2

Assessment of Speech Quality in VoIP

Zdenek Becvar, Lukas Novak and Michal Vondra
Czech Technical University in Prague, Faculty of Electrical Engineering
Czech Republic

1. Introduction

In VoIP (Voice over Internet Protocol), the voice is transmitted over the IP networks in the form of packets. This way of voice transmission is highly cost effective since the communication circuit need not to be permanently dedicated for one connection; however, the communication band is shared by several connections. On the other hand, the utilization of IP networks causes some drawbacks that can result to the drop of the Quality of Service (QoS). The QoS is defined by ITU-T E.800 recommendation (ITU-T E.800, 1994) as a group of characteristics of a telecommunication service which are related to the ability to satisfy assumed requirements of end users. The overall QoS of the telecommunication chain (denoted as end-to-end QoS) depends on contributions of all individual parts of the telecommunication chain including users, end devices, access networks, and core network. Each part of the chain can introduce some effects which lead to the degradation of overall speech quality. Lower speech quality causes user's dissatisfaction and consequently shorter duration of calls (Holub et al., 2004) which reduces profit of telecommunication operators. Therefore, both sides (users as well as operators or providers) are discontented.

The end device decreases speech quality by coding and/or compression of the speech signal. The speech quality can be also influenced by a distortion of the speech by its processing in the end device e.g. in the manner of filtering. It can lead to the saturation of the speech, insertion of a noise, etc. The processed speech is carried in packets via routers in the networks. Individual packets are routed to the destination as conventional data packets. Therefore, the packets can be delayed or lost. According to ITU-T G.114 recommendation (ITU-T G.114, 2003), the delay of speech should be lower than 150 ms to ensure high quality of the speech. Each packet is routed independently; therefore the delay of packets can vary in time. The variation of packet delay is usually denoted jitter.

The impact of all above mentioned effects on the speech quality can be evaluated either by subjective or objective tests. The first group, subjective tests, uses real assessments of the speeches by users. Therefore it cannot be performed in real-time. The second set of tests, objective tests, tries to estimate the speech quality by speech processing and evaluation.

The rest of chapter is organized as follows. The next section gives an overview on the related work in the field of VoIP speech quality. The third one describes basic principles of the speech quality assessment. The speech processing for all performed tests are described in section four. Section five presents the results of realized assessments of the speech quality. Last section sums up the chapter and provides major conclusions.

2. Related works

Voice packets transmitted over the IP network may be lost or delayed. In non-real-time applications, packet loss is solved by appropriate control protocol, e.g., Transfer Control Protocol (TCP) by retransmission. This solution is not suitable for voice transmission since it increase the delay of voice packets (Linden, 2004).
Clear advantage of VoIP is the ability to use wideband codecs. However, higher transmission bandwidth results in high sampling frequency and escalates requirements on hardware components. Bandwidth of roughly 10 kHz is sufficient for sampling a speech signal. Nevertheless, 8 kHz bandwidth (16 kHz sampling frequency) is the best trade-off between bit rate and speech quality (Benesty et al., 2008). The extension of the bandwidth improves the intelligibility of fricative sounds such as '*s*' and '*f*' which are very difficult to distinguishing in conventional narrowband telephony applications (Benesty et al., 2008).
The impact of random packet losses for different packet sizes on the speech quality is evaluated in (Ding & Goubran, 2003). The results show that MOS (Mean Opinion Score) decreases more rapidly if larger packet size is used. These results are confirmed also in (Oouchi et al., 2002). The paper (Oouchi et al., 2002) presents the negative dependence between packet loss ratio and the speech quality. Moreover, the authors performed speech quality tests to show that the tolerance to packet losses is getting lower with higher duration of packets. In this chapter, we compare not only the speech quality over the length of lost packet as in previous paper, but also, the impact of losses in narrowband speeches is compared with wideband.
In real networks, bursts of packets are lost more frequently than single packets, due to effects such as network overload or router queuing (Ulseth & Stafsnes, 2006). The paper (Clark, 2002) presents a review of the effect of burst packet loss compared with random packet loss. The results are similar for small packet loss ratio (approximately up to 3 %). However, the quality of the burst packet loss is decreasing more significantly than in case of the random packet loss for higher packet loss ratio. In this chapter, we additionally consider two different bandwidths of speech, i.e., 3.1 kHz and 7 kHz.
Packet losses can be eliminated by using Packet Loss Concealment (PLC) algorithms. These algorithms can replace missing part of the speech signal and make a smooth transition between the previous decoded speech and lost segments. Several PLC algorithms is described, e.g., in (Kondo & Nakagawa, 2006); (Tosun & Kabal, 2005). The PLC algorithms are based on various methods, each of them more or less suitable for specific use. All PLC algorithms work with the frequency characteristic of speech. Therefore, one of the criterions for the right choice of the proper PLC algorithm can be its frequency characteristic since each frequency band is perceived individually by human ear (Fastl & Zwicker, 1999); (Robinson & Hawksford, 2000). In this chapter, influence of the harmonic distortion on the speech quality is analyzed. Harmonic distortion cause unequal transfer of all frequency components. With the knowledge of this influence, the suitable frequency characteristic of PLC method can be chosen. The harmonic distortion analysis is based on the Mel-cepstrum (Molau et al., 2001) as it follows the psychoacoustic model of human sound perception.
In general speech quality assessment, the random placement of packet losses is assumed (Ulseth & Stafsnes, 2006). Nevertheless, the placement of the packet loss can significantly influence the final speech quality evaluation since each phonetic element can have different significance for the voice intelligibility. For example, speech sound carrying high energy (e.g., voiced sound) is more important than the low energy one (e.g., unvoiced sound) (Sing &

Chang, 2009); (Bachu et al., 2010). The knowledge of an impact of losses of individual phonetic elements can be exploited in design of codecs or for speech synthesis. In the paper (Sun et al., 2001), the authors proofs the more noticeable impact of losses placed in the voiced sounds than unvoiced sounds on the speech quality. All phonetic elements are classified into voiced and unvoiced without any further division. In this chapter, we consider further classification on smaller groups of phonetic elements. Moreover, the subjective listening tests are performed.

3. Speech quality assessment in VoIP

Speech quality can be measured and rated either by using by R-factor or by MOS scale (ITU-T P.800.1, 2003). The R-factor is based on E-model (ITU-T G.107, 2005). Its range is from 0 to 100. The R-factor represents level of user's satisfaction with the speech. The expression of user's satisfaction related to the R-factor is presented in Table 1.

