other side is difficult. If a frame got lost, it causes a distortion, resulting in a segment of speech which can be louder or fainter than the previous segment. If it is louder, the previous segment is pre-masked, if it is fainter, the loss is post-masked. Thus, if one considers only distortion, it is not possible to distinguish pre- and post-masking. Instead, the same amount of distortion can cause pre- or post-masking or can be effected itself by pre- or post-masking, depending on the loudness of the resulting speech segment.

This perceptual effect denoted as temporal masking had been considered as an addition to the PESQ algorithm. However, after implementing it, it never showed any improvements to the prediction performance of PESQ. Thus, it was not included.

### 2.3 Single Frame Loss and Importance

In [8], a measurement procedure was presented, which determines the impact of single frame losses. It is based on PESQ and consists of two speech quality measurements: First, a speech sample is encoded and decoded again. PESQ compares this degraded sample with the original to estimate the impact of the encoding process. Next, the same speech sample is encoded and one (or multiple frames) are dropped. Then, the encoded frames are decoded or concealed, depending whether the frames are lost. Again, PESQ calculates the MOS value. The impact of the frame loss is now identified by comparing the pure coding-degraded MOS value with the MOS value containing additionally the frame loss. The authors have conducted two million measurements with speech samples containing deliberately dropped packets. For example, Fig. 4 displays the distribution of MOS values, varying the sample content and the location of the lost packet in samples having a length of 8 s. The results show that the encoding has a large impact on the speech quality as well as the sample content (e.g., speaker and sentence). One can see that the *coding distortion* varies widely and depends on the sample content. Figure 4 shows also the impact of losing one and two speech frames. The *frame distortion* remains small.

Originally, PESQ has not been designed for measuring the impact of single packet losses and in such case works outside its specification [9]. Therefore this application has been verified with formal listening-only tests. Humans' subjective ratings and the predictions of PESQ have a cross correlation of R=0.94 [7]. PESQ therefore reflects well the single frame measurements.

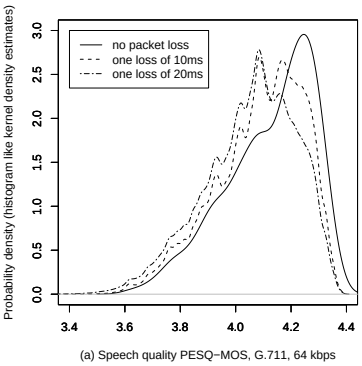



**Fig. 4.** Impact of coding distortion and packet loss on speech quality. The probability density functions (PDF) are displayed for varying sample contents.

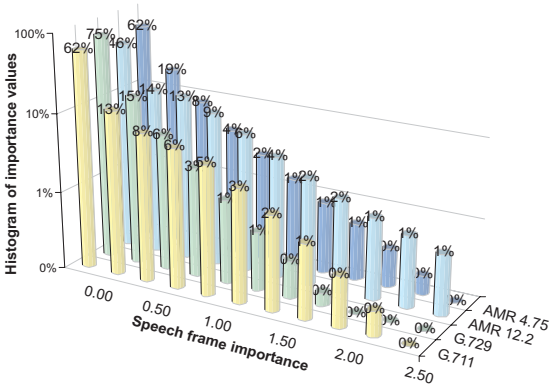



**Fig. 5.** Histogram of speech frame importance values during voice activity. Most speech frames are not important at all.

