Signal Envelop Criterion for Passive Voice Quality Analyzing

Angel Garabitov, Aleksandar Tsenov
Technical University of Sofia
8 Kl. Ohridski Blvd, Sofia 1000
Bulgaria
{angelsg@mail.bg} {aleksandar.tsenov@fdiba.tu-sofia.bg}

ABSTRACT: *In this work, we have studied the VoIP quality signal waveform analysis. We have presented models for objective prediction of the voice quality for IP networks and also monitor the quality control of VoIP networks. We studied the methods to find the quality of audio. The study has analysed the quality of VoIP connection and indicate during quality decrease. This exercise provided the chance of solving the issues in VoIP network before users are affected by VoIP specific connection problems (echo, noise or breaks in the conversation). The signal waveform envelope distortion is reviewed; practical questions of its numerical implementation are discussed. We supported the study with many illustrations.*

Keywords: VoIP Quality, Signal Waveform Analysis

Received: 5 September 2020, Revised 29 November 2020, Accepted 7 December 2020

DOI: 10.6025/jcl/2021/12/1/9-16

1. Introduction

For low speed WAN links that are not well-provisioned to serve voice traffic, problems such as delay, jitter, and loss become even more pronounced. In this particular network environment, the following factors can contribute to poor voice quality:

- Large data packets sent before voice packets introduce long delays.
- Variable-length data packets sent before voice packets make delays unpredictable, resulting in jitter.
- Narrow bandwidth makes the 40-byte combined RTP, UDP, and IP header of a 20-byte VoIP packet especially wasteful.
- Narrow bandwidth causes severe delay and loss because the link frequently is congested.
- Many popular QoS techniques that serve data traffic very well, such as WFQ and RED, are ineffective for voice applications:

Unlike the elastic data traffic that adapts to available bandwidth, voice quality becomes unacceptable after too many drops and too much delay. Perfect sound quality (QoS) in telecommunications systems depends on absence or insignificant influence of

impairments affecting encoding, transmission, and amplification. To implement QoS on a network requires the configuration of QoS features that provide better and more predictable network service by supporting bandwidth allocation, improving loss characteristics, avoiding and managing network congestion, metering network traffic, or setting traffic flow priorities across the network. There are many solutions for QoS assessing. At first stage the software detects impairments and at the second stage uses proprietary algorithms to convert them into MOS score prediction according to ITU-T P.800 standard.

Recently, objective speech quality assessment has become a very active research area. This is an attempt to circumvent the limitations of subjective testing by simulating the opinions of human testers algorithmically. There are two distinct approaches to objective testing: intrusive and non-intrusive.

Intrusive speech quality estimation techniques compare the test (i.e., network distorted) speech signal, as reconstructed by the decoder, to the reference, input speech, basing their estimation on the measured amount of distortion. ITU-T.

On the other hand, non-intrusive schemes assess the quality of the distorted signal in the absence of the reference signal. This approach is effective in environments where the reference speech signal is not accessible. P.563 is the new ITU-T Recommendation for non-intrusive evaluation speech quality in narrowband telephony applications [3]. Intrusive models are more reliable than the nonintrusive ones as the former have access to a reference speech signal to compare the distorted speech signal with.

However, the afore-mentioned models are compute intensive as they base their results on the time and/or frequency domain analysis of the speech signal under test. They also require the test call to be recorded for a considerable duration before it can be analyzed. Hence, they are not suitable for real-time and continuous monitoring of speech quality.

2. Waveform Envelope Distortion Criterion

The delayed packet may come late or may not come at all, in case it is lost. QoS (Quality of Service) considerations for voice are relatively tolerant towards packet loss, as compared to text. Besides, voice smoothing mechanism regulates it so that you don't feel the bump. When a packet is delayed, you will hear the voice later than you should. If the delay is not big and is constant, your conversation can be acceptable. Unfortunately, the delay is not always constant, and varies depending on some technical factors. This variation in delay is called jitter, which causes damage to voice quality. Damages in quality sound reflects on the sound signals and can be seen in signal waveform envelop.

