Impact of Additional Noise on Subjective and Objective Quality Assessement in VoIP

Zdenek Becvar, Lukas Novak, Jan Zelenka, Miloslav Brada, Pavel Slepicka
Department of Telecommunication Engineering
Czech Technical University
Prague, Czech Republic
{becvaz1, novakl5, zelenj2, bradam2, slepicp}@fel.cvut.cz

Abstract—The main requirement in the Voice over IP technology is a good quality of received voice signal during communication between subscribers. The signal quality can be influenced by many factors such as packet loss, jitter, packet delay, noise etc. and it can be measured by number of methods. The main purpose of this paper is the investigation of an impact of different noise types and different noise levels on the quality assessment in VoIP. The artificial generated noises and real noises obtained from real telecommunications networks were used for testing. The next goal is a comparison of the results obtained by subjective listening tests and objective measuring methods. PESQ and 3SQM were used for objective testing in this paper.

Keywords—Speech quality; Noise; VoIP; MOS

Topic area—Quality of service in multimedia communication.

I. INTRODUCTION

The quality of the audio signal, which is perceived by a human hearing system, directly depends on several characteristics of the signal that is reconstructed on a receiving side of a telecommunication chain. This analog signal can be affected by many kinds of disturbances that are produced at an output of the voice decoder. These disturbances such as sound dropouts, frequency disruptions, phase discontinuity, noises etc. are products of not appropriate signal processing of reconstruction methods. These methods can not recover the correct waveform of the signal. That is caused by effects like packet loss, jitter and so on [1] [2].

Especially, noise is a special kind of disturbance which has not always the disturbing effect, but in some cases it can mask some undesirable effects like dropouts are. We can distinguish many kinds of noises and study their influences on a human perception. In this paper will be examined three types of noises with different SNR (Signal to Noise Ratio) levels. The influence of noises will be inspected with subjective and objective testing methods. As the objective methods will be used PESQ (Perceptual Evaluation of Speech Quality) according to ITU-T P.862.1 [3] recommendation and the method noted as 3SQM (Speech Quality Measurement) according to ITU-T P.563 [4]. The results conformity was evaluated by using Pearson's correlation coefficients [5] and residual errors [5].

The rest of paper is organized subsequently. Next section describes characteristics of tested noises and the subjective tests composition. The third section provides the results obtained by subjective test and two objective methods and their discussion. Last section presents our conclusions and future work plans.

II. TESTS COMPOSITION

In the tests were used two artificial generated noises: the white noise and the grey noise. The third kind of noise signal was represented by two noise signals obtained from the receiving side of the telecommunication chain of local telecommunication operator. There were used 4 SNR levels: 0dB, 10dB, 20dB, 30dB. Each combination of SNR and type of noise was evaluated five times to increase the accuracy of final quality value.

A. *Types of Used Noises*

A first type of artificial noise was white noise with normal Gaussian distribution density of probability. White noise has flat spectrum distribution of energy (Fig. 1). This noise is inherently generated by all metallic lines due to its nonzero resistance and therefore some thermal noise there will be always present. Another source of white noise is a quantization error caused during analog to digital conversion of electrical signals due to the finite number of used quantization levels. Moreover, there are a lot of other sources in telecommunication networks which generate white noise; this was the reason of choice of this type of noise.

Grey noise was chosen as a second type of artificial noise. Grey noise has spectrum characteristic (Fig. 1) which corresponds to the psychoacoustic equal loudness curve of human ear [6]. A sensitivity of human ear is dependant on frequency; in the band of telecommunication channel is in the interval of 300–2000Hz rising with the local maximum located around 2000Hz and in the band of 2000–3400Hz is slightly decreasing. Therefore, grey noise has such spectrum distribution that an auditor perceives all tones with same loudness.

This work has been supported by internal grant No. CTU0712413 funded by CTU in Prague and by FRVS grant No. 1439/2007/G1.