R-factor	Quality	Users' satisfaction
90 – 100	Best	Very satisfied
80 – 89	High	Satisfied
70 – 79	Medium	Some users dissatisfied
60 – 69	Low	Many users dissatisfied
50 – 59	Poor	Nearly all users dissatisfied

Table 1. Users' satisfaction with the speech quality according to R-factor.

The parameter MOS is an average value from given range, which is used for assessment of the speech quality by subjects (persons). It is in the range from 1 to 5 (see Table 2). Both subjective and objective methods give their outputs in specific MOS scale; nevertheless it is possible to convert them. The scales for objective and subjective listening tests are denoted MOS-LQO (MOS-Listening Quality Objective) and MOS-LQS (MOS-Listening Quality Subjective) respectively. Another type of MOS, MOS-LQE (MOS-Listening Quality Estimated) is derived from R-factor.

MOS	Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 2. Speech quality expressed by MOS scale.

In this chapter, the MOS scale is considered since it is largely used in praxis.
The speech quality can be assessed either by a subjective or by an objective test. The practical utilization of both types is related to their principle.

3.1 Subjective tests

The first group of test is subjective speech quality assessment. This test uses statistical evaluation of assessments of the several speeches by real persons. To obtain valid and precise results, several requirements and parameters must be fulfilled. These requirements

for subjective tests are defined by ITU-T recommendations P.800 (ITU-T P.800, 1996) and P.830 (ITU-T P.830, 1996). The documents define procedures of selecting speakers for recording of source speeches, methods for recording and preparing of input samples, required quantities and formats of speeches, parameters of testing environment and methods for selection and guidance of test attendants. The content of individual phases of the experiment and the content of the speeches are not defined.
Two types of subjective tests can be performed: conversational and listening. The conversational tests are executed in laboratory environment. Two tested objects (persons) are placed in the separated sound proof rooms with suppressed noise. Both persons are performing phone call and assessing its quality.
The requirements on the listening tests are not too high as on conversational test. The process of listening speech quality assessment is as follows. First, the speeches are transmitted over telecommunication chain. Then, the tested objects listening several speeches and assess it according to MOS-LQS. The final value of MOS is calculated as an average over all tested objects and all speeches obtained under the same conditions.
In all of our subjective and objective tests, every individual speech is modified in MATLAB software to obtain specific degradation of the speech. The speech processing is applied to be inline with conventional degradation of the speech in the common telecommunication equipments. More details on the speech processing are described in section four of this chapter.

3.2 Objective tests

The realization of the subjective test is very time consuming. Therefore, the objective tests are defined for speech quality measurement. These tests are based on substitution of the subjective tests by appropriate mathematical models or algorithms. The objective tests can be divided on intrusive and non-intrusive. The intrusive one uses two speeches for determination of final speech quality. The first speech is original non-degraded speech and the second one is the same speech degraded by transmission over the telecommunication chain. It enables to obtain more precise results, however this method cannot be used for real-time speech quality measurement. On the other hand, non-intrusive methods do not require the original source speech, since the evaluation of speech quality is based only on the degraded speech signal processing. Hence, the non-intrusive methods are suitable for real-time monitoring of speech quality. The non-intrusive methods are standardized by series of recommendations ITU-T P.56x such as ITU-T P.561 recommendation known as INMD (In-service Non-intrusive Measurement Device), its enhancement ITU-T P.562, or ITU-T P.563 denoted as 3SQM (Single Sided Speech Quality Measurement) which contains all procedures defined by ITU-T P.561 and ITU-T P.562 and additionally, it considers several new aspects such as additional noise, distortion, or time alignment.
One of the first largely expanded intrusive methods was PSQM (Perceptual Speech Quality Measurement) defined by recommendation ITU-T P.861. This recommendation was replaced by another one, labeled ITU-T P.862, in 2001 since the former one cannot evaluate some effects such as jitter, short losses, signal distortion, or impact of low speed codecs in compliance with human perception. The latter recommendation is generally denoted PESQ (Perceptual Evaluation of Speech Quality) (ITU-T P.862, 2001). The PESQ is one of the most widely used objective methods developed for end-to-end speech quality assessment in a conversational voice communication.

The principle of PESQ is depicted in Fig. 1. The PESQ method is based on the comparison of original (non-degraded) signal *X(t)* with degraded signal *Y(t)*. The signal *Y(t)* is result of a transmission of the signal *X(t)* through a communication system. The PESQ method generates a prediction of the quality which would be given to the signal *Y(t)* in subjective listening test.

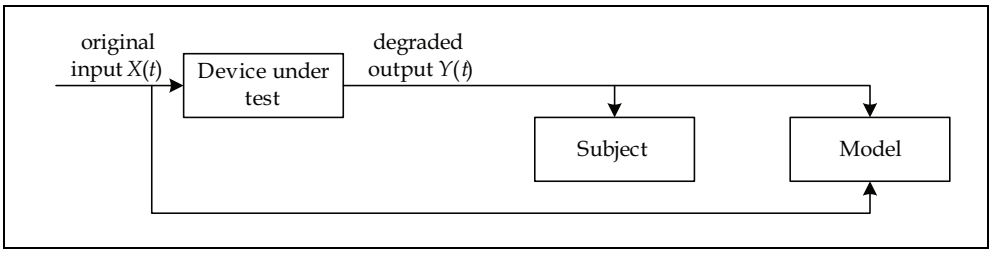

Fig. 1. Principle of PESQ method for speech quality assessment.

The range of the PESQ MOS score (according to ITU-T P.862) is between -0.5 and 4.5. Since this range does not match to a scale used for the subjective test, the ITU-T P.862.1 recommendation (ITU-T P.862.1, 2003) enables to recalculate the PESQ MOS to better comport the results of subjective listening test. The range of converted value is from 1 to 4.55. The recalculation is defined by the next formula:

$$y = 0.999 + \frac{4.999 - 0.999}{1 + e^{-1.4945 \times x + 4.6607}} \tag{1}$$

where x is an objective PESQ MOS score and y is a matching ITU-T P.862.1 MOS score.
The ITU-T P.862 recommendation describes all requirements on the tested speeches (e.g. character of speech signals, duration of speech, etc.). Frequency characteristics of the speech signal and signal level alignment must be in accordance with recommendation ITU-T P.830.

4. Speech processing for speech quality assessment

The speeches used for subjective and objective tests are original digital studio recordings, obtained by separation from dialogs of two people. Their content does not imply any emotional response at the listeners. All recordings are in Czech language spoken by natural born Czech speakers without speech aberrations.
The length of utterances is selected to fulfil the requirements of both types of tests. Therefore it is between 8 an 12 s. Tested signals is coded by 16-bit linear PCM (Pulse Code Modulation) sampled with 8 kHz or 16 kHz sample rate (down sampled from the original studio quality recordings sampled with 48 kHz).
Speeches are modified in MATLAB in line with speech processing in conventional telecommunication chain. Furthermore, the analysed phenomenons, such as distortion or packet losses are also incorporated in MATLAB.
Three individual cases are analysed: i) comparison of the impact of individual and consecutive packet losses of conventional telecommunication frequency bandwidth (3.1 kHz) with wideband systems (7 kHz); ii) analysis on the impact of harmonic distortion; iii) impact of the losses of individual phonetic elements on the speech quality.