The paper is discussing problems connected with tools to non-intrusively evaluate VoIP quality by waveform analysis. The method detects impairments of quality of audio for human perception. It enables to see the quality of VoIP connection at a glance and warns when quality deteriorates. This gives the option to troubleshoot your VoIP network even before users are affected by VoIP specific connection problems (echo, noise or breaks in the conversation).

The mechanics behind human voice production are unique and in many ways quantifiable. Understanding human speech and its

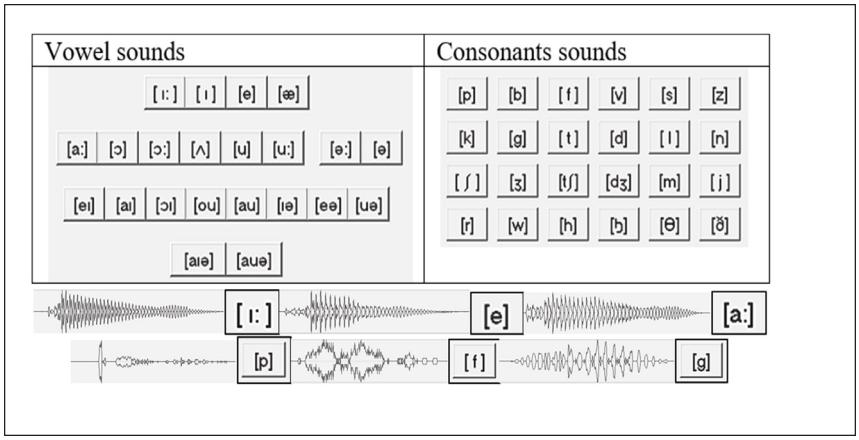

Figure 1. Vowel and consonants sounds

perceived properties are an important factor when it comes to the development and engineering of communications equipment. Speech is made up from a number of different types of sound which include voiced sound, unvoiced and plosive. All of these sounds are influenced by the person's sinuses and nasal cavities and all make up what we understand as normal human speech. Some basic sounds in the English language and their sonogram are shown in Figure 1.

Reference are voiced Bill Shephard, coordinator of the Syndicate examinations in English as a foreign language at the University of Cambridge. All referenced samples have continuous and smooth signal envelopes [10].

3. Sound Samples Envelope Analyzing

3.1. Basics Envelop Curve Smoothness Analyses

After comparing the number of significant deviations from the smoothness with an average conversation can assess the quality of a call. The idea is to apply voice quality prediction model to achieve optimum end-to-end voice quality.

A smooth function or a continuously differentiable function is a function that has a continuous derivative on the entire definition set.

It is possible to make analogy with the physical movement and signal envelope. The first derivative or rate of change of envelope's amplitude will be analogous to the speed of the physical object. The second derivative or the velocity change rate will be the acceleration.

Any sudden change in speed and acceleration are a reflection of a hard or soft impact.

In the signal envelope, each sharp jump on the first or the second derivative speaks of distortion of the smoothness of the shape, and hence of the possibility that it may be due to interruption or jitter.

The sound attack front may be due to some of the specificity of the speech. Another simple way of describing the attack phase, consists in estimating the amplitude difference between the beginning and the end of the attack phase. Another description of the attack phase is related to its average slope.

The number of jumps above a certain value can definitely be interpreted as a disturbance and disruption of the speech intelligibility.

The quality of a telecommunication voice service is largely influenced by the quality of the transmission system. Nevertheless, the analysis, synthesis and prediction of quality should take into account its multidimensional aspects.

After comparing the number of significant deviations from the smoothness with an average conversation can assess the quality of a call. The idea is to apply voice quality prediction model to achieve optimum end-to-end voice quality.

3.2. Envelope Extraction Using the Signal

The envelope extraction is based on two alternate strategies: either based on a filtering of the signal, or on a decomposition into frames via a spectrogram computation.

