Computer Standards & Interfaces 36 (2014) 626–630

Contents lists available at ScienceDirect

Computer Standards & Interfaces

journal homepage: www.elsevier.com/locate/csi

Evaluation of objective speech transmission quality measurements in packet-based networks

Oldřich Slavata *, Jan Holub

Dept. of Measurement, Faculty of Electrical Engineering, Czech Technical University in Prague, Technicka 2, CZ-166 27 Praha 6, Czech Republic

ARTICLE INFO

Article history:
Received 7 October 2011
Received in revised form 9 September 2013
Accepted 10 September 2013
Available online 17 September 2013

Keywords:
POLQA
speech quality
MOS
packet loss
jitter

ABSTRACT

This paper presents an analysis of the relation between IP channel characteristics and final voice transmission quality. The NISTNet emulator is used for adjusting the IP channel network. The transmission quality criterion is an MOS parameter investigated using the ITU-T P.862 PESQ, future P.863 POLQA and P.563 3SQM algorithms. Jitter and packet loss influence are investigated for the PCM codec and the Speex codec.

1. Introduction

At the beginning of the 21st century, increasing transmission capacity of the network and improved digital processing methods for video and acoustic signals enabled the Internet to be used for real time voice and video communication. VoIP (voice-over Internet Protocol) allows the transmission of voice in digital form in UDP/TCP/IP packets. Using the IP network to transfer a telephone call poses particular difficulties. Network parameters such as delay variations (jitter), packet loss and bandwidth affect the quality and clarity of the transferred audio signal. Other parameters do not affect the transmitted speech waveform directly but contribute to a decrease in the conversational quality score (e.g. delay).

To assess the quality of voice transmission we used the MOS (mean opinion score) scale (Table 1). The term MOS is defined in Recommendation ITU-T P.800 [15].

Several methods can be used to obtain MOS values. The most accurate method is a subjective test, where the MOS value is obtained directly from users. However, conducting subjective tests is time-consuming and expensive. It is therefore replaced by objective methods based on computer algorithms.

Intrusive methods provide results nearest to those provided by subjective tests. They are based on a comparison of the original and transferred sample. These algorithms use psychoacoustic models of human perception, seeking to offer a mathematical description of the human perception of sound, and to find variables which have a direct impact on the perceived quality of a voice signal. Intrusive methods include PAMS (Perceptual Analysis Measurement System), developed by British Telecommunications, PSQM (Perceptual Speech Quality Measurement), described in Recommendation ITU-T P.861, PESQ (Perceptual Speech Quality Evaluation of), according to ITU-T P.862 (P.862.1) and newly ITU-T P.863 — POLQA (Perceptual Objective Listening Quality Analysis) [5].

Non-intrusive methods are another type of quality measurement. These methods do not use the reference signal, and the final MOS is calculated using the parameters of the transferred sample only. A disadvantage of these methods is their lower accuracy and reliability. An example of a non-intrusive method is 3SQM, which is defined in recommendation ITU-T P.563.

2. Methods used to obtain MOS values

2.1. ITU-T P.862 — PESQ

PESQ is intrusive method of measuring speech transmission quality. It works on the principle of comparing the original and transferred sample.

Before the comparison, the amplitude equalization and time alignment of both samples must be done. Amplitude compensation only adjusts the volume to the level needed for further processing. It does not correct any errors caused by too high or low volume when recording the sample. For the final result of the PESQ algorithm is very important to have matched the corresponding sections of the signal. Therefore, it is important to align any delays of the degraded signal against the original. This part of the algorithm operates on the basis of correlation between the original and degraded signal. The algorithm first calculates the

* Corresponding author.
E-mail address: slavao1@fel.cvut.cz (O. Slavata).

0920-5489/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csi.2013.09.003

Table 1
MOS scale.

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

delay of the entire sample. Then the sample is divided to its sub-parts and the correlation is calculate for each part separately. Consequently, the sample is divided into shorter periods again and the various delays are recalculated until a segment is too short or the correlation is not better than in the previous step.

The most important part of the PESQ algorithm is psycho-acoustic transformation. Parameters of the original and degraded signal are evaluated using a mathematical model of the human auditory system.

- The sample is divided into Sections 16 ms long with a 50% overlap.
- For each segment 256-point FFT is calculated.
- Series of FFT results are divided into 17 frequency bands called "bark bands".
- For each of the seventeen bands the energy contained therein is summed.
- The energy is converted back to the volume level.
- Results are further threshold and weighted according to the sensitivity of the human ear to different frequencies.

The result of the transformation is a vector with 17 values for each 16 ms period. These vectors are then sorted into a matrix according to the time sequence in the signal. Matrixes of original and degraded signals are compared. Positive and negative differences are summed separately because the human ear is more sensitive to the added disturbance than the missing signal. The weighted sums of the differences are then subtracted from the maximum value of five and the resulting value is MOS for a given sample.