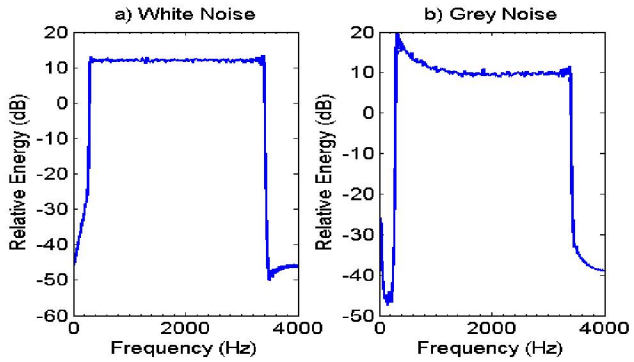

Figure 1. Spectrum characteristics of used white noise and grey noise.

Both types of noise were filtered with the low pass filter to be adapted into the band of telecommunication channel 300 – 3400Hz.

The third and the fourth type of noise were real noises recorded in two different points in telecommunication network. These noises are consisted of all noise contributors within the telecommunication chain, but moreover, they are also affected by transfer characteristics of signal processing electronic circuits, e.g. SLIC (Subscriber Line Interface Circuit). Power spectral density of these noises can be seen in Fig.2.

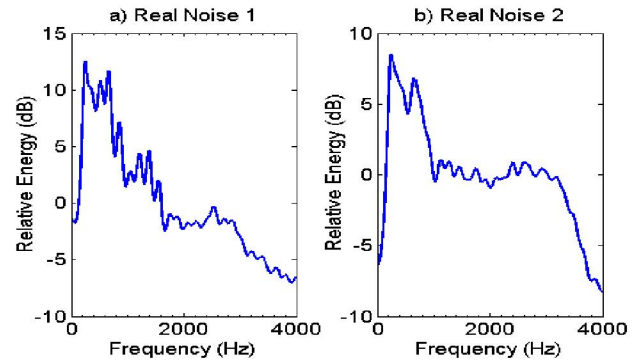

Figure 2. Spectrum characteristics of used real noise 1 and real noise 2

B. *Noises generation*

Modification of all speech files was implemented in MATLAB® v7.1 software. For generation of noise signals was used LPC (Linear Predictive Coding) [7]. A degree of LPC model was set to 200 in the case of artificial noise signals and to 50 in the case of real noise signals. LPC was excited by white noise with gauss distribution. There were chosen 4 levels of SNR (1) and for each level there was generated noise accordingly.

$$SNR = 10\log\left(\frac{P_s}{P_n}\right) \tag{1}$$

Therefore, noise generated by the LPC model x_n was scaled by k and added to original speech signal x_s to form the distorted signal x_d (2).

$$x_d[n] = x_s[n] + k \cdot x_n[n] \tag{2}$$

The scalar k was calculated (3) to reach the desired SNR of the signal x_d.

$$k = \sqrt{\frac{P_s}{P_n} 10^{-\frac{SNR}{10}}} \tag{3}$$

C. *Test*

The test is in accordance with recommendations ITU-T P.800 [8] and P.830 [9]. These standards cover a set of requirements that are necessary to fulfill to obtain reasonable subjective test results. The sound samples properties, number of testing subjects, conditions in the testing area, acceptable length of the test and methods to evaluate results are parts of the recommendations.

The original recordings were acquired from the radio Akropolis. They are the studio quality sound samples recorded by professional speakers. The basic set of utterances consists of 80 emotively neutral sentences spoken by 5 male and 5 female speakers.

The average utterance length was approximately 8 seconds with defined voice to silence ratio. The speech fulfill between 40% and 80% of the time of utterance, the rest of time the silence is present. In the next processing the samples were downsampled to the 8 kHz, converted to 16-bit linear PCM (Pulse Code Modulation) and disturbed by noise.

The test was performed by 18 listeners, the employees and students of university. This is more than the acceptable minimum of listeners defined in [8].

The test was performed with high quality headphones in the quiet laboratory at Czech Technical University. The room was guarded against ambient noise and disturbances. Also the conditions in the room were selected not to disturb the listeners.

The test consists of 80 utterances and its realization takes about 30 minutes.

The results were calculated as an average from each 5 utterances distorted by the same type of noise and SNR. The same evaluation method was used in processing the 3SQM and PESQ results.

To express the conformity between the subjective and objective methods we used the Pearson's coefficients defined in (4) and residual errors (5).

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \quad (4)$$

$$e_i = x_i + y_i \quad (5)$$

III. Experimental Results and Discussion

The results obtained by objective methods (PESQ and 3SQM) and by subjective testing (in figures and tables noted as SUBJ) for four SNR levels and for different types of noises (white, grey and two real noises) are shown in Tab. 1.