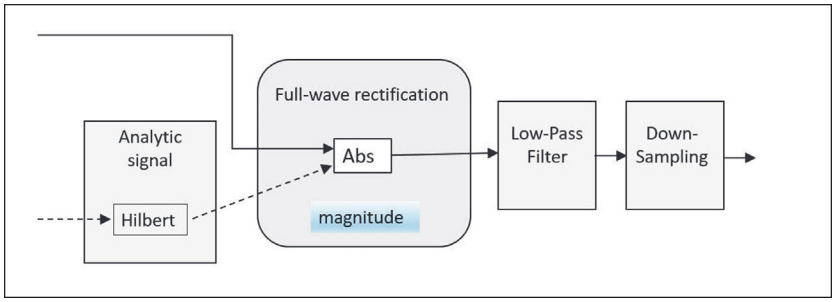

Figure 2. Envelop extraction process

The envelope of the signal is a feature that was built on the characteristic points of the signal, for example, on the extremes. Each (discrete or continuous) signal are local extremes: the local maxima and local minima. As a result, it is possible to build two envelopes: the lower envelope constructed by local minimum points, and the upper envelope constructed by local maximum points. This example shows how to extract the signal envelope using the signal.

The waveform envelope distortion is reviewed. In normal telephone signal amplitude has no abrupt changes and the curve of the waveform envelope is smooth. Large jumps occur in case of problems such as jitter lost packets, and so on. Jumps in the value of the first derivative defined numerically is an indication of a problem. The proposed method is based on analysis of the smoothness of the waveform envelope by numerically determining the first derivative. The phone sound usually has enormous volumes and is easily affected by noises. Furthermore, for reasons of the complex and highly non-stationary nature of phone sound signals, they should be segmented into components for the first step of automatic analysis and classification. To obtain proper information, signal is divided into small portions – which are processed independently. After exhausting the entire length of the processed signal received items of jumps in the differential are added together. In value of the amount compared with the averages can assess the quality of the conversation.

Different segments of filter length equal to 300 to obtain a smoother shape.

Here is an example of audio file with its envelope and corresponding first derivative:

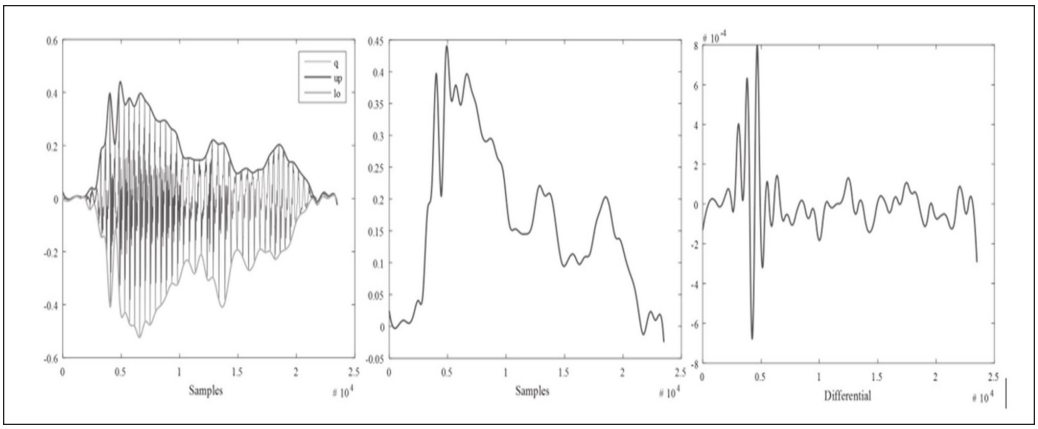

Figure 3. Envelop and corresponding first derivative of vowel [au]

3.3. Experimental Environment

We compare the output (out signal) with an input (in signal). The algorithm must include monitoring and measurement (or calculation) the basic parameters of the language.