In [8], an importance metric has been firstly introduced. It its defined as follows: If a sample is encoded, transmitted and decoded, the maximum achievable quality of a transmission is limited by the coding performance, which depends on the codec algorithm, its implementation, and the sample content. Some samples are more suitable to be compressed than others (see Fig. 4). For a sample $s$, which is coded with the encoding and decoder implementation $c$, the quality of transmission is $MOS(s, c)$. The sample $s$ has a length of $t(s)$ seconds. As explained above, the quality is not only degraded by encoding but also by frame losses. If such losses occur, the resulting quality is described by $\mathrm{MOS}\,(s, c, e)$. The vector $e$ describes a loss event. The following empiric equation (4) describes how to calculate the importance. If this equation is applied on the data displayed in Fig. 4, one can see that most speech frames during voice activity are not important at all (Fig. 5).

$$\mathrm{Imp}\,(s, c, e) = (\mathrm{MOS}\,(s, c) - \mathrm{MOS}\,(s, c, e)) \cdot t\,(s) \tag{4}$$

Still, one drawback remains. Equation 4 can only measure the effect of a single frame loss. If it is used to add the impact of two or more lost frames, it does not scale linearly with the number of frames [8]. Thus, the aim of this paper is to develop an "additive" metric.

## 3 Example: Frame Dropping Strategies

We have described a method to classify the impact of single and multiple frame losses. But how cam this be applied? In the following we assume a scenario in which we know the importance of each frame and in which we can control the

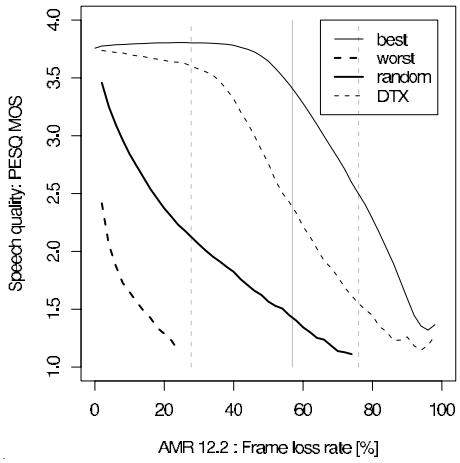



**Fig. 6.** Impact of dropping strategy on the speech quality. The gray, vertical lines refer the minimal, mean, and maximal percentage of silent frames in the sample set.

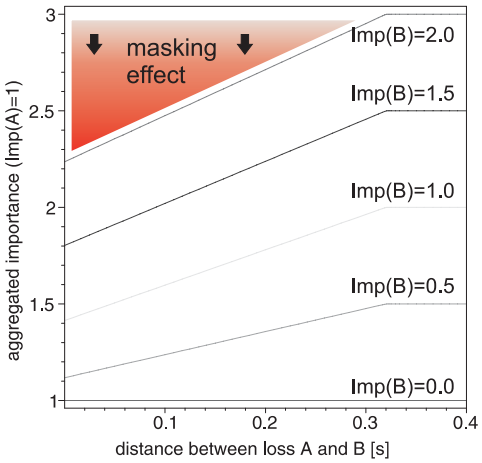



**Fig. 7.** Behavior of Equation 12 depending on loss distance and importance.

loss process - e.g. which frames can be dropped. We assume that we can drop frames at any position and show, how the *frame dropping strategy* influence the speech quality.

In our simulations, we increase the dropping rate from 0% to 100% in steps of 2% using adaptive-multi rate (AMR) coding at 12.2 kbit/s (other codecs give similar results). We use 832 different samples but consider only the mean MOS value over all sample variations. In Fig. 6 we have displayed the speech quality depending on the frame loss rate. If frames have to be dropped, which frames should be dropped? Classic approaches chose the frames randomly (line "random"). Discontinuous Transmission (DTX) algorithms detect the voice activity (VAD) to interrupt the frame generation. They reduce the transmission rate during inactive speech periods while maintaining an acceptable level of output quality. Thus, a DTX algorithm would first drop silent frames, and then active frames (line "DTX"). Using the metric of frame importance, we introduce two novel strategies: As a worst case we consider the dropping strategy that drops the most important frames first. The second called "best" loss strategy preferentially drops the less important frames, and only at high loss rates important frames are dropped. One can see that the "best" loss strategy performs better than the DTX and random case. In case of the worst strategy, the speech quality drops very fast.

