

END-TO-END SPEECH QUALITY ANALYSIS FOR VOIP

L. Sun¹, G. Wade¹, B. Lines¹, E. Ifeachor¹, D. Le. Foll²

Abstract

The paper first presents the factors affecting end-to-end speech quality in VoIP, such as end-to-end frame loss/concealment and end-to-end jitter. It then discusses how these factors affect the current objective speech quality measurement algorithms and suggests essential modifications for VoIP. It also highlights extensions to the basic objective measurement, such as time-varying speech measures, and the use of non-intrusive techniques.

1. Introduction

Common VoIP network connections normally include the connection from phone to phone, phone to PC (IP Terminal or H.323 Terminal) or PC to PC, as shown in Figure 1. The Switched Communication Network (SCN) can be a wired or wireless network, such as PSTN, ISDN or GSM. The end-to-end speech transmission quality will depend on the quality of the gateway (G/W) or IP/H.323 terminal and IP network performance.

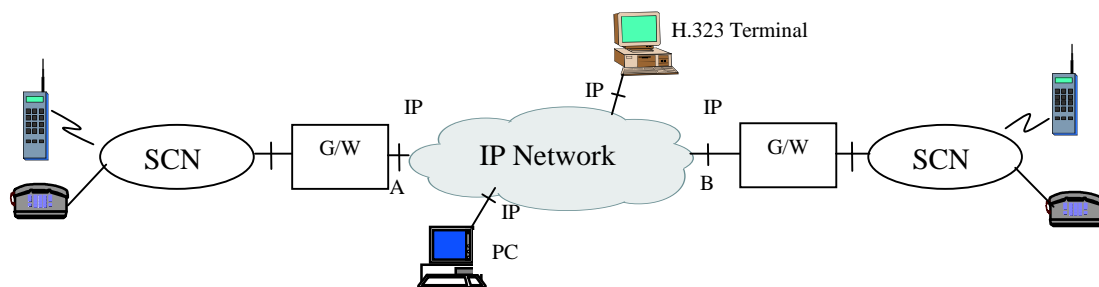


Figure 1. VoIP network architectures

Current research is concentrating on how to guarantee IP Network performance in order to achieve the required Quality of Service (QoS). Also, the impact of network parameters such as packet loss and jitter on speech quality have been broadly analyzed [1][2]. On the other hand, research is underway to improve the speech quality for “best effort” IP networks, and different compensation strategies for packet loss [3][4] and jitter [5][6] have been proposed to improve speech quality even under poor network conditions. Some simple and efficient compensation strategies have been used in current VoIP gateways or terminals and more sophisticated techniques are likely in the near future.

Different packet loss compensation strategies (such as packet loss concealment, PLC) and different adaptive jitter buffer adjustment strategies, will lead to different end-to-end packet/frame loss and delay jitter even under the same network impairments. The end-to-end packet/frame loss and jitter will be more important for perceived speech quality than the network packet loss and jitter, although the latter are two important factors that reflect network performance.

¹ School of Electronic, Communication & Electrical Engineering, University of Plymouth, United Kingdom

² Wavetek Wandel Goltermann, Plymouth, United Kingdom

Regardless of the strategy that is used to improve IP network performance or gateway/terminal performance, the purpose is to achieve a satisfactory speech transmission quality. The final judgement of speech quality still depends on the end user's perception. Subjective speech quality MOS (Mean Opinion Score) scores are considered the most powerful and recognised measurement of speech quality. Since subjective measurement is time-consuming and expensive, objective speech quality measurement has been proposed to estimate the subjective quality of a network. Typical objective measurement methods include PSQM (Perceptual Speech Quality Measurement [7]) and PAMS (Perceptual Analysis/Measurement System [8]). PSQM has been chosen as the ITU standard (P.861, 2/98) for objective speech quality measurement. Since these objective measures were originally developed for the assessment of speech quality for low bit rate codecs, the impact of packet loss or variable delay (two important impairments in VoIP) were not considered in their first versions. Current work in ITU Study Group 12 therefore focuses on new objective speech quality assessment methods for VoIP, GSM and other networks. Modified PSQM or PAMS (e.g. PSQM+[9], PSQM99, PAMS release 2.0 and 3.0) and other new algorithms have been proposed for the competition of the new ITU standard, which is expected to be available at the end of this year.

In this paper, we will first discuss the major end-to-end speech quality impairment factors for VoIP, such as end-to-end frame loss/concealment and end-to-end delay jitter. This is followed by a summary of current objective speech quality assessment methods and some new developments. Then the limitations and possible improvements to these algorithms for VoIP applications are analyzed. The paper concludes with a review of possible extensions to the basic objective measure.