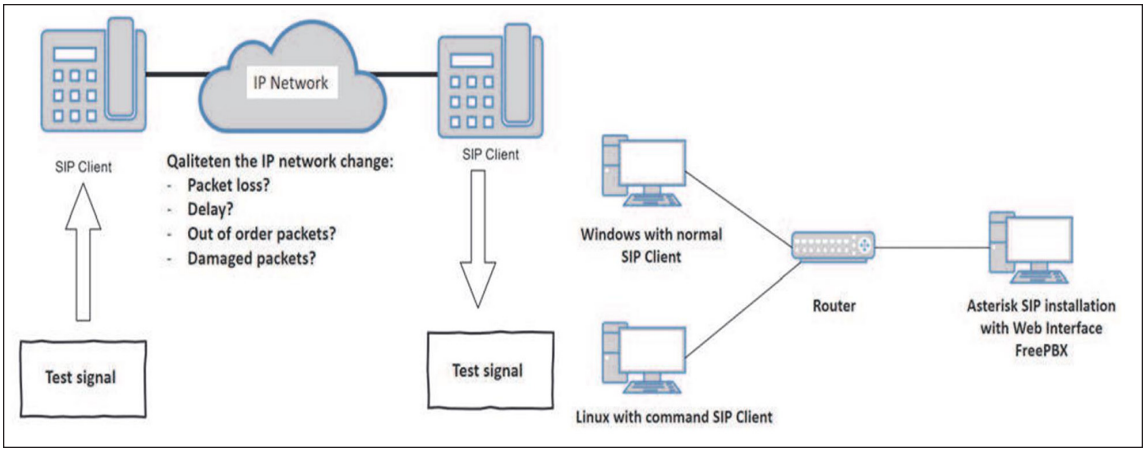

Figure 4. Experimental environment

3.4. The Input Test Signal is Sawtooth 1kHz

Envelop and histogram of the signal without defects is shown in Figure 5.