4.1 Speech processing for losses of packets for narrow and wideband channels

Forty speeches are used for the analysis of the impact of bandwidth and packets length on speech quality. Each of the speeches is encoded using PCM. The speeches sampled with 8 kHz frequency are filtered according to recommendation ITU-T G.711 with bandwidth of 3.1 kHz (from 300 Hz to 3400 Hz) (ITU-T G.711, 1988). The speeches sampled with frequency 16 kHz (ITU-T G.711.1, 2008) are filtered according to recommendation ITU-T G.711.1 with bandwidth of 7 kHz.

At the beginning, the analyzed speech is split into sections with the same length. The individual sections can be understood in terms of packet that will be transmitted through a network. The length of all packets is chosen, in each round of analysis, to be equal to the following values: 10, 20, 30, or 40 ms. These lengths are the most frequently used in practice for the transmission and the evaluation (Hassan & Alekseevich, 2006). The lengths of packets correspond to 80, 160, 240, or 320 samples per packet for the speeches sampled with 8 kHz and 160, 320, 480, or 640 samples for the speeches sampled with 16 kHz.

After the division of certain analyzed speech to packets, random vector V_L=$\{v_{L1}, v_{L2}, ..., v_{LR}\}$ is generated. The number of elements in the vector (R) corresponds to the number of packets, to which the speech is divided. The random vector V_L contains random numbers in the range from 0 to 1 generated with uniform distribution. The number of vector elements does not depend only on the length of the speech, but depends also on the length of individual packets.

The packet losses are randomly determined and the number of lost packets is calculated according to the Packet Loss Ratio (PLR) and the total number of packets in the speech. The original random vector V_L is consequently recalculated to vector of packet losses (V_{PL}=$\{v_{PL1}, v_{PL2}, ..., v_{PLR}\}$) according to subsequent formula:

$$\begin{aligned} v_{PLr} &= 1 \qquad \forall v_{Lr} \geq T_0, \;\; 0 < r < R \\ v_{PLr} &= 0 \qquad \forall v_{Lr} < T_0, \;\; 0 < r < R \end{aligned} \tag{2}$$

where T_0 is the threshold for setting up the element to zero. The threshold is determined according to the required PLR. The vector V_{PL} contains only numbers "one" and "zero", where zeros represent the losses.

The final speech is a product of multiplication of the modified vector V_{PL} and the relevant speech split into R packets. Packets, which are multiplied by zero corresponds to the lost parts, and packets multiplied by the number one remained unchanged.

The total duration of lost packets is the same for all speeches with the same PLR, regardless of the amount of packets contained in a speech (R). For example, if the speech is divided into packets with 10 ms length, the number of these packets is twice the number of packets created in case of the packets with 20 ms length.

Random vector v_L is generated twenty times for each value of PLR, each sections length, and for each of the speeches. Repetition of generation of the random vector limits negative effects of random drop of losses, which could affect results.

The random losses of packets, described above, are characterized by random appearance in time. Beside this, the consecutive losses frequently occur in real networks. For example, during the short outage in communication e.g. during overload of a node, the loss of only one packet is not very likely to happen. More probable is the loss of several packets. Therefore, the consecutive packet loss can be expressed as a loss of sequence of subsequent packets.

Speech processing of consecutive packet losses is similar to the generation of the individual losses, with slight modification to follow the effect of losing packets in groups. Randomly generated vector V_L is thus shortened to the length r. The length r of the new reduced vector $V_R = \{v_{R1}, v_{R2}, ..., v_{Rr}\}$ is calculated according to the next formula:

$$r = R - \left(P \cdot \left(n_{CP} - 1 \right) \right) \tag{3}$$

where R is the number of packets in the whole speech; n_{CP} is the amount of packets in the consecutive loss; and P represents the number of elements in the reduced vector that will be set to zero. The P can be determined by the following equation:

$$P = \frac{PLR \cdot r}{n_{CP}} \tag{4}$$

In all cases, the speech quality assessments are performed for up to twenty consecutive lost packets due to the limitation by the length of speeches according to the ITU-T P.862 recommendation. The considered PLR for this test is 10 %. The shortest of the analyzed speeches has length of 8 s. For the packets with duration of 40 ms in groups of twenty packets, it gives the overall duration of the 800 ms. It is 10 % of the speech with duration of 8 s. Hence, the higher length of consecutively lost packet cannot be accommodated into these speeches.

To eliminate the random factor of position of consecutive packet loss, twenty repetitions for each of the speeches and for each of the length of group of consecutive packet loss are performed, as in the case of the random individual packet losses to suppress the affection of results by effect of placement of losses into different parts of speeches.

4.2 Speech processing for investigation of harmonic distortion

Individual frequency bands of speech are suppressed for the analysis of harmonic distortion influence i.e. all speech components with frequencies of that band. For our analysis, the narrowband telecommunication channel (300 – 3400 Hz) is separated to four sub-bands. Lower count of frequency bands would give less information on individual frequency influence and higher count would distort the speech too little as it would be undistinguishable from the original non-distorted speech. Another parameter which has to be set is the choice of corner frequencies of these bands. One possibility is to take them linearly according to the telecommunication channel band and the other one is to choose corner frequencies nonlinearly with unequal bandwidth of each band. An ear perceives each frequency differently, which means that if the frequency of perceiving tone will increase linearly, the listener will subjectively sense only logarithmic increase. Relating to this fact, the corner frequencies are chosen with geometrical interval. Mel-frequency (*mel*) band for the calculation of corner frequencies is defined by the subsequent equation:

$$mel = 2595 \cdot \log\left(1 + \frac{f}{700} \right) \tag{5}$$

The communication band is divided into four Mel-frequency bands with equal wide. The resulting Mel-frequencies are presented in Table 3. Note that the lowest band (band 1) is extended to 50 Hz.

Band	1	2	3	4	
Mel [mel]	402	800	1197	1595	1992
Frequency f [Hz]	300	723	1325	2181	3400
Designed f_c [Hz]	50	723	1325	2181	3400

Table 3. Frequency bands used for harmonic distortion.