2. End-to-end Speech Quality Impairment Factors for VoIP

2.1 End-to-end frame loss/concealment

Packet loss is the major cause of speech quality impairment in VoIP networks, and clearly, this is more serious if a packet contains a relatively large number of speech frames. The effective end-to-end packet loss is a little different from the network packet loss, which is normally caused by network congestion and is detected by the application protocol, such as RTP. For example, some packets will be discarded at the terminal (gateway) due to the limited size of the playout buffer, or if they arrive too late, or are recovered too late by the PLC.

Lost packets may be compensated for by a variety of recovery strategies. Packet loss compensation normally includes local repair (interpolation of missing data using the surrounding packets), or interleaving, or the use of packet-level FEC for sending redundant information. For example, a low quality redundant LPC frame can be used to replace a lost main stream speech frame (e.g. a G.729 frame) in an FEC scheme. Some CELP-based speech codecs (e.g. G.729 and G.723.1) have built-in "bad frame masking" or PLC. If external FEC recovery and a codec with built-in recovery are both used in a gateway/terminal, there will be almost no end-to-end frame loss under reasonable packet loss conditions.

Clearly, given some loss compensation strategy it is apparent that the effective end-to-end frame loss, and not the network packet loss, is the more important parameter for perceived speech quality. We can also classify a received (recovered) speech frame into one of three types:

- Normal frame (including error-corrected frame)
- Lost frame (silence or comfort noise)
- Concealment frame (the recovered frame is different from the original)

Speech quality for a normal frame is degraded mainly by codec impairment in a gateway/terminal. A lost frame will be silence or comfort noise. Concealment frames will differ, depending on the PLC strategy used, and in general the quality of a concealment frame will lie between that of a normal frame and a lost frame.

2.2 End-to-end delay jitter and delay

As IP networks are based on "best effort" policies, different IP packets from the same source may have different delays because of differing traffic conditions and routes. The variation in delay is known as network jitter or delay jitter. Gateways or terminals use playout buffers (jitter buffers) to compensate for network jitter

or to integrate “late packets”. By increasing the size of the jitter buffer, a terminal is able to resynchronize more packets, but with more delay. There is therefore a trade off between jitter compensation and packet loss.

Jitter compensation is accomplished primarily by using adaptive playout buffer algorithms [5][6]. These generally work by using some measurement of packet delay to update the playout buffer on a talkspurt by talkspurt basis i.e. a variable delay occurs in silent periods. Clearly, this leads to distortion of silence, or variable speech burst delay. Some algorithms also adjust the buffer in mid-talkspurt (variable delay during talkspurt), leading to distortion of talkspurts or speech clipping.

If a fixed-size playout buffer is used there will simply be a fixed playout delay for each speech frame, but no end-to-end delay jitter. Conversely, buffer adjustment results in end-to-end delay jitter, which is apparently different from the network jitter. If adjustment is carried out during speech silent periods it will cause a variable duration of the silent intervals between speech talkspurts as shown in Figure 2. Subjectively the adjustment may be imperceptible, but it is problematic for comparison based objective speech quality measurement methods. If the adjustment takes place in mid-talkspurt, it will either insert a frame or several frames of silence, or “slip” a frame or several frames of speech (speech clipping). This again complicates comparison- based objective speech quality assessment methods (see next section).

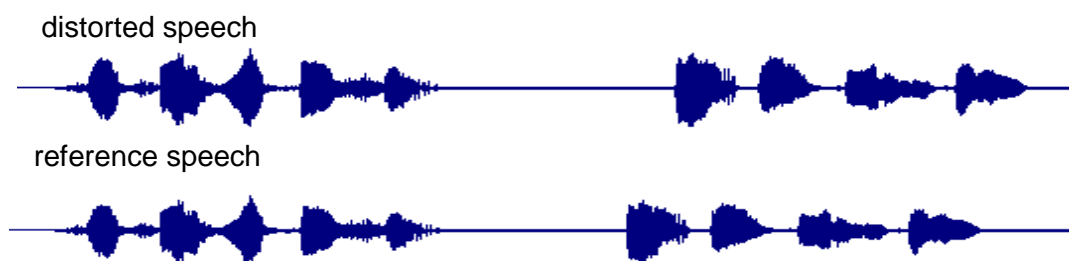


Figure 2. End-to-end delay jitter caused by buffer adjustment in a silent period

The end-to-end transmission delay is also different from the network transmission delay. For a phone-to-phone connection the network transmission delay is associated with section A to B (Figure 1) whereas the actual end-to-end transmission delay will also include playout buffer delay, codec delay, and PSTN transmission delay. End-to-end delay is a critical issue for VoIP network and has a direct effect on conversational speech quality.