## 4 Additive Metric

A metric that describes the importance of frames shall fulfil the following requirements: First, it should be easily deployable to quantify the impact of frame losses. Consequently, the loss distortion should be measurable with off-the-shelf instrumental measurement methods like PESQ (or any other successor). For example, it should be able to calculate the metric with two speech quality measurements: with loss and without loss. Second, the metric shall be one-dimensional. Of course, the distortions caused by frame loss can have many effects. However, it shall be modeled as a one-dimensional quality scale because this would simplify the development of algorithms that utilize this metric. Last, it should be possible 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". In a mathematical sense, this requirement is called additive property. It is of importance when frame loss impacts are to be applied in analytical contexts such as the rate-distortion multimedia streaming framework by Chou and Miao [6].

The development of such a metric is based on the idea to study the internal behavior of PESQ and to remodel it for frame losses. PESQ predicts the impact of frame loss rather well but is far too complex to be applied on a frame basis. Thus, a simpler model is required that only contains the issues that are relevant. The proposed approach is based on the following three principles. First, we assume that the importance of frames is known. For example, the method described in section 2.3 can be used for offline purposes. A real-time classification of frame importances is beyond the scope of this paper and is addressed in [16]. Second, if two or more frame losses have a distance of more than 320 ms, the importance values, as calculated by (10), can simply be added. Last, if two losses occur shortly after each other, then (12) is required to add the importance values. In the following it is described how we have developed this approach by remodelling PESQ's behavior.

*Asymmetric effect:* PESQ judges the impact of distortion with two factors, the asymmetric and the normal distortion. In case of frame loss, the overall coding spectrum, which influences strongly the asymmetric effect, is not changed because the impact of a frame loss is limited to the position of its loss and does not change the rest of the sample. Also, the asymmetric effect is mainly caused by the encoding and not by frame losses. Therefore it is reasonable to neglect the difference between asymmetric and normal distortion and consider just the sum of them.

*Long-term aggregation:* In general, the weighting of disturbances over time is determined as in PESQ. Thus, the syllable disturbances are added up as described in (3). Contrary, we consider disturbances of speech frames instead of syllables. The disturbance consists of coding as well as loss distortion as shown in (5). If frame losses are not present, the term $dist_{loss}[i]$ is zero.

$$sylablle_{indictor}[i] = dist_{coding}[i] + dist_{loss}[i]\,. \tag{5}$$

Combining (3) and (5) we can write

$$(AorD_{indicator})^2 = \frac{1}{N}\sum_{n=1}^{N}\left(dist_{coding}[n] + dist_{loss}[n]\right)^2 \tag{6}$$

and transform (6) to (7):

$$\begin{aligned} &N \cdot (AorD_{indicator})^2 - \textstyle\sum_{n=1}^{N} dist_{coding}[n]^2 \\ &= \textstyle\sum_{n=1}^{N}\left(dist_{loss}[n]^2 + 2 \cdot dist_{coding}[n] \cdot dist_{loss}[n]\right) \end{aligned} \tag{7}$$

As an approximation we combine both asymmetric and symmetric disturbances. Then, (1) can be simplified to:

$$MOS = 4.5 - AorD_{indicator} \tag{8}$$

with $AorD_{indicator} = 0.1 \cdot D_{indicator} - 0.0309 \cdot A_{indicator}$.Combining (7) and (8), we get:

$$\begin{aligned} &\left((4.5 - \mathrm{MOS}(s,c,e))^2 - (4.5 - \mathrm{MOS}(s,c))^2\right) N \\ &= \textstyle\sum_{n=1}^{N}\left(dist_{loss}[n]^2 + 2 \cdot dist_{coding}[n] \cdot dist_{loss}[n]\right) \end{aligned} \tag{9}$$

with $\mathrm{MOS}(s,c)$ being the speech quality due to coding loss and $\mathrm{MOS}(s,c,e)$ being the speech quality due to coding as well as frame loss. Equation 9 is the basis of our new importance metric. One can see that if a loss distortion does not overlap within one syllable, the distortions can simply be added. We define (10), which approximates a linear scale better than (4).