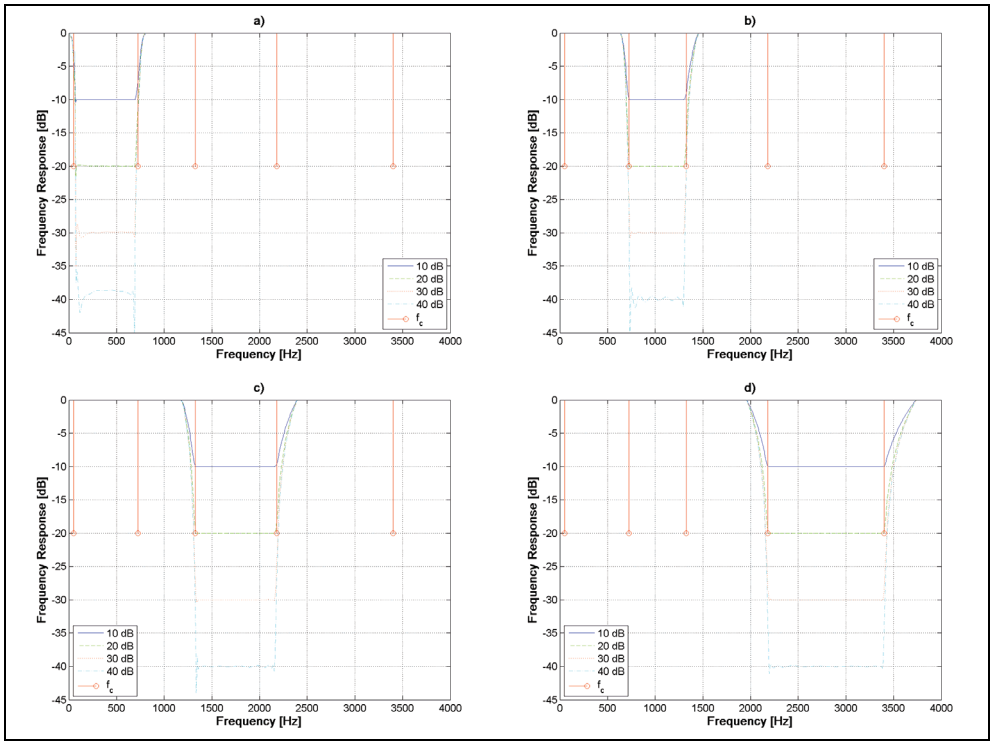

Fig. 2. Frequency responses of band stop filter for all four levels of suppression and four Mel-bands a) band 1; b) band 2; c) band 3; d) band 4.

Band stop filters designed in MATLAB by Yulewalk method (Friedlander & Porat, 1984) is used for filtering of speeches. The Yulewalk method designs recursive IIR (Infinite Impulse Response) digital filters by fitting a specified frequency response. Each band is suppressed in four levels: 10, 20, 30, and 40 dB. Frequency responses of band stop filters for all four bands and suppression levels are shown in Fig. 2.

4.3 Speech processing for losses of individual phonetic elements

Following approach is considered for the analysis of the relevance of individual group of phonetic elements on the speech quality. The algorithm for generation of packet losses with exact ratio is modified to place the lost packet only to the specific location of the particular phonetic elements. It requires the phonetic transcription of the speech and subsequent exact definition of the first and the last samples of each phonetic element. Since no reliable

automatic algorithm is available, it was done manually. The speech is transcript to CTU IPA (Hanzl & Pollak, 2002) which is well transparent and optimized for processing by computer. All individual speech sounds of the speech are considered as analyzed phonetic elements. All phonetic elements are classified according its lexical meaning into four groups: i) vowels & diphthongs; ii) nasal & liquids; iii) plosives & affricates; iv) fricatives. The vowels and diphthongs are always voiced sounds, therefore they carry higher energy. Also the nasals and liquids are predominantly voiced with higher energy level. The plosives and affricates as well as fricatives contain either voiced or unvoiced consonants, which have the same mechanism of its origination in voice organs. The classification of the speech sounds into groups is presented in Table 4.

Group	Speech sounds
Vowels & diphthongs	‘a‘,‘á‘,‘e‘,‘é‘,‘i‘,‘í‘,‘y‘,‘ý‘,‘o‘,‘ó‘,‘u‘,‘ú‘,‘au‘,‘eu‘,‘ou‘
Nasal & liquids	‘m‘,‘n‘,‘ň‘,‘r‘,‘l‘
Plosives & affricates	‘p‘,‘b‘,‘t‘,‘ť‘,‘d‘,‘ď‘,‘k‘,‘g‘,‘c‘,‘č‘,‘dž‘,‘dz‘
Fricatives	‘f‘,‘v‘,‘s‘,‘z‘,‘ř‘,‘Ř‘,‘š‘,‘ž‘,‘j‘,‘ch‘,‘h‘

Table 4. Classification of the speech sounds into groups for speech quality assessment purposes.

For the speech processing in MATLAB, we assume 10 ms packet length; it corresponds to 80 samples per packet. The speech is furthermore processed in the following way.
First, the amount of packets contained only in speech sounds of investigated group of phonetic elements is determined (denoted R). Then, the coefficient K, which expresses the ratio of all packets in the speech (denoted T) to the amount on packets in the specific phonetic elements, is derived as follows: $K=T/R$. Subsequently, the coefficient K is recalculated to the new one (denoted H), which corresponds to the loss ratio only among selected group of phonetic elements; $H=K*PLR$, where PLR is a packet loss ratio related to the overall speech. The coefficient H expresses the probability of loss of each packet in investigated group of phonetic elements. Next, the vector of losses $V_g=\{v_{g1}, v_{g2}, ..., v_{gR}\}$ is randomly generated according to the coefficient H. The length of V_g is equal to R and each element of V_g represent whether the packet belonging to an element of a selected group will be lost or not. Furthermore, the new vector $V_s=\{v_{s1}, v_{s2},...,v_{sT}\}$ is created from vector V_g by filling up vector V_g to the length of complete speech T by insertion of “no loss” labels to the position of phonetics elements not belonging to the group of investigated phonetic elements. At the end, the packets labeled as “lost” are replaced by zeros (silence).

5. Results of speech quality assessment

As mentioned in previous section, three types of speech modification are investigated. This section provides the results of all performed tests.

5.1 Impact of losses of packets for narrow and wideband channels

Since the PESQ is not designed to evaluate the wideband speeches, the recalculation of output of conventional ITU-T P.862 PESQ according to the subsequent equation is required (Barriac et al., 2004):

$$y = 1 + \frac{4}{1 + e^{-2x+6}} \tag{6}$$

The results of objective tests for five packet lengths and two bandwidths are presented in Fig. 3. While maintaining all speech packets (no packets are lost), the speech quality reach the maximum. Of course this maximum is independent on the packet length. By increasing the *PLR* the significant speech quality degradation can be observed from Fig. 3.