2.3 Codec

Codec impairment is one of the major aspects of speech quality degradation. Its quality can be assessed easily by objective speech quality measurement. Codecs for use in VoIP include G.723.1(5.3/6.3 Kbps), G.729 (8 Kbps, CS-ACELP), G.729A (low complexity version), G.728 (16Kbps, LD-CELP) , G.721 (32Kbps, ADPCM) and G.711(64Kbps, PCM). Codecs G.723.1 and G.729 are in normal use.

2.4 Echo

There are two types of echo in VoIP applications. The first is the usual far-end echo caused by the 4-to-2-wire hybrid conversion (line echo, or hybrid echo). The user will hear his/her own voice signal reflected back from the remote central office or gateway’s line-card hybrid. The second form of echo occurs when free-air microphone and speakers are used, as is the case for most PC endpoints. The remote user’s voice signal produced by the speaker is picked up by the microphone and echoed back to the remote user. This echo is normally called acoustic feedback echo.

The echo canceller in a gateway or an IP terminal is used to alleviate the echo. The residual echo is perceived by the end user and so is a factor when measuring conversational speech quality. As echo signal in VoIP scenarios also suffers from packet loss and variable delay, this makes echo canceller difficult to predict echo and makes echo impairment a problem in VoIP.

3. Objective Speech Quality Measurement

3.1 Objective perceptual speech quality measurement

Objective perceptual speech quality measurement systems normally use two input signals, namely a reference signal and the degraded signal measured at the output of the network or system under test. Due to non-linearity arising from the codec, the signals should be speech recordings or artificial speech-like test signals. Typical measurement methods are PSQM and PAMS. Signal processing normally includes pre-processing, psycho-acoustic modelling, and a speech quality estimation model. The differences between these algorithms lie in differences between models. For example, the ITU P.861 PSQM algorithm consists of a perceptual model and a cognitive model (Figure 3), whilst PAMS includes an auditory transform (psychoacoustic model) and perceptual layer processing [8].

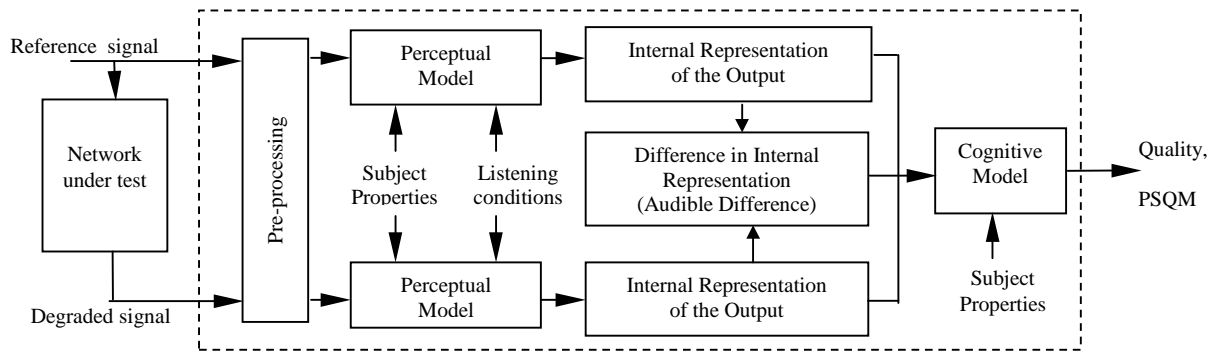


Figure 3. Structure of PSQM

As an example, we summarise the processing of PSQM as follows:

- Pre-processing: including delay adjustment (time alignment), loudness adjustment (equates loudness) and duration adjustment.
- Transformation: transforming the signal into a psychophysical representation that approximates human perception.
- Calculation of perceptual difference distance: calculating noise disturbance, N_i , for frame i or N (PSQM value) for the whole speech segments.
- Mapping to objective MOS: mapping from the PSQM value to MOS score.

3.2 Limitations and improvements for VoIP applications

The PSQM specification (ITU P.861) assumes that the source speech is “clean” and that there are no channel degradations such as transmission bit errors, frame erasures, cell loss, or packet loss. In PSQM’s pre-processing unit, only the fixed network delay is estimated to keep the two signals in time alignment at beginning of a comparison.