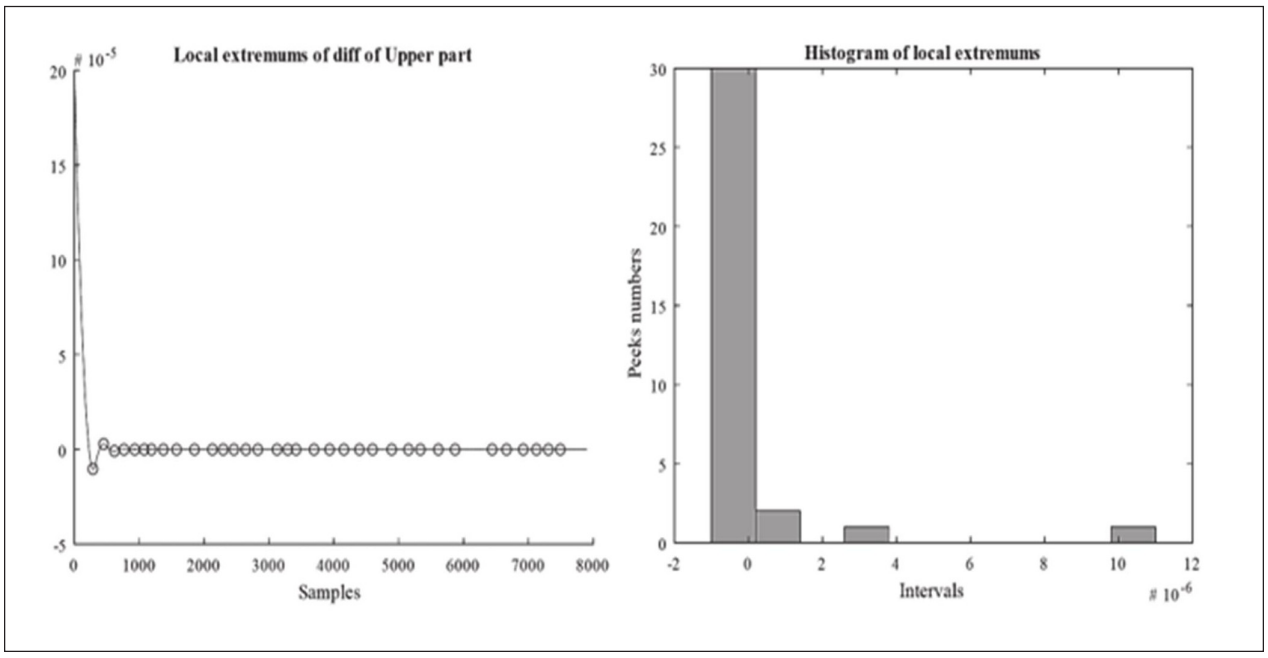

Figure 5. Sample without defect

The range of values is 10^{-5}. The histogram of local extrema shows only one great value due to first jump of the signal in moment on start.

3.5. Output Test Signal - Sawtooth 1kHz with Defects

We set standards and stepwise parameter degradation network to affect the sound quality - from RTP (protocol for the transmission of sound) using Linux ***"tc"*** command. The result is judged what kind of degradation of network parameters (delay, loss, jitter) as it affects most intelligibility.

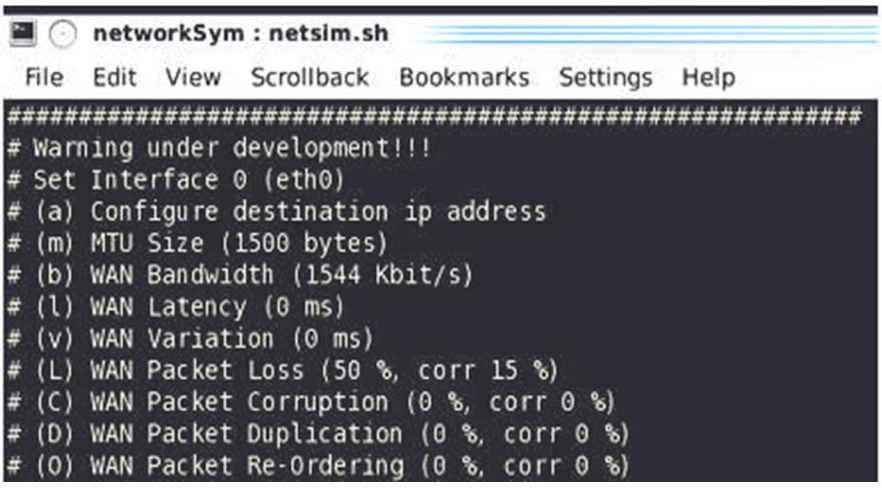

Figure 6. Settings of experimental environment

Test signal with defects Increase the filter length to 300 to obtain a smoother shape.