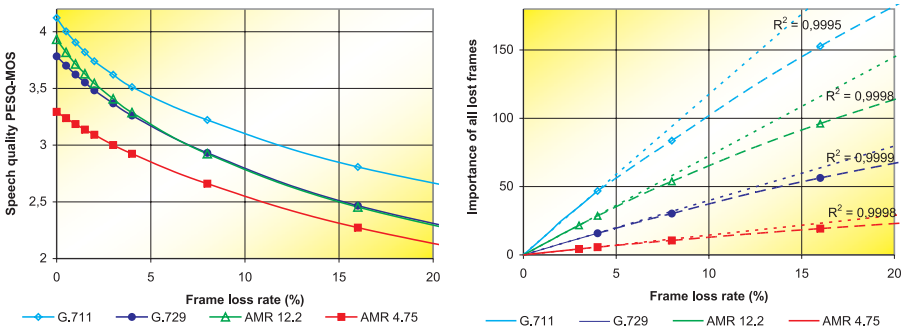
$$\begin{aligned} &Imp(s,c,e) = (cl - c) \cdot t(s) \\ &\text{with } cl = (4.5 - MOS(s,c,e))^2 \text{ and } c = (4.5 - MOS(s,c))^2 \end{aligned} \tag{10}$$

*Short-term aggregation:* For the short term aggregation, we first model the impact of two frame losses with two delta impulses at time $t_a$ and $t_b$ with the height of $imp_a$ and $imp_b$ representing the importance. If the distance $t_{width} = t_b - t_a$ is larger than 320 ms, adding of the importance values is done as described in the previous section. Otherwise, it is calculated as explained below.

First, we calculate the probability that both losses occur in the same syllable. We assume that syllables start at $0, 320, \ldots$ ms and have a length of $t_{syll} = 320$ like in PESQ. Because of the re-occurrence pattern of syllables, it is sufficient to consider only the period of $0 \leq t_a < t_{syll}$. The overlapping of syllables can be neglected, too. The probability that the two losses are within a syllable is

$$\begin{aligned} P_{in.syll}(t_{width}) &= \tfrac{1}{t_{syll}} \textstyle\int_{t_a=0}^{t_{syll}} \left\{ \begin{matrix} 0 \text{ if } t_a + t_{width} \geq t_{syll} \\ 1 \qquad \text{otherwise} \end{matrix} \right\} \mathrm{d}t_a \\ &= \left\{ \begin{matrix} 0 & \text{if } t_{width} \geq t_{syll} \\ 1 - \frac{t_{width}}{t_{syll}} & \text{otherwise} \end{matrix} \right. \end{aligned} \tag{11}$$

If two losses are within a syllable, PESQ adds them not with an exponent of $p = 2$ but with $p = 6$ (2). Because it is not simple to remodel PESQ's algorithm,




(a) Impact of random frame losses on speech quality

(b) Importance for many losses: measured (PESQ, slashed) and predicted (new metric, dotted)

**Fig. 8.** Impact of random frame losses on the speech quality. For loss rates (AMR <3%, ITU G.711/729 <4%) the correlation between measured and predicted importance is very high.

we introduce the following heuristic function (Eq. (12) and Fig. 7), which shows a similar behavior as PESQ. For a loss distance longer than the length of a syllable, it simply sums up the importance values. If it is lower, the importance values are added but the sum is leveled with a factor $1-P_{in.syll}$. Also, if the distance is short, we use an another addition, which sums up the square importance values. Again, this later addition is leveled by the probability of $P_{in.syll}$. Actually, we also tested to add cubics of importance values to model the effect of $p = 6$ but this solution did not let to higher correlation coeffiency.