Attempts have been made to modify the method, and to make it more robust to a variety of applications. For example, PSQM+ has been proposed to consider the effect of loud distortion and time clipping by introducing scaling factors for the frames of time clipping and loud distortion. Variable delay or time-warping problems have been considered in PSQM99 or PAMS’s new versions[10].

According to our analysis for end-to-end jitter and frame loss/concealment in the previous section, here we suggest some modifications/improvements for the normal objective speech quality measure methods for VoIP applications.

- End-to-end delay jitter

Playout buffer adjustment algorithms may make the buffer adjustment during either silent periods or mid-talkspurt. Either way, it is necessary to keep strict time alignment for the two signals being compared. If an adjustment happens in a silent period, it is necessary to perform realignment before the next talkspurt. If it happens in mid-talkspurt (Figure 4), time-alignment strategies e.g. cross-correlation could be used to find the matching signal frames for the calculation of the Noise Disturbance for each frame after the adjustment.

Obviously if buffer adjustment (end-to-end delay jitter) is very small, the effect could be imperceptible by the end user. In this situation, we only need to remove delay jitter to keep the two signals aligned and do not need to consider the corresponding subjective impairment. If the end-to-end delay jitter is greater than a subjective threshold, the playout impairment itself should be considered as one factor to weight the speech quality measurement algorithm.

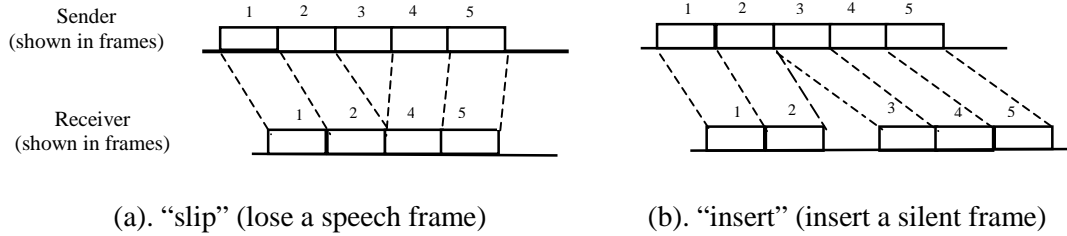


Figure 4. Buffer adjustment in mid-talkspurt (adjustment size = 1 frame)

- End-to-end frame loss and concealment

PSQM treats all frames (normal frame, lost frame and concealment frame) with the same processing methods or the same weighting factors. For the lost frame in Figure 5(a), the power spectrum components of the frame are very small leading to a very high noise disturbance N_i which, after averaging over all frames, gives higher PSQM value. PSQM+ was proposed by introducing an additional scaling factor for lost frames, thereby compensating for the lost frame effect.

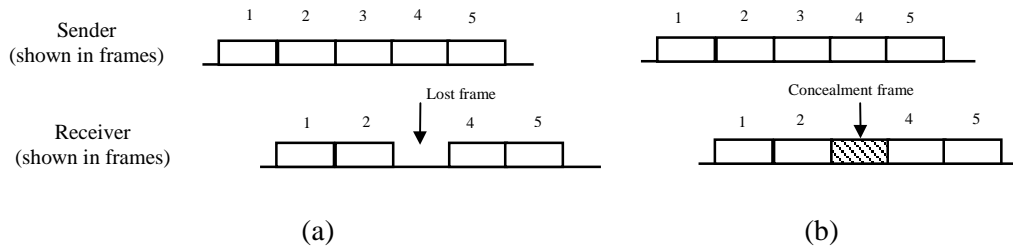


Figure 5. Lost frame and Concealment frame

If packet loss recovery strategies are used for loss compensation, then the lost frame will be replaced by a concealment frame (Figure 5(b)). There will generally be little difference between the power spectrum components of a normal frame and a concealment frame, and the Noise Disturbance Distance N_i for the concealment frame will lie between that for a lost frame and that for normal frame. According to research on PLC strategy [4], PLC normally works well for single packet loss recovery, but degrades rapidly for loss of more than one packet (burst frame loss). The latter situation may be perceptible to the end user. In this case, if N_i is consistently higher than that for normal frame (above a perceptual threshold), a new weighting factor for concealment frame compensation could be added in the objective speech quality measurement algorithm.

3.3. Other extensions

Objective speech quality measurement methods mentioned above are listening-only quality assessment and based on normal stationary transmission conditions. Currently there are several extensions under study.

- From listening-only to conversational speech quality assessment

Conversational tests are the preferred method of network evaluation and unidirectional models are being extended to bi-directional models for conversational performance assessment. A possible idea is to have two devices, one at each end, holding a conversation with each other. Two conversational models based on PAMS have been proposed [11][12]. The first applies an echo/delay weighting to PAMS and the second implements a full conversational model using PAMS for core speech quality predictions. Conversational related factors, such as delay, talker echo, listening level and vocal level equilibrium, have been considered in their models.