TABLE I. Results of subjective test and objective methods

SNR [dB]	White noise			Grey noise		
	SUBJ	3SQM	PESQ	SUBJ	3SQM	PESQ
0	1,60	1,00	1,41	1,27	1,00	1,31
10	2,68	2,58	1,72	2,68	2,11	1,82
20	3,50	3,30	2,58	3,48	3,40	2,76
30	4,35	3,34	3,66	4,45	3,60	3,61
SNR [dB]	**Real noise 1**			**Real noise 2**		
	SUBJ	**3SQM**	**PESQ**	**SUBJ**	**3SQM**	**PESQ**
0	1,42	1,00	1,52	1,22	1,00	1,30
10	2,58	2,39	2,03	2,72	2,06	1,78
20	3,62	3,27	2,93	3,48	3,09	2,72
30	4,68	2,81	4,11	4,40	3,40	3,78

The results of subjective test are plotted for all types of noise in Fig. 3. As TABLE I 1 and Fig. 3 show, the behavior of subjective MOS score in the function of SNR is very close to the linear behavior. Subjective listening quality is nearly independent on the type of used noise. The maximal difference among all noise types was 0.38 MOS (white noise and real noise 2 for SNR=0 dB).

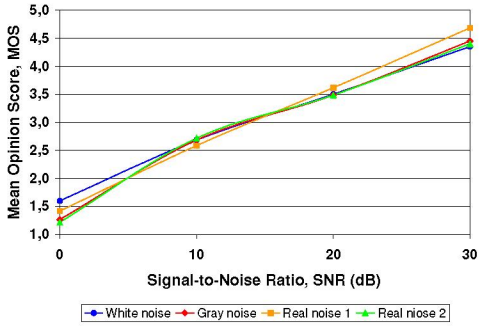

Figure 3. Subjective results for four different noises.

Fig. 4 shows the comparison of subjective and objective results for white noise and grey noise. Results obtained by PESQ method were very close to the subjective results only for SNR=0 dB. Afterwards, the difference was around 0.8 MOS for all SNRs. The 3SQM results were closer to the subjective (difference less than 0.5 MOS) than PESQ results up to SRN=20 dB.

The results obtained by PESQ for both noises were similar to each other (difference between them was about 0.1 MOS). Opposite to the PESQ, 3SQM results for white and grey noise was slightly closer to the subjective ones.

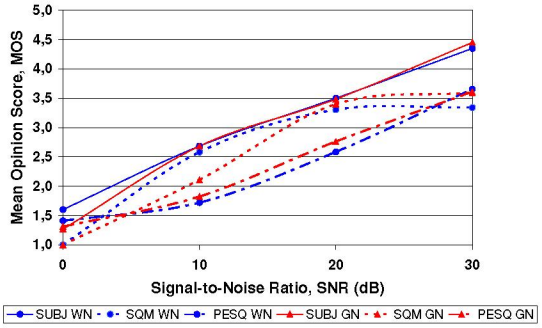

Figure 4 Comparison of subjective an objetive (PESQ and 3SQM) results for white noise (WN) and grey noise (GN).

The subjective and objective results for two types of real noises are shown in Fig. 5. As it can be seen from Fig. 4 and Fig. 5, the behaviors of objective methods are analogical to the white and grey noises. Nearly all (excepting two PESQ values for SNR=0 dB) objective MOS scores was less then subjective.

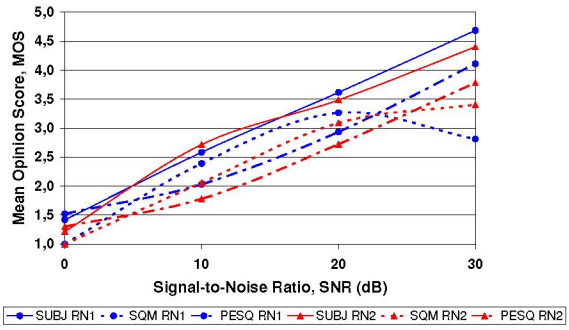

Figure 5. Comparison of subjective an objetive (PESQ and 3SQM) results for two different real noises.