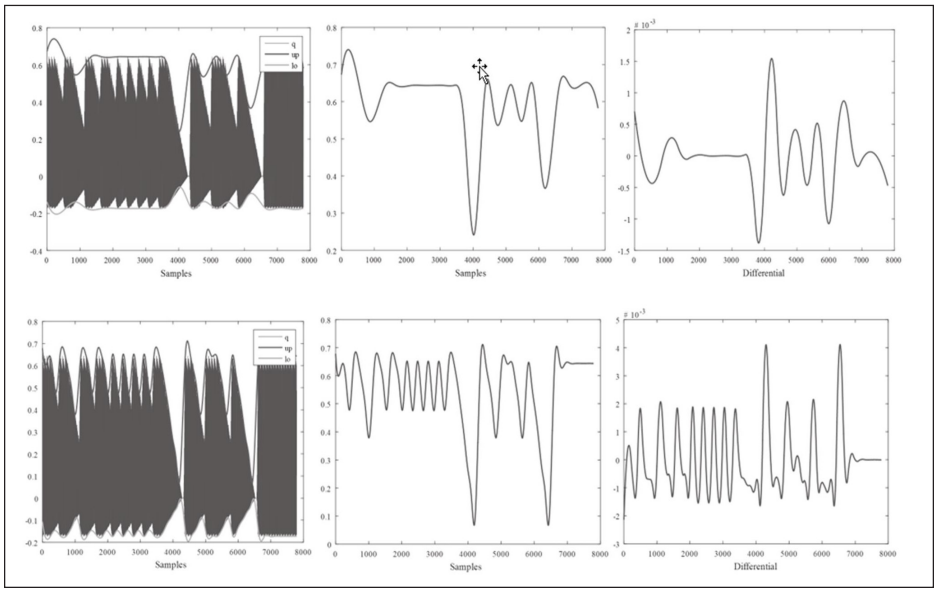

Figure 7. Sample size 300 and 100

Histogram of first derivative Values: [7 14 6 12 9 1 0 0 0 2]. The range of values is 10^{-3}. Second derivative Values: [13 9 4 5 2 2 12 3 0 2]. The range of values is 10^{-5}. After comparing the number of significant deviations from the smoothness with an average conversation can assess the quality of a call.

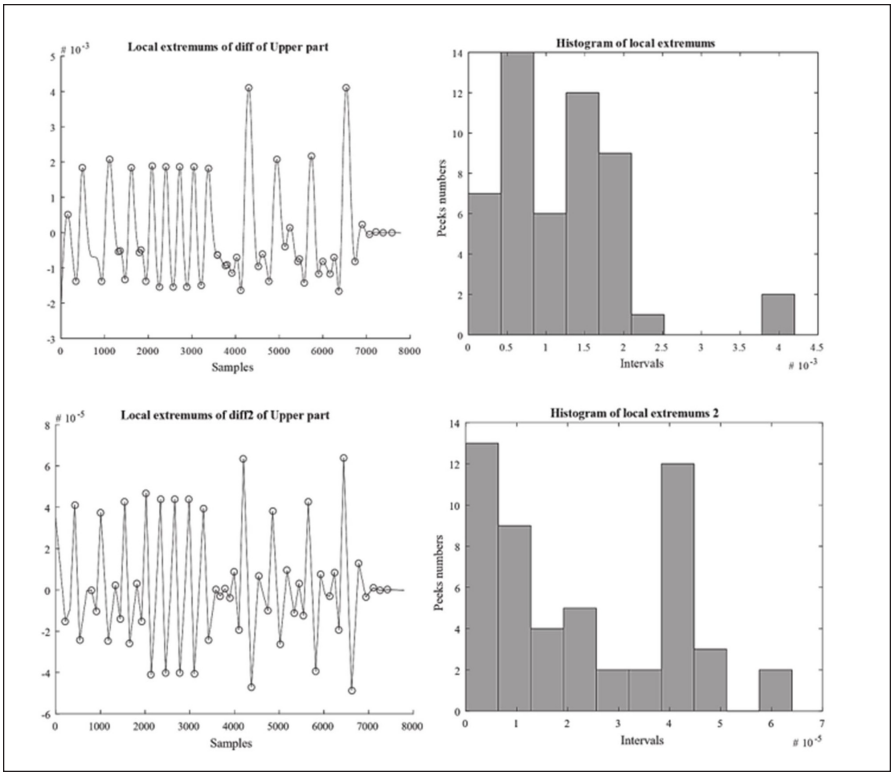

Figure 8. Histograms and local extremums for sample size 300 and 100

The histogram of local extrema shows many great values.

4. Proposed Algorithm

The derivative of function $f(t)$ must be defined for all t .

$$f'(t) = \frac{f(t+h) - f(t)}{h} \tag{1}$$

$$f' = \frac{f(t+h) - 2\ f(t)\ +\ (t-h)}{h^2} \tag{2}$$

Formula (2) is a variation of the numerical differentiation formula using three adjacent values. It results in a smoother and more accurate value of the first derivative.

In general terms, the method of the algorithm is as follows.

1) Located the signal extremes. They must be sought between every two consecutive sign changes.

2) Build two envelopes signal: lower and upper. Obtain the analytic signal. Extract the envelope, which is the magnitude (modulus) of the analytic signal. Plot the envelope along with the original signal.