The comparison of speeches with 3.1 kHz bandwidth and speeches with 7 kHz bandwidth show a higher rating of speech with lower frequency bandwidth (3.1 kHz) for $PLR \geq 4$ %. On the other hand, the speeches with bandwidth of 7 kHz achieve higher score than speeches with 3.1 kHz bandwidth (by approximately 0.3 MOS) for $PLR < 4$ %. For PLR > 4%, the difference between both frequency bandwidths grows with *PLR* up to *PLR* = 12 % (for packet length 10 ms), where speech quality for both bandwidths differs the most. The maximum gap between results for both bandwidths is approximately up to 0.8 MOS. With further increase of the PLR, both bandwidths perform in closer way. The impact of bandwidth is also decreasing with higher duration of packets. For packet duration of 40 ms, the maximum difference between both bandwidths is up to 0.6 MOS. Only the results of objective test are presented in this section since the results of subjective one are included in (Becvar et al., 2008).

Based on the results of objective as well as subjective tests, the consideration of wideband codecs is profitable only for systems with very low PLR (up to 4%).

Fig. 3 also shows that the same *PLR* imply lower degradation of speech divided in shorter sections although the total ratio of lost part of the speech is always the same. Thus the separation of the speech into packets with 40 ms length results in a higher impairment than splitting the speech into 10, 20, or 30 ms segments. The largest divergence (comparing 10 and 40 ms packet lengths) in the speech quality rating over the packet length is at *PLR* = 4 % for speeches with bandwidth 7 kHz. This discrepancy is roughly 1.4 MOS. The largest difference for speeches with bandwidth 3.1 kHz is at *PLR* = 8 %, it is approximately 1.2 MOS). This analysis clearly shows that it is profitable to utilize shorter speech packets.

The evaluation of the quality of speech influenced by consecutive packet losses is performed under the same conditions as the tests of individual packet losses.

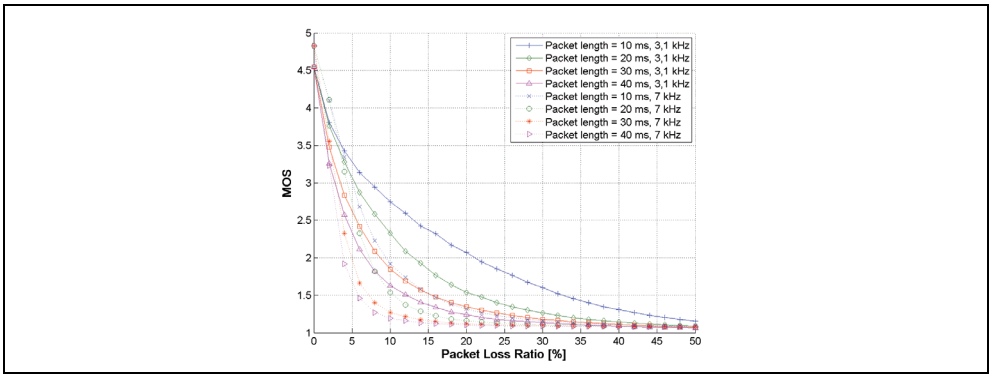

Fig. 3. Objective assessment of the different frequency bandwidth and lengths of packets over packet loss ratio.

The results plotted in Fig. 4 show the impact of the length of consecutive packet losses on the speech quality for objective assessment by PESQ method for 10 % *PLR*. Several lengths of packets are considered for analysis.
The longer duration of consecutive losses firstly lower the speech quality from approximately 2.7 MOS for 10 ms packet length and 3.1 kHz bandwidth. Then, the speech quality gradually increases with length of consecutive packet losses to 600 ms, where the MOS rating is again approximately 2.7 MOS. Further increase of the consecutive packet loss leads to only insignificant speech quality improvement. This effect is not noticeable for packet length over 40 ms, where the MOS score is continuously rising over the length of consecutive loss. From Fig. 4 can be determined the minimum amount of consecutive packet losses to obtain higher speech quality than in case of individual losses. This limit is sixty, ten and, three losses for packets with length of 10, 20, and 30 ms respectively for 3.1 kHz channel. The situation for 7 kHz channel is the same, however the final speech quality is lower (between 0.2 and 0.8 MOS) than the quality of 3.1 kHz.
Note that the PLR is equal to 10 %for all lengths of packets.

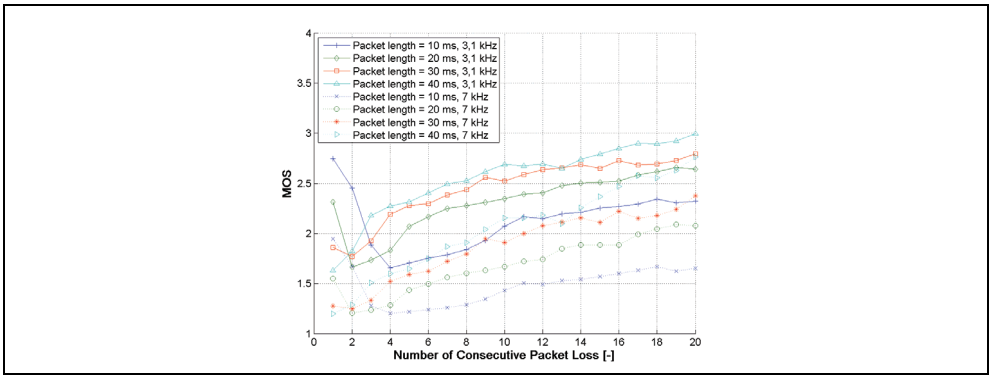

Fig. 4. Objective assessment of the impact of consecutive packet losses.

From Fig. 4 can be further seen that the lowest value of the objective speech quality is achieved for the overall duration of consecutive losses of 40 ms duration. This is achieved at four, two, or one consecutively lost packets with length of 10, 20, and 40 ms respectively.
The results also show that the same summarized duration of lost parts of speech are almost identical and independent on the length of individual packets. For example, MOS score for the loss of sections with 40 ms length is the same like for consecutive loss of bursts of two packets with 20 ms length or loss of bursts of four packets of 10 ms length. This fact is more noticeable in Fig. 5. This figure presents the impact of individual packet lengths over the overall length of consecutive losses. The results presented in Fig. 4 are converted into Fig. 5 with the new scale on x-axis by recalculation of x-axis according to the subsequent formula:

$$D_{total} = n_{CP} \cdot d \tag{7}$$

where d represents the length of packets.
The results presented in Fig. 5 depict that the worst rating is achieved for packet losses (either consecutive or individual) with overall duration of 40 ms. Hence, the division of packets to 40 ms length is not appropriate from the point of view of the individual losses.