The behavior of PEQS test results have linear behavior for SNR greater then 10dB, but there is nearly no difference between SNR=0 dB and SNR=10 dB. The situation is opposite than in the case of 3SQM. The 3SQM's results are almost linear for the SNR up to 20 dB, but for the greater value, the MOS is invariable or slightly decreasing.

The correlation among subjective test and two objective methods is presented in TABLE II. The Pearson's correlation

coefficient was used for conformity detection. The results were generally very good and show good correspondence among subjective and objective results; only one coefficient (for 3SQM method and real noise 1) was less than 0.9. Nevertheless, wide differences between results are noticeable in Fig. 4 and Fig. 5. It is caused by not taking the direct component into account. This difference is expressed by residual error.

TABLE II.COMPARISON OF RESULTS OBTAINED BY SUBJECTIVE TEST AND OBJECTIVE METHODS BY PEARSON'S CORELLATION COEFFICIENTS

Pearson coefficient	White noise	Grey noise	Real noise 1	Real noise 2
3SQM	0,94	0,97	0,84	0,98
PESQ	0,96	0,97	0,98	0,95

The average residual errors are summarized in the TABLE III. The residual errors show wide difference in the objective and subjective results. 3SQM shows slightly better conformity with subjective test from the residual errors point of view. Residual error was smaller for the case of white, grey and real noise 2 in results obtained by 3SQM and subjective test. Opposite to this, the best similarity for PESQ and subjective test was achieved for real noise 1. But it is necessary to consider that all average residual errors take relatively high value.

TABLE III. AVERAGE RESIDUAL ERRORS BETWEEN SUBJECTIVE TEST AND OBJECTIVE METHODS

Average residual error	White noise	Grey noise	Real noise 1	Real noise 2
3SQM	0,48	0,45	0,71	0,57
PESQ	0,69	0,62	0,48	0,60

IV. CONCLUSIONS AND FUTURE WORK

The subjective test shows that the most important factor influencing the speech quality (in the term of noise disturbance) is the SNR. Noise with SNR 30 dB has nearly no impact on the MOS score, SNR 0 dB leads to the MOS between 1 and 1.5. The relation between SNR and MOS score is almost linear.

Further, nearly no dependence between noise type and speech quality results from the subjective test. The different noise characteristics make less than 0.38 MOS difference. The variance in the perceived quality comparing miscellaneous noises is not fundamental, even the spectral characteristics of the noises differ.

The SNR is factor having major effect on the quality of speech signal. The influence of noise type is not so important to be considered in the area of Voice over IP quality research. This fact can be useful for Comfort Noise Generator (CNG) designing.

The correlation between subjective and objective test, measured by the Pearson's correlation coefficients, shows very good tests conformity. The Pearson's correlation coefficients was better than 0.9 (except one case). However, the Pearson's correlation coefficients compare the similarity of the curves shape, but do not take the direct components into account. The direct component is evaluated by residual errors. The residual errors show considerable difference (from 0,45 to 0,71) between subjective and both objective tests. Generally, the objective methods tend to undervalue the MOS score.

In the future, we will investigate the impact of the other kinds of disturbances such as phase discontinuity, losses of individual phonemes and harmonic and disharmonic distortions.

REFERENCES

[1] Z.Becvar, J. Zelenka, M. Brada, T. Valenta, "Comparison of Subjective and Objective Speech Quality Testing Methodsin the VoIP Networks," In proceeding of IWSSIP06. Budapest, 2006.

[2] Z. Becvar, J. Zelenka, M. Brada, L. Novak, "Comparison of PLC methods used in VoIP networks," unpublished.

[3] ITU-T Recommendation P.862.1, "*Mapping function for transforming P.862 raw result scores to MOS-LQO* ," 2003.

[4] ITU-T Recommendation P.563, "*Single Ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications*," 2004.

[5] ITU-T Recommendation P.862, "*Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-toend Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs*," Genava, 2001.

[6] E. Zwicker, H. Fastl, *Psychoacoustics – Facts and Models*, 2nd ed., Berlin, 1990.

[7] L. R. Rabiner, R.W. Schafer, *Digital processing of speech signals.* Prentice Hall, New Jersey, 1978.

[8] ITU-T Recommendation P.800, "*Methods for Subjective Determination of Transmission Quality*," 1996.

[9] ITU-T Recommendation P.830, "*Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs*," 1996.