$$add\left(imp_a, imp_b, t_{width}\right) = \begin{cases} imp_a + imp_b & \text{if } t_{width} > t_{syll} \\ \left(imp_a + imp_b\right)\frac{t_{width}}{t_{syll}} + \sqrt{{imp_a}^2 + {imp_b}^2}\left(1 - \frac{t_{width}}{t_{syll}}\right) & \text{otherwise} \end{cases} \quad (12)$$

Equation 11 partially models the time-frequency masking effect, which causes a masking of minor distortions by nearby louder ones. However, PESQ models the temporal masking effect only in the statistical mean. PESQ's masking is stronger – or at least longer – than the pre- or postmasking effect. It can be seen if one compares Fig. 3 with 7. This observation explains why it was not necessary to add time masking effects to PESQ: It is already included.

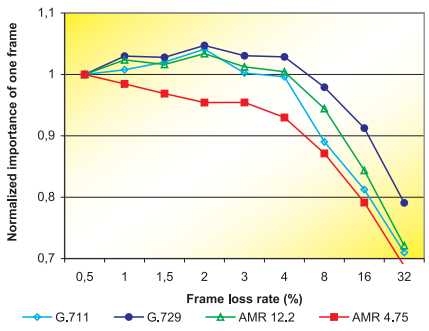



**Fig. 9.** Impact of loss rate on the mean, normalized importance using the new metric.

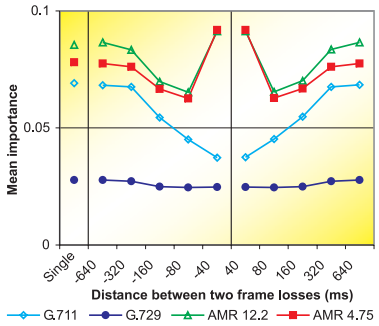



**Fig. 10.** Mean importance of a single lost frame is given on the very left and the importance of the second lost frame is displayed in relation to its distance from the first lost frame.

## 5 Validation

We consider a scenario, in which frame losses occur randomly, and we determine the speech quality for a given frame loss rate. The same scenario has been conducted in [8] with the old metric based on (4). The experimental set-up is described briefly: We follow recommendation ITU P.833, conduct many instrumental speech quality measurements, and vary the coding, the sample, and the loss pattern: A speech sample is encoded, frame losses are enforced depending on the experimental requirements (e.g., random frame loss), the frames are decoded or concealed, and finally PESQ calculates the MOS value by comparing the original sample with the degraded version.

Figure 8a displays the relation between the rate of random frame losses and speech quality for different codecs: The higher the loss rate the worse are PESQ's speech quality ratings. Next, we calculate the importance of all frame losses (Fig. 8b). At a loss rate of 0% the importance value is 0. As long as the loss rate is low, the importance increases linearly with the loss rate.

If the impact of frame losses can be added, the following statement is valid: The overall importance of the loss of $N$ frames can be calculated by multiplying the mean importance with $N$ (13). In Fig. 9, the mean importance depending on the loss rate is displayed. For low loss rates the importance is a bit underestimated. In case of loss rate above 8%, it is clearly underestimated. One should note that in this experiment the masking is not considered.

$$Imp\,(s, c, l_{mean}) \cdot N = Imp\,(s, c, \{l_1, \ldots, l_N\}) \tag{13}$$

The next experiments reseembles the measurements of single frame losses described in [8], but this time we dropped two speech frames instead of one. Between both losses there is a lossless gap of 40, 80, 160, 320, or 640 ms. In Fig. 10