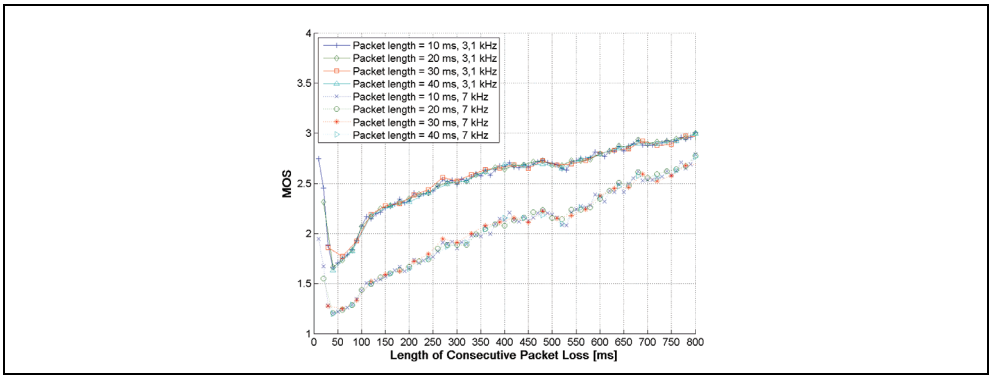

Fig. 5. The converted results of the objective tests of consecutive packet losses.

5.2 Impact of harmonic distortion

The subjective as well as the objective tests are considered for the evaluation of the impact of harmonic distortion. The subjective tests are prepared in accordance with ITU-T P.800 and ITU-T P.830 recommendations. Five different speeches for four bands and four suppression levels are processed for each distortion. It leads to 80 speeches (5 speeches * 4 bands * 4 levels) in total. Therefore, the overall duration of the subjective listening test is approximately 20 minutes per a listener.

Overall, 26 listeners participates on the subjective listening test. Software Tester (Brada, 2006) developed at Czech technical University in Prague is used for the listening test. The listeners participated on the subjective testing have been selected from students and employees of the university with respect to above mentioned recommendations.

The results are presented in Fig. 6 for the subjective tests and Fig. 7 for objective tests over four suppression levels (10, 20, 30 and 40 dB) in four bands.

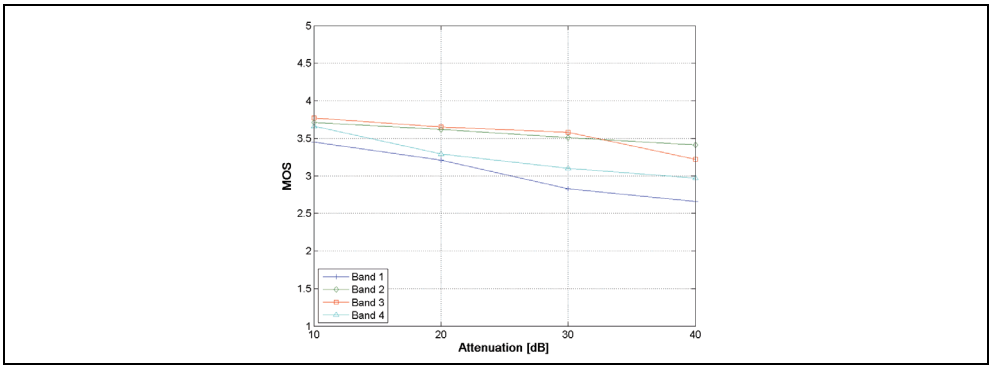

Fig. 6. Subjective assessment of the harmonic distortion.

The results show that the biggest influence at the reception quality of speech is obtained by the components contained in the first band (50 – 723 Hz). This phenomenon is caused by higher energy carried by lower frequency components in comparison with lower energy of higher frequency components. Therefore, the suppression of low frequencies causes

significant decrease in the speech quality. Only slightly lower impact is caused by the frequency components contained in the fourth frequency band (2181 – 3800 Hz). The lowest degradation of the speech quality is noticeable in the second and in the third bands (723 – 1325 Hz and 1325 – 2181 Hz). In all cases, the higher attenuation of the signal in individual bands leads to the decrease of the speech quality.
While the objective method PESQ is used for the speech quality assessment (see Fig. 7), only minor differences in the quality can be noticed for all four bands. The suppression of all bands has the similar impact on the speech quality according to PESQ. Also the impact of the level of attenuation is negligible since the drop of the speech quality is only between 0.25 and 0.5 MOS for all bands while attenuation varies between 10 to 40 dB.

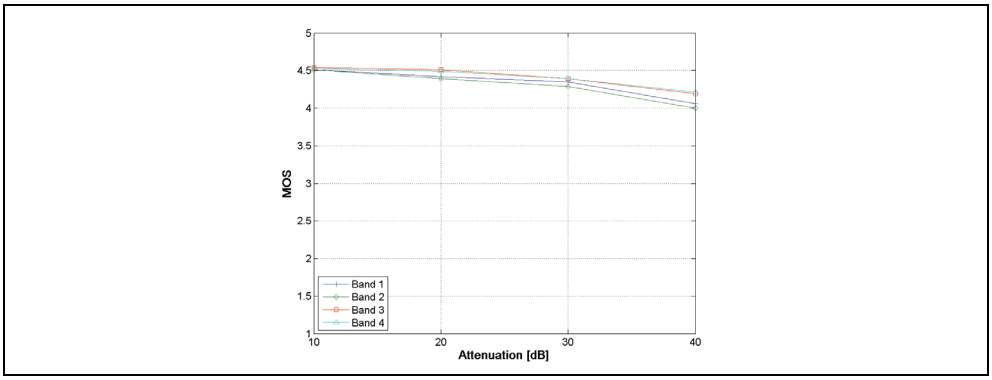

Fig. 7. Objective assessment of the harmonic distortion by PESQ method.

The average MOS score of subjective listening test and objective speech quality assessment by PESQ for individual bands is summarized in Table 5. The average subjective score is always lower than objective one. The difference between results of both tests is very significant for all bands (between 0.74 and 1.31 MOS).

Group	Subjective score (MOS)	Objective PESQ score (MOS)
Band 1	3.03	4.34
Band 2	3.56	4.30
Band 3	3.56	4.40
Band 4	3.26	4.41

Table 5. Average MOS score of each bands over levels of suppression.