- From stationary to time-varying speech quality assessment

Current objective speech measure methods are based on normal stationary transmission conditions and are typically designed and calibrated using short sentence stimuli to give predictions of subjective opinions. These methods are not suitable for the assessment of a realistic time-varying system, as might occur for example in fading radio transmission, variable rate speech coding by DCME or VoIP.

For VoIP scenarios even without a mobile network connection or use of an adaptive coding algorithm, variable network packet loss and jitter make the speech quality time-varying. Prediction of a MOS score for a longer time period (paragraph test) will be more meaningful to reflect the overall speech quality than an instantaneous speech quality score (sentence test). Traditional objective speech quality assessment methods need further modifications to achieve this goal.

- From intrusive perceptual quality measurement to non-intrusive speech quality evaluation

The objective speech quality measurement methods mentioned above are comparison based intrusive tests, and so use a reference signal. They are not suitable for long-term monitoring of actual networks. Non-intrusive speech quality measurement can be performed in-service during ordinary calls and can be used for long-term speech quality monitoring in a network. In-service Non-intrusive Measurement Devices (INMDs), as specified in ITU-T P.561 [13] and Call Clarity Index (CCI) [14] have been used in current PSTN networks. The investigation of INMD/CCI concepts for VoIP applications is under study.

4. Conclusions

The major end-to-end speech quality impairments in VoIP networks have been presented in this paper. The impairments of end-to-end frame loss/concealment and end-to-end delay jitter depend on the loss and jitter compensation strategy employed. There are currently no standard loss and jitter compensation methods in VoIP networks and so various strategies are currently used. The normal objective measurement methods must be modified for VoIP in terms of frequent time alignment and the quality of concealment frames. Long term, speech quality measurement in VoIP will also involve conversational quality, time-varying assessment, and non-intrusive measurement.

5. References

- [1]. TR 101 329 V 2.2.2 (1999-05), (Technical Report), Telecommunication and Internet Protocol Harmonization Over Networks (TIPHON); General aspects of Quality of Service (QoS)
- [2]. L.A.R. Yamamoto, J.G.Beerends, Impact of network performance parameters on the end-to-end perceived speech quality, Expert ATM Traffic Symposium, Mykonos, Greece, Sep. 1997
- [3]. H. Sanneck, Adaptive Loss Concealment for Internet Telephony Applications, Proceedings INET'98, Geneva/Switzerland, July 1998
- [4]. J. Rosenberg, G.729 Error Recovery for Internet Telephony, Project Report, 1997, Columbia University (<http://www.cs.columbia.edu/~jdrosen/e6880/index.html>)
- [5]. J. Rosenberg, L. Qiu, H. Schulzrinne, Integrating Packet FEC into Adaptive Voice Playout Buffer Algorithms on the Internet, Proceedings of IEEE Infocom 2000, March 2000, Tel Aviv, Israel
- [6]. Po L. Tien and Maria C. Yuang, Intelligent Voice Smoother for Silence-Suppressed Voice over Internet, IEEE Journal on Selected Areas in Communications, Vol. 17, No.1, Jan. 1999
- [7]. ITU-T P.861 (02/98) Objective quality measurement of telephone-band (300-3400 Hz) speech codecs
- [8]. EG 201 377-1 V1.1.1 (1999-04), (ETSI Guide) Specification and measurement of speech transmission quality; Part 1: Introduction to objective comparison measurement methods for one-way speech quality across networks
- [9]. ITU-T Contribution COM 12-20-E, Improvement of the P.861 Perceptual Speech Quality Measure, KPN Research, Netherlands, Dec. 1997
- [10]. Antony Rix, Richard Reynolds and Mike Hollier, Perceptual Measurement of End-to-end Speech Quality over Audio and Packet-based Networks, AES 106th Convention, Munich, 8-11, May 1999
- [11]. ITU-T Contribution COM12-Q13 Workshop, Innovations in Modelling, BT, Sep. 1998
- [12]. A W Rix, A Bourret and M P Hollier, Models of human perception, BT Technol J Vol. 17, No.1, January 1999, pp.24 –34
- [13]. ITU-T P.561(02/1996), In-service, non-intrusive measurement device - Voice service measurements
- [14]. Simon Broom, Philip Coackley and Phil Sheppard, Getting the Message, Loud and Clear – Quantifying Call Clarity, British Telecommunication Engineering, Vol. 17, April 1998