New approach in subjective and objective speech transmission quality measurement in TCP/IP networks

New approach in subjective and objective speech transmission quality measurement in TCP/IP networks

Pavel Souček, Oldřich Slavata, Jan Holub

Dept. of Measurement, FEE CTU Prague, Technicka 2, CZ 166 27 Prague 6, Czech Republic

E-mail: soucepa3@fel.cvut.cz, slavao@fel.cvut.cz, holubjan@fel.cvut.cz

Abstract. This paper deals with problems of speech transmission quality measurement in modern telecommunication networks. It focuses on problems caused by specific types of distortions and errors caused present in transmissions using TCP/IP networks.

1. Introduction

Measurement methods used in telecommunications have evolved rapidly in last few years due to communication technologies developments, caused by the need of better utilization of existing transmission channel capacities or the need of minimization of new communication channels establishment, respectively. At the same time, the amount of transmitted data increases significantly. This trend requires many partial problems to be solved. Very actual topic is objective (algorithmic) measurement of transmitted speech, video or multimedia quality in (common) case of transmission channel errors including packet loss and jitter delay for packet-based transmissions. Moreover, the used algorithms must be continuously calibrated for given conditions like language and user portfolio as subjective quality assessment varies in time based on developing user experience (and thus expectation). Considering the fact that subjective service acceptability threshold approaches in some cases the offered service quality, even tiny details, frequently neglected in the past, must be taken into account, e.g. nativity or non-nativity of the language used for the communication. It was shown that non-native listeners have higher demands for telecommunication channel quality that can have significant (roaming in mobile networks) or even fatal (radio communication during multi-national military operations) impacts to the service users.

The main problem of present approaches in speech transmission quality measurement is that they are partially based on facts that are in some ways obsolete. Most problematic is their basis in analog telephony and they have difficulty in reflecting modern trends which lead to new type of distortions in signal and even more dramatic change in listeners perception.

Like technical part of communication, the listener's perception evolved over the years. This evolution is consequent of widespread availability of free communication tools like ICQ, Skype and many others, which are able to employ modern coders much more quickly than common telephone network or mobile operators. This is even more important with use of free communication software like Viber and Linphone, which are using cell phones data connection to make call instead of standard service. This allows much easier employment of modern coders and leads to shift in perceived speech transmission quality. One of the consequences of this widespread usage of such coders, which often

deploy psycho-acoustic compression, is that respondents in research made in [1] often evaluated mp3 like higher quality record over lossless formats.

2. Conventional speech transmission quality measurement

Subjective speech transmission quality measurement is standardized in ITU Recommendation P.800 [2]. This recommendation defines listening conditions including listening chambers, its reverberation and length of testing session. It also defines the parameters of speech samples and scale used for evaluation of speech transmission quality. This scale is named Mean Opinion Score (MOS) and it is five point scale where 1 is the worst and 5 the best.

Subjective tests are divided into the conversational test and the listening test. Listening tests consist of a session with a group of subjects from whom an opinion on the speech transmission quality of the samples is gathered. Conversational tests are similar, but they are more demanding on time and organization than listening tests.

According to this recommendation the speech material should consist of simple, meaningful, short sentences, chosen at random as being easy to understand (from current non-technical literature or newspapers, for example). Every sample then should contain from two to five sentences, each from two to three seconds in length.

The specification of suitable sample is further narrowed by ITU-T Recommendation P.862.3[3]. For use with PESQ algorithm samples from eight to twelve seconds should be used.

Another part of P.800 defines number of votes needed to maintain high precision and to obtain statistically significant results. In case of listening tests it makes at least 100 votes per condition. It is usually difficult and impractical to have 100 respondents in one test, so in order to achieve this precision we usually use several samples for same condition.

The second way to obtain MOS values is by objective testing, which can be divided into two categories - intrusive and non-intrusive. These methods show good performance under most conditions, but it is also known that they fail under certain special circumstances.

Intrusive methods usually deliver results nearest to the results of subjective tests. They are based on a comparison between the original sample and the transferred sample. These tests are based on algorithms that use psychoacoustic models of human perception. Intrusive methods include several standardized algorithms, e.g. PESQ (Perceptual Evaluation of Speech Quality- ITU-T P.862 [4]) and POLQA (Perceptual Objective Listening Quality Assessment ITU-T P.863 [5]).

Non-intrusive methods are another type of objective measurement. These methods do not need to use a reference signal, and the final MOS is calculated only from the parameters of the transferred sample. A disadvantage of these methods is that they are less accurate and less reliable than intrusive methods. An example of a non-intrusive algorithm is 3SQM (Single Side Speech Quality Measurement - ITU-T P.563 [6]).