5.3 Impact of losses in individual phonetic elements

The impact of individual phonetic elements is investigated by subjective listening tests and by PESQ objective method. The subjective listening test, executed in accordance with ITU-T P.800 a ITU-T P.830 recommendations, performs 25 listeners. The software Tester is also used for the subjective speech quality assessment. As well the listeners participated on the subjective testing have been selected from students and employees of the Czech Technical University in Prague.

We have considered following parameters: four groups of phonetic elements and five ratios of packet losses (2, 4, 6, 8, and 10 %). Four speeches are generated for each pair of parameters (group and packet loss ratio) to eliminate effect of random drop of losses. Above mentioned assumptions results into 80 speeches utilized for speech quality testing (4 groups * 5 ratios * 4 speech). The overall duration of the subjective listening test is approximately 20 minutes per a listener.

The results of subjective test are presented in Fig. 8. From this figure can be observed that the most significant group of phonetic elements from the speech quality point of view are groups containing vowels and diphthongs. This fact is caused by two reasons. The first one, based on lexical aspect, says that the vowels are basement of nearly all syllables and words; hence its modification or unintelligibility can cause the change of the meaning of the whole word. The second aspect is the signal processing. From this side, the vowels as well as sound voices contains high amount of energy. Therefore, its loss leads to the loss of major part of information.

The second most important group consists of nasals and liquids since all speech sounds included in this group are voiced and thus they carry high energy. The difference between this group and group with vowels and diphthongs is marginal; it is roughly 0.15 MOS in average in subjective tests.

The next most perceptible impact is caused by fricatives. This group contains voiced as well as unvoiced consonants. Therefore the speech quality in comparison to the first and second group is higher (roughly from 0.4 to 0.55 MOS).

The lowest important group consists of plosives and affricates. This group contains also voiced and unvoiced consonants; however its energy is the lowest of all groups. Its impact on the speech quality is less perceptible by users. The average MOS score is by roughly 0.5 MOS higher than the score of speech with losses in fricatives.

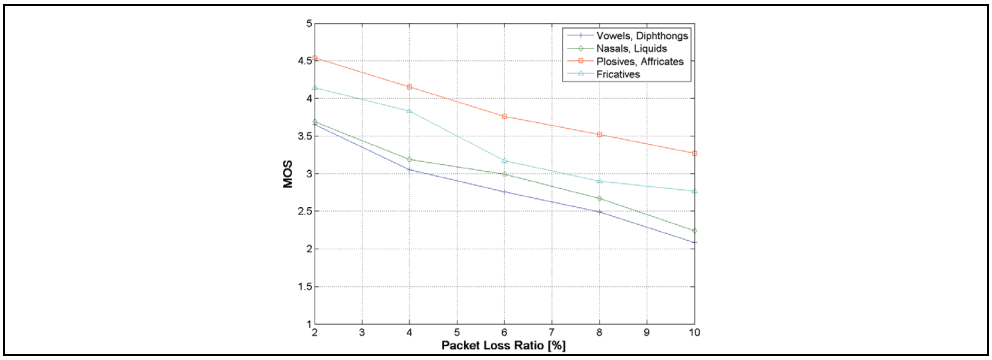

Fig. 8. Subjective assessment of an impact of phonetic elements.

The results of objective speech quality measurement are depicted in Fig. 9. These results show slightly lower speech quality than subjective test. Contrary to the subjective tests, the second group (nasals & liquids) seems marginally more important (but by only 0.18 MOS in average) than the first group (vowels & diphthongs) at higher packet loss. Nevertheless, the impact of both groups is more perceptible than another two groups (fricatives, plosives & affricates) since fricative, plosives and affricates carry less energy in general. The significance of the losses in fricatives is very close to the impact of plosives & affricates. The slightly higher negative impact (less than 0.1 MOS) is achieved by losses in fricatives.

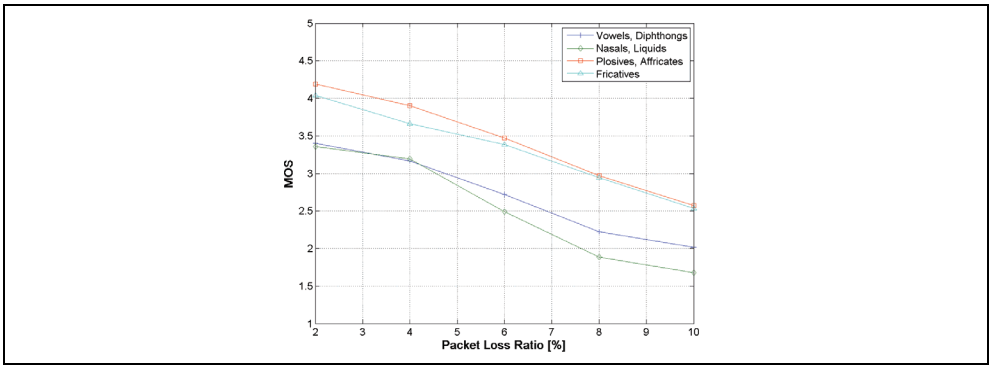

Fig. 9. Objective assessment of an impact of phonetic elements by PESQ method.

The average MOS score of subjective listening test and objective speech quality assessment by PESQ is summarized in Table 6. The average subjective score is always higher than the results of objective evaluation. The difference is negligible for vowels & diphthongs; however it is more significant in case of nasals & liquids and plosives & affricates.

Group	Subjective score (MOS)	Objective PESQ score (MOS)
Vowels& diphthongs	2.81	2.70
Nasal & liquids	2.96	2.52
Plosives & affricates	3.85	3.42
Fricatives	3.36	3.31

Table 6. Average MOS score of each group of phonetic elements.

The difference in speech quality of groups containing only voice sounds and groups containing also unvoiced sound is considerable in results of both subjective as well as objective tests. For example, the speech quality is roughly by 1 MOS higher if the packet losses hit only plosives and affricates than if the losses are in vowels and diphthongs. This fact should influence the design of packet loss concealment mechanisms to put more focus on elimination of losses of vowels, diphthongs, nasals or liquids.

6. Conclusions

This chapter provides an overview on the speech quality assessment in VoIP networks. Several effects that can influence the speech quality are investigated by objective PESQ and/or subjective tests.

The results of objective tests show advantage of wideband communication channel only for high quality networks (with PLR up to 4%). On the other hand, while the speech is affected by consecutive packet losses or by individual losses with higher packet loss ratio, the narrowband channel reaches better score. The most significant difference between wide and narrow band speeches is at 12 % of lost packets.

The consecutive packet losses can leads to the higher speech quality while the duration of losses is long enough comparing to the individual losses. The exact duration of loss that reaches higher score than individual one depends on the length of packets.