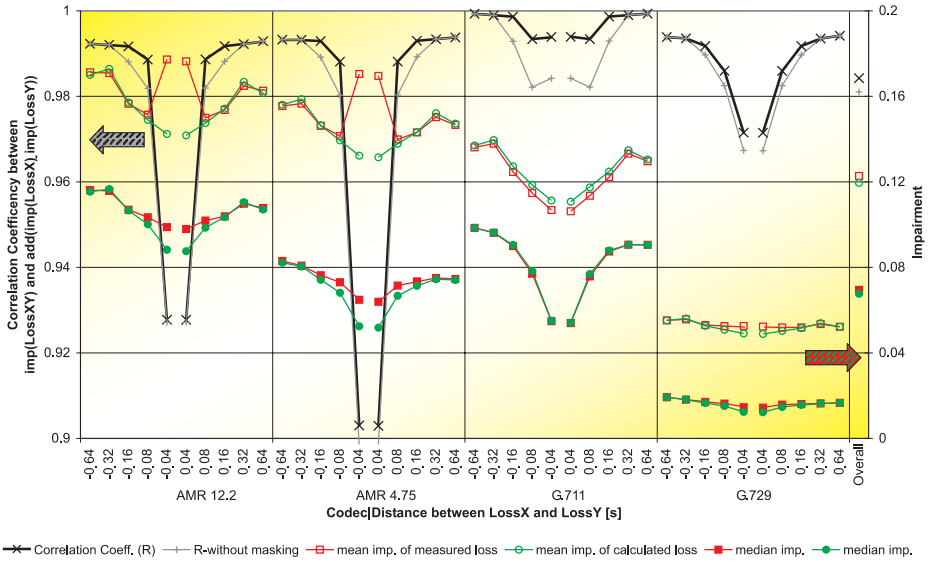



**Fig. 11.** Median and mean impairment due to two frames losses. Estimated with PESQ displaying $Imp\left(s,c,\{l_1,l_2\}\right)$ and with our model: $add\left(Imp\left(s,c,l_1\right),Imp\left(s,c,l_2\right),t_{width}\right)$. We also show the cross correlation between PESQ ratings and the rating of our model (R value). The R-without-masking compares PESQ with the long-term only aggregation function.

we display the importance averaged over all single frame losses, vertically sorted according to the encoding scheme and marked with Single on the horizontal axis. Also, we display the importance of the second frame $l_2$, if the first frame $l_1$ is lost already. The importance value is calculated using (14).

$$Imp\left(s,c,\{l_2 \mid l_1\}\right) = \Big(\left(4.5 - MOS\left(s_i,c,\{l_1,l_2\}\right)\right)^2 - \left(4.5 - MOS\left(s_i,c,\{l_1\}\right)\right)^2\Big) \cdot t\left(s\right) \tag{14}$$

Considering the G.711 results, one can see that the nearer the frame losses are, the lower the importance of a frame becomes. This effect can be explained with the temporal masking effect [14]. However, the mean importance for two AMR frame losses increases significantly, if the loss distance is 40 ms. We assume that this effect is due to a mismatch between the encoder's and decoder's internal state. The first loss results into to a desynchronized decoder. The applied loss concealment leads to a wrong prediction of frames' content. Since the de-synchronisation of the decoder can last for multiple following frames (up to 700 ms [16]), the mean impairment due to the concealment of the second loss can be significantly higher. This effect occurs only with the AMR codec, thus we will not consider it in this work any further.