3) Determine whether a function is continuous by numerically differential. Finde first derivative of upper part (1) or (2).

4) Finding the absolute value of the first derivative function and find the peaks.

5) Finding all local extremums.

6) Building a histogram.

7) Comparing the result with ***Network performance objectives for IP-based services - Y.1541 (12/11).***

8) Assessing the quality of a call.

5. VoIP and QoS. is the Signal Acceptable?

For enterprise VoIP to compete successfully with the Plain Old Telephone System, the voice quality should be at least equal to analog phones or better. Audio quality was a significant concern in the earliest implementations of VoIP, when the technology was fairly new.

Audio calls will thus be subject to high levels of jitter, degrading the quality of conversations. If the QoS settings are correct and network traffic is at its usual levels, there should not be any significant problem with intelligibility. The sound quality of VoIP calls drops dramatically when UDP packets are not received in a timely fashion, if packets are lost or reordered.

QoS may be measured in a number of different ways, several of which are detailed in various IETF standards for RTP such as RFC 3550 and RFC 3611. The QoS usual common monitoring program monitors the quality of a network connection by looking at “quality of service” parameters like VoIP jitter, packet loss, packet delay variation, duplicate packets and other readings.

Several telephony phenomena, further exacerbated by VoIP processing, affect the character of voice conversations without really affecting sound quality at all. These phenomena include end-to-end and round-trip network delay, delay variance (jitter), and echo. Intelligibility is directly impacted by noise or other types of distortion.

The clarity of a voice signal or voice channel has been measured subjectively according to ITU-T Recommendation P.800

resulting in a mean opinion score (MOS).

It is very difficult to separate the quantification of voice quality (the evaluation or measurement of noise and distortion) from the subjective experience of the human talker and listener. Voice quality can really only be judged relative to the situation being assessed and the human experience of it [8].

6. Conclusions and Future Work

The problem of real-time quality estimation of VoIP is of significant interest. This paper has shown an approach for solving this problem by employing the envelope of the signal. One of the main objectives of this research was to estimate the effect of fragmentation on speech quality. This is due to the fact that the analytical algorithms do not model the effect of fragmentation on speech quality [8] [9]. Hence, the effect of fragmentation can be mapped only by conducting suitably designed formal subjective tests.

The focus of the current research has been on estimating the effect of all VoIP traffic parameters that affect the listening quality of a telephone call in combine. A future objective would be to derive a neural network model for conversational quality estimation of a call. Conversational quality suffers due to increase in the end-to-end delay of a call. Clearly, the next objective would be to estimate the particular effect of VoIP traffic parameters and their impact on the signal quality.

Acknowledgement

Research, the results of which are presented in this publication are funded by internal competition, TU-Sofia 2017, CONTRact № 172ПД0010-07 for a research project to help PhD.

References

[1] Black, U. (2000). *Voice Over IP*, Upper Saddle River, New Jersey: Prentice Hall PTR, 2000.

[2] Cray, A. (1998). *Voice Over IP: Hear's How*, 1998.

[3] ITU-T Recommendation P.563.

[4] T-REC-Y.1541 IPDV

[5] P.800: Methods for Subjective Determination of Transmission Quality - ITU -T Recommendation P.800

[6] ITU-T Rec. G.1020

[7] Pennock, S. (2002). Accuracy of the Perceptual Evaluation of Speech Quality (Pesq) Algorithm, *Measurement of Speech and Audio Quality in Networks (MESAQIN)*, 2002.

[8] Sun, L. F., Ifeachor, C. (2002). Subjective and Objective Speech Quality Evaluation Under Bursty Losses, *Measurement of Speech and Audio Quality in Networks (MESAQIN)*, 2002.

[9] Moller, S. (2000). *Assessment and Prediction of Speech Quality in Telecommunications*, Kluwer Academic Publishers, Boston/Dordrecht/London, 2000, 116-117.

[10] A Linguistic Feature Representation of the Speech Waveform http://www.mit.edu/~mitter/SKM_theses/93_9_Eide_PhD.pdf