The tests of harmonic distortion performed in the means of a suppression of a part of bandwidth, leads to the conclusion that the most important parts of the frequency band are the lowest and the highest bands. The objective method PESQ is not able to handle with the harmonic distortion and its results do not match the subjective one.
The evaluation of the importance of the groups of phonetic elements shows that the most considerable elements are vowels and diphthongs. On the other hand, the speech quality is affected only slightly by losses of plosives or affricates.

7. References

Bachu, R. G.; Kopparthi, S.; Adapa, B. & Barkana B. D. (2010). Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy, In: *Advanced Techniques in Computing Sciences and Software Engineering*, Khaled Elleithy, 279-282, Springer, ISBN 978-90-481-3659-9.

Barriac, V.; Saout, J.-Y. L. & Lockwood, C. (2004). Discussion on unified objective methodologies for the comparison of voice quality of narrowband and wideband scenarios. *Proceedings of Workshop on Wideband Speech Quality in Terminals and Networks: Assessment and Prediction*, June 2004.

Becvar, Z.; Pravda, I. & Vodrazka, J. (2008). Quality Evaluation of Narrowband and Wideband IP Telephony. *Proceeding of Digital Technologies 2008,* pp. 1-4, ISBN 978-80-8070-953-2, November 2008, Žilina, Slovakia.

Benesty, J; Sondhi, M. M. & Huang Y. (2008). *Springer handbook of speech processing*, Springer-Verlag, pp. 308, ISBN: 978-3-540-49125-5, Berlin Heidelberg, Germany.

Brada, M. (2006). Tools Facilitating Realization of Subjective Listening Tests. *Proceedings of Research in Telecommunication Technology 2006,* pp. 414-417, ISBN 80-214-3243-8, September 2006, Brno, Czech Republic.

Clark, A. D. (2002). Modeling the Effects of Burst Packet Loss and Recency on Subjective Voice Quality. *The 3rd IP Telephony Workshop 2002,* New York, 2002.

Ding, L. & Goubran, R. A. (2003). Assessment of Effects of Packet Loss on Speech Quality in VoIP. *Proceedings of The 2nd IEEE International Workshop on Haptic, Audio and Visual Environments and Their Applications, 2003*, pp. 49–54, ISBN 0-7803-8108-4, September 2003.

Fastl, H. & Zwicker, E. (1999). *Psychoacoustics. Facts and Models, Second edition*, Springer, ISBN 3-540-65063-6, Berlin.

Friedlander, B. & Porat, B. (1984). The Modified Yule-Walker Method of ARMA Spectral Estimation, *IEEE Transactions on Aerospace Electronic Systems*, Vol. 20, No. 2, March 1984, pp. 158-173, ISSN 0018-9251.

Hanzl, V. & Pollak, P. (2002). Tool for Czech Pronunciation Generation Combining Fixed Rules with Pronunciation Lexicon and Lexicon Management Tool. *In Proceedings of 3rd International Conferance on Language Resources and Evaluation*, pp. 1264-1269, ISBN 2-9517408-0-8, Las Palmas de Gran Canaria, Spain, May 2002.

Hassan, M. & Alekseevich, D. F. (2006). Variable Packet Size of IP Packets for VoIP Transmission. *Proceedings of the 24th IASTED international conference on Internet and multimedia systems and applications*, pp. 136-141, Innsbruck, Austria, February 2006.

Holub, J.; Beerend, J. G. & Smid, R. (2004). A Dependence between Average Call Duration and Voice Transmission Quality: Measurement and Applications. *In Proceedings of Wireless Telecommunications Symposium*, pp. 75-81, May 2004.

ITU-T Rec. E.800 (1994). Terms and definitions related to quality of service and network performance including dependability. August 1994.

ITU-T Rec. G.107 (2005). The E-model, a computational model for use in transmission planning. March 2005.

ITU-T Rec. G.114 (2003). One-way transmission time. May 2003.

ITU-T Rec. G.711 (1988). Pulse Code Modulation of Voice Frequencies. 1988.

ITU-T Rec. G.711.1 (2008). Wideband embedded extension for ITU-T G.711 pulse code modulation. March 2008.

ITU-T Rec. P.800 (1996). Methods for Subjective Determination of Transmission Quality. August 1996.

ITU-T Rec. P.800.1 (2003). Mean Opinion Score (MOS) terminology. March 2003.

ITU-T Rec. P.830 (1996). Subjective Performance Assessment of Telephone-Band Wideband Digital Codecs . February 1996.

ITU-T Rec. P.862 (2001). Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. February 2001.

ITU-T Rec. P.862.1 (2003). Mapping function for transforming P.862 raw result scores to MOS-LQO. November 2003.

Kondo, K. & Nakagawa, K. (2006). A Speech Packet Loss Concealment Method Using Linear Prediction. *IEICE Transactions on Information and Systems*, Vol. E89-D, No. 2, February 2006, pp. 806-813, ISSN 0916-8532.

Linden, J. (2004). Achieving the Highest Voice Quality for VoIP Solutions, *Proceedings of GSPx The International Embedded Solutions Event*, Santa Clara, September 2004.

Molau, S.; Pitz, M.; Schluter, R. & Ney, H. (2001). Computing Mel-frequency cepstral coefficients on the power spectrum, P*roceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 73-76, ISBN 0-7803-7041-4, Salt Lake City, USA, August 2001.

Oouchi, H.; Takenaga, T.; Sugawara, H. & Masugi M. (2002). Study on Appropriate Voice Data Length of IP Packets for VoIP Network Adjustment. *Proceedings of IEEE Global Telecommunications Conference*, pp. 1618-1622, ISBN 0-7803-7632-3, November 2002.

Robinson, D. J. M. & Hawksford, M. O. J. (2000). Psychoacoustic models and non-linear human hearing, In: *Audio Engineering Society Convention 109*, September 2000.

Sing, J. H. & Chang, J. H. (2009). Efficient Implementation of Voiced/Unvoiced Sounds Classification Based on GMM for SMV Codec. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, Vol. E92–A, No.8, August 2009, pp. 2120-2123, ISSN 1745-1337.

Sun, L. F.; Wade, G.; Lines, B. M. & Ifeachor, E. C. (2001). Impact of Packet Loss Location on Perceived Speech Quality. *Proceedings of 2nd IP-Telephony Workshop*, pp. 114-122, Columbia University, New York, April 2001.

Tosun, L. & Kabal, P. (2005). Dynamically Adding Redundancy for Improved Error Concealment in Packet Voice Coding. *In Proceedings of European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey, September 2005.

Ulseth, T. & Stafsnes, F. (2006). VoIP speech quality – Better than PSTN?. *Telektronikk*, Vol. 1, pp. 119-129, ISSN 0085-7130.