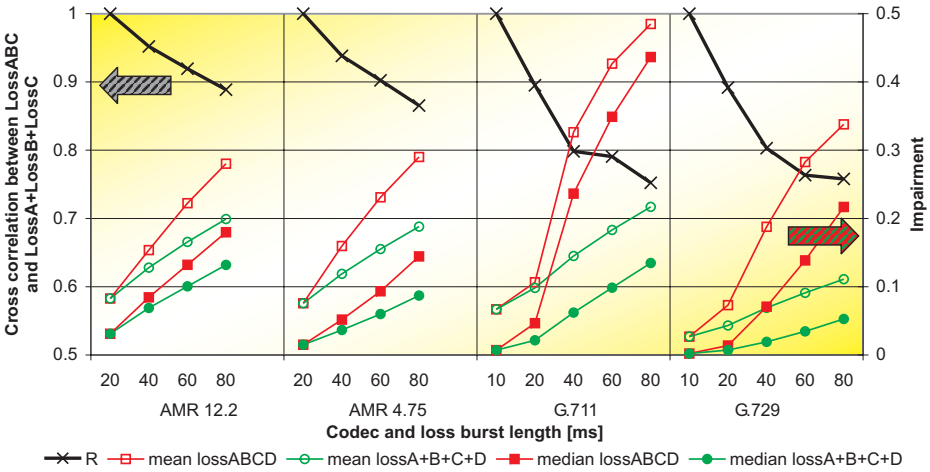



**Fig. 12.** Importance of a block of frame losses (red/square: PESQ, green/circle: our approach, black lines/cross: cross correlation R)

Figure 11 displays the prediction performance of losing two speech frames compared to the sum of losing two individual frames. The correlation coefficient between the importance of the double loss case and the sum of both single loss case is calculated and displayed. The correlation for a distance of >320ms is about R>0.98 and drops down to a minimum of R=0.78 at a distance of 40ms. This effect can be explained with concealment and error propagation effects that are not modeled in our model.

In the next experiment we study the effect of bursty frame losses. We dropped one block of continuing speech frames within a sample length of 8 s. The duration of the complete block was between 10 to 80 ms (in Fig. 12 the red lines marked with a square). Also we used our model to add the importance of the corresponding single frame loss (the green lines marked with a circle). To calculate the importance of the burst loss we use (15) with $N$ being the number of continuously lost frames, $pos$ the position of the first lost frame, and $Imp^*_{pos}$the importance of a frame loss at position $pos$.

$$Imp^*\left(N, pos\right) = \begin{cases} Imp^*_{pos} & \text{if } N = 1 \\ add\left(Imp^*\left(N-1, pos\right), Imp^*_{pos+N-1}, 0\right) & \text{if } N > 1 \end{cases} \tag{15}$$

The correlation (R) between PESQ and our model is displayed with black lines. The longer the loss burst, the worse the cross correlation. Our model give a lower impact of bursty losses as PESQ. This modelling is in line with the indications that PESQ displays an obvious sensitivity to bursty losses judging them worse than humans do [2].

## 6 Conclusion

This paper describes the impact of speech frames loss by considering their temporal relation. It is based on the concept of the *importance of speech frames* and models psychoacoustic aggregation behavior over time. Thus, our model covers an important aspect in the relation between speech frame losses and speech quality. Our model shows a high prediction accuracy for many loss patterns when compared to PESQ. Additionally, our time aggregation function has a very low complexity. However, before it can be fully applied, three issues have to be addressed:

First, the measurement of speech frame importance in [8] has a high computational complexity and delay. Thus, it cannot be applied online. We provide solutions in [16] that decrease delay and complexity at cost of a lower prediction accuracy. The question remains whether the lower prediction accuracy influences the performance of our time aggreation function.

Second, we compare the performance of our aggregation algorithm to the same PESQ algorithm, which we used to derivate and remodel our algorithm. We achieve a high prediction performance. However, it is still an open point how well our algorithm performs, if it is compared to subjective listening-only test results. The verification with databases containing subjective results is subjected of future studies.

Last, further studies are required to see how our metric scales at high loss rates. Definitely, the effects of concealment and error propagation play an important role if losses are frequent or bursty and need to be modeled.

Nevertheless, the given results contribute to research and standardization: First, they enable researchers developing communication protocols to model the impact of frame loss with a high accuracy. For example, it can be applied for algorithms that prevent frame loss burstiness. Second, this work provides also feedback to the developers of PESQ or similar algorithms, as it explains why PESQ does not require temporal masking: It was already included. Third, it identifies weaknesses of frame loss concealment algorithms (e.g. AMR). Last but not least, our work is directly intended for the standardization process of ITU-T P.VTQ, as it can been seen as an alternative or complementary algorithm to the ITU's E-Model, Telchemy's VQmon and Psytechnics' psyVOIP algorithms [17,18], which relate VoIP packet loss and delay to service quality.