2.2. ITU-T P.863 — POLQA

POLQA is the successor of PESQ. Principle of the algorithm is similar to PESQ but it removes some of its disadvantages. Time alignment algorithm of POLQA can recognize new features of modern codecs such as "time warping" which PESQ evaluates as errors.

Similarly to PESQ POLQA supports measurements in the common telephony band (300–3400 Hz), but in addition it has a second operational mode for assessing HD-Voice in wideband and super-wideband speech signals (50–14000 Hz).

POLQA also examines the original signal and its possible errors (too much timbre, noise or reverberation) are taken into account in the final evaluation. This approximates the results of subjective tests where users compare the transmitted signal, with their subjective vision of the ideal.

2.3. ITU-T P.563 — 3SQM

3SQM is non-intrusive method for measuring listening quality of the voice signal. The algorithm consists of three separate parts which have different methods of calculating the MOS.

Part 1 In the sample are calculated parameters typical for computer signal processing such as: signal-to-noise ratio (SNR), the length of suspension and damping, time cropping … Range of values of these parameters are then used to estimate the value of MOS.

Part 2 A complex "cleaning" function is applied at the degraded sample. Missing parts are recalculated; the sample is filtered and further regulated. This purified sample, together with the original, is used as input signal for the simplified PESQ (without time alignment) and its output is an estimate of MOS.

Part 3 The main part of this block is a precision LPC model of the human vocal tract. This 'synthesizer' attempts to pronounce the degraded sample. The result is compared with the original sample. Everything different in the original sample is considered as unnatural to the human vocal tract and considered as damage caused during sample transfer. The sum of this added disturbance is used to calculate the MOS estimation.

The most distant of these three estimates of MOS is dropped and the arithmetic mean of the remaining two is the resulting estimate of MOS for the entire algorithm.

3. Experiment description

3.1. Test-bed

The test-bed (Fig. 1) consisted of three computers, an Opera audio analyser, and interconnecting cables. A concatenated speech file in WAV format (8kSa/S, 16bit), 16.75 s in length, was used. The file contained 4 short sentences spoken by 4 different speakers (two men, two women), and adequately covered the entire human speech spectra. Due to this fact, the concatenated file was used as an effective replacement for testing using multiple speech samples.

The signal was transferred from an audio output "line 1 out" of the Opera analyzer to an audio input (microphone) of PC 1. PC 1 and PC 2 were connected by a UTP network cable (subnet 192.168.0. X), as were PC 2 and PC 3 (subnet 192.168.1. X). PC 2 was therefore fitted with a two-port network interface card. The test signal was transferred from PC 1 to PC 3 using a VoIP call in the Linphone program, using PCM (G.711) and Speex codecs. From PC 3, the audio output (headphone) signal was led back into the audio input "line 2 in. of the Opera analyzer. The NISTNet emulator [16] was running on PC2, which (according to the specific settings) introduced transmission errors between PC 1 and PC 3. The results depend on the accuracy and repeatability of the network simulation. We proved by several experiments [8,9] that NISTNet suits these requirements satisfactorily. It was also used in other experiments [12]. The measured samples were adjusted in Adobe Audition 3.0 (converting stereo → mono) and then tested using the POLQA (ITU-T P.863) PESQ (ITU-T P.862) and 3SQM (ITU-T P.563) algorithms [7]. The PESQ algorithm output was recalculated to the value of MOS-LQO (Listening Quality Objective) according to a mathematical prescription defined in ITU-T P.862.1. According to the official wording of P.862,

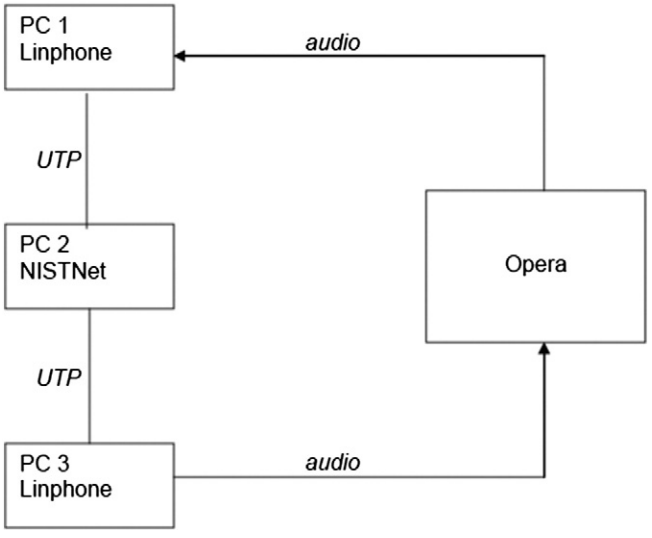

Fig. 1. Test-bed.