3. Specific problems with TCP/IP networks

Nowadays more and more of the calls is transferred using data and TCP/IP networks. Compared to original analog telephone network, these networks are relatively reliable and in normal conditions errors are rare. For example in case of packet loss the error rate should not exceed few percent.

If we take into consideration, that for most coders one TCP/IP packet carries from one to three voice packets, each containing 20 milliseconds of speech signal, we get 400 TCP/IP packets for whole sample. Error rate under 2 percent means that in worst case we will lose 8 packets. This together makes 480 milliseconds of speech signal, but not necessarily continuous.

With error rate this low, it is not difficult to imagine that lost packets will contain silent breaks between words or even a word that will not change the meaning of message. If only objective tests are used to evaluate speech transmission quality, the resulting MOS score will be below the actual quality, compared to subjective tests.

Another effect caused by use of these samples is, that the type A uncertainty will be very high and even though that objective tests should always give exactly same result. The way to eliminate this

effect, used in current research, is to transfer every sample multiple times and as output speech transmission quality average of these acquired values is used.
In [7] researchers transfer every sample ten times and for every condition they use same speech material. However this approach cannot be used to gain subjective data for comparison, because subjects in such test will be affected by repetitiveness of the speech material and wouldn't be able to evaluate quality properly, if we played same sample to them multiple times.

4. New methodology using long samples

Based on our previous research we found out that samples long approx. 80 - 120 seconds would allow us to decrease time needed for objective tests and thus would lead to significant improvement. These samples would also make it possible to carry out subjective tests for the same set of samples and to compare the directly instead of comparing only similar samples.
In comparison with concatenated sample we can run only one test session and it would reflect real life conditions better. In this case sample is long enough for network related errors to affect quality in a way, listeners will be able to register it and distortion will be probably present in every sample.
The length of samples recommended by P.800 has its meaning. Samples should be short enough, so tested subjects can reliably evaluate their speech transmission quality. When we use longer samples, there is always a risk, that results will be influenced by so called recency effect. This means, that listeners are able to asses quality only of the end of sample, it leaves biggest impression. In order to eliminate this effect we are trying two approaches in order to gain reliable results:

- Continuous evaluation based on beep marks – samples are enhanced with beep marks placed equidistantly in sample every 10 seconds
- Continuous evaluation with custom measurement application – evaluation is realized with help of custom application and the time of change in score is fully in hands of subject

The second method is inspired by [8], where similar algorithm is used for video quality assessment. Application records speech transmission quality set by listener every second and computes average quality. There might be some improvements in the future, depending on gathered data, like weighting the values according to their position in sample. It is reasonable to assume that the beginning and the end of sample should have another importance than the rest of sample.
In order to evaluate, that these two new methods don't suffer same weakness as method currently used for speech transmission quality evaluation used with samples defined in P.800, we also plan to use reference test where original methodology is used, in order to confirm the recency effect with samples this long.
At this time we are carrying out initial tests in order to evaluate correctness of this methodology using speech material from four speakers (two males and two females). This speech material was processed using network simulator to emulate distortions.

5. Conclusion

Methodology currently used for speech transmission quality measurement is not entirely appropriate for use with conditions that include TCP/IP networks or any similar transfer method using packetization and buffers. Moreover speech material used for objective tests cannot be used for subjective testing in. Our new proposed methodology should eliminate this disadvantage and enable subjective tests to be carried out with same speech material as objective tests and without risk of results being affected by recency effect.

6. References

[1] Sterne, Jonathan,"MP3: The meaning of format", Duke University Press, 2012
[2] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," International Telecommunication Union, Geneva, 2001
[3] ITU-T Rec. P.862.3, "Application guide for objective quality measurement based on Recommendations P.862, P.862.1 and P.862.2", International Telecommunication Union,

Geneva, 2007
[4] ITU-T Rec. P.862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”, International Telecommunication Union, Geneva, 2001
[5] ITU-T Rec. P.863, “Perceptual Objective Listening Quality Assessment (POLQA)”, International Telecommunication Union, Geneva, 2011
[6] ITU-T Rec. P.563, “Single-ended method for objective speech quality assessment in narrow-band telephony applications”, International Telecommunication Union, Geneva, 2004
[7] Holub, J. - Slavata, O.: Impact of IP Channel Parameters on the Final Quality of the Transferred Voice In: Wireless Telecommunications Symposium 2012 Papers and Presentation [CD-ROM]. Pomona (CA): California State Polytechnic University, 2012
[8] ITU-T Rec. P.910, ”Subjective video quality assessment methods for multimedia applications”, International Telecommunication Union, Geneva, 2008