To show the relevance of the questions addressed in this paper, we demonstrate the impact of the packet dropping strategy on speech quality: Using the knowledge about frame importance, simulations and informal listening-only tests show that only a fraction of all speech packets need to be transmitted if (at least) speech intelligibility is to be maintained. Knowing the importance of speech frames might allow significant energy savings on wireless phones, because fewer packets need to be transmitted.

## 7 Acknowledgements

We like to thank A. Raake, J. G. Beerends and C. Schmidmer for their feedback, E.-L. Hoene for the revision of this paper, and last not least the reviewers for their excellent comments.

## References

1. Mohammed, S., Cercantes-Perez, F., Afifi, H.: Integrating networks measurements and speech quality subjective scores for control purposes. In: Infocom 2001. Volume 2., Anchorage, AK, USA (2001) 641–649
2. Sun, L.: Subjective and objective speech quality evaluation under bursty losses. In: MESAQIN 2002, Prague, CZ (2002)
3. Jiang, W., Schulzrinne, H.: Comparison and optimization of packet loss repair methods on voip perceived quality under bursty loss. In: NOSSDAV. (2002) 73–81
4. Sanneck, H., Tuong, N., Le, L., Wolisz, A., Carle, G.: Intra-flow loss recovery and control for VoIP. In: ACM Multimedia. (2001) 441–454
5. Petracca, M., Servetti, A., De Martin, J.C.: Voice transmission over 802.11 wireless networks using analysis-by-synthesis packet classification. In: First International Symposium on Control, Communications and Signal Processing, Hammamet, Tunisia (2004) 587–590
6. Chou, P., Miao, Z.: Rate-distortion optimized streaming of packetized media. Technical Report MSR-TR-2001-35, Microsoft Research Technical Report, Redmond, WA (2001)
7. Hoene, C., Dulamsuren-Lalla, E.: Predicting performance of PESQ in case of single frame losses. In: MESAQIN 2004, Prague, CZ (2004)
8. Hoene, C., Rathke, B., Wolisz, A.: On the importance of a VoIP packet. In: ISCA Tutorial and Research Workshop on the Auditory Quality of Systems, Mont-Cenis, Germany (2003)
9. ITU-T: Recommendation P.862 - Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-To-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs (2001)
10. Rix, A.W., Hollier, M.P., Hekstra, A.P., Beerends, J.G.: Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part I - time alignment. Journal of the Audio Engineering Society **50** (2002) 755–764
11. Beerends, J.G., Hekstra, A.P., Rix, A.W., Hollier, M.P.: Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment, part II - psychoacoustic model. Journal of the Audio Engineering Society **50** (2002) 765–778
12. Hoene, C., Wiethölter, S., Wolisz, A.: Predicting the perceptual service quality using a trace of VoIP packets. In: QofIS'04, Barcelona, Spain (2004)
13. Beerends, J.G.: Measuring the quality of speech and music codecs: An integrated psychoacoustic approach. presented at the 98th Convention of the Audio Engineering Society, preprint 3945 (1995)
14. Zwicker, E., Fastl, H.: Psychoacoustics, facts and models. Springer Verlag (1990)
15. Beerends, J.G., Stemerdink, J.A.: A perceptual speech-quality measure based on a psychoacoustic sound representation. Journal of the Audio Engineering Society **42** (1994) 115–123
16. Hoene, C., Schäfer, G., Wolisz, A.: Predicting the importance of a speech frame (2005) work in progress.
17. Telchemy: Delayed contribution 105: Description of VQmon algorithm. ITU-T Study Group 12 (2003)
18. Psytechnics: Delayed contribution 175: High level description of psytechnics ITU-T P.VTQ candidate. ITU-T Study Group 12 (2003)