Accuracy Analysis on Call Quality Assessments in Voice over IP

Yi Han*, John Fitzpatrick†, Liam Murphy‡, Jonathan Dunne§
*†‡ School of Computer Science and Informatics
University College Dublin
Email: *yi.han@ucdconnect.ie, †john.fitzpatrick@ucd.ie, ‡liam.murphy@ucd.ie
§IBM Software Group, Dublin, Ireland
Email: jonathan_dunne@ie.ibm.com

***Abstract*—Voice over IP (VoIP) now has tremendous influence on the telecommunication market with its flexibility and price advantage. Users of VoIP expect call quality to be as good as, if not better than the traditional Public Switched Telephone Network (PSTN). However in VoIP, factors that are related to the IP transport network such as packet loss, delay, bandwidth, jitter, and voice encoding (codec) all affect call quality. Call quality assessment in VoIP systems is mainly conducted with off-line tests using the Perceptual Evaluation of Speech Quality (PESQ) [1] methodology. Another method that can be utilised is an on-line approach using the E-Model, which can be used in real time. However, these two methods have limits and inaccuracy, and often do not give the same results. Call quality assessment is often used to adjust system and codec parameters. Therefore, given inaccurate results, the system would decrease the adjustment efficiency or even inadvertently decrease call quality. The primary contribution of this paper is a comparison between the accuracy of PESQ and the E-Model investigated by conducting an extensive set of experiments in a real enterprise network using a widely deployed Voice over IP (VoIP) product. Experiments were conducted under varying controlled network conditions. The results show that under various conditions, loss rates, codecs and across a range of languages that there can be significant differences between the call quality measurement obtained when using the E-model versus a PESQ analysis.**

***Index Terms*—Voice over IP, PESQ, E-MOdel, call quality**

I. Introduction

Voice over IP plays an important role in the telecommunication field, and can be especially useful for enterprise users. VoIP utilises IP networks for transport. Thus it fundamentally reduces the cost of network operations and deployment leading to increased cost savings for both users and operators when compared to traditional Public Switched Telephone Network (PSTN) systems. One of the main users of VoIP technology has been enterprises, VoIP has allowed enterprises to reduce costs by combining both their IP network and their fixed line voice networks. However, VoIP is increasingly being used in a mobile environment; both by proprietary applications and as an operator delivered service. This usage will continue to grow, particularly has cellular networks migrates towards all-IP infrastructure and Voice over LTE (VoLTE) becomes more widely deployed.

Although VoIP is now widely deployed, the call quality and reliability of VoIP when compared to PSTN can still be a weak point. The PSTN has been developed over the last a few decades and utilises dedicated networks allowing it to provide guaranteed levels of quality. VoIP on the other hand is a relative newcomer with commercial systems only being released since 1995 [2]. Due to IP network limitations, factors like packet delay, loss, and jitter inevitably degrade the call quality. Furthermore, the choice of speech codec (coder-decoder) for a particular scenario can also contribute to this degradation. Moreover, implementation specific parameters can vary between systems, for example the jitter buffer size and codec implementation, and this means that speech quality metrics obtained from different systems are not directly comparable.

This paper investigates how these factors affect the audio quality in VoIP as interpreted by two of the most popular voice quality metrics and how these methods can produce different values when analysing the same voice call. A large set of experiments were performed on a widely deployed VoIP system under varying network conditions, codecs and languages to produce comprehensive results on how the various factors degrade the audio quality.

Traditionally, MOS results were obtained using subjective listening tests in which a large number of people listened to the decoded sound and scored it from 1 to 5, with a higher score indicating higher voice quality. The relation of MOS and human perception is shown in Table I. Subjective tests require large human resources and are quite time consuming, and for this reason not feasible for many situations.

Audio quality in VoIP software can be assessed by both off-line and on-line testing. Off-line tests predominantly use PESQ (Perceptual Evaluation of Speech Quality). This utilises real voice samples and is based on the comparison between an original audio file and the encoded, transmitted and decoded audio file at the receiver. PESQ measures the degradation between these two files and produces a quality metric: Mean Opinion Score (MOS), which matches the range of subjective test MOS results but slightly shifts due to the degradation of the digital coding process. The PESQ MOS is designed to range from 0.5 to 4.5 as voice over IP systems, a MOS result of 4.5 is considered as highest achievable score and 0.5 as the lowest.

For example, the most popular codec G.711 that was standardized by International Telecommunication Union-

TABLE I
MEAN OPINION SCORE (MOS)

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

Telecommunication Standardization Sector (ITU-T) gives a MOS of nearly 4.2 under ideal network conditions which means packet delay, loss and jitter are minimal. A MOS of 3.0 is considered to be slightly annoying to users and becomes unacceptable to users when the MOS further decreases.

However, these off-line tests require the original audio file and therefore cannot be done in real time for ongoing calls. On-line testing can be performed using the E-Model [12] which can be used to estimate call quality in real time for ongoing calls. The E-Model takes into account network factors that impact call quality and outputs an R score that ranges from 0 to 100 which can be easily mapped to MOS; this is further discussed in the next section.

This paper compares the voice quality scores obtained from both PESQ and the E-Model analysis. Experiments were conducted to investigate the correlation between the two assessment methods under a variety of conditions. Analysis of the obtained results shows the correlation between both. The goal of this work is to first assess the accuracy of using the E-Model in real time when compared to PESQ analysis such that the E-Model can be used for the real time analysis of live calls and subsequently show its potential for fault analysis, debugging, and codec selection and adaptation.

The remainder of this paper is organised as follows. Section II provides some technical background on both PESQ and the E-Model. Related work is described in Section III. Section IV describes the testing environment and experiments performed. Section V analyses and discusses the obtained results. Section VI concludes the paper and describes the next steps in this work.

II. PESQ & THE E-MODEL

The purpose of call quality assessment methods is to accurately determine the call quality as perceived by human listeners and speakers. Obviously, the most accurate and reliable method for this are subjective tests [3] performed by human assessors as it directly reflects human perception. However subjective tests must be conducted in a controlled environment to mitigate background noise and consider all parameters such as speech codec in use and listening equipment. These factors make it both time and labour intensive and not suitable for an automated system.

A. PESQ

According to the results presented in [1], PESQ has demonstrated acceptable accuracy for factors including codec evaluation, codec selection and so on. In voice quality assessment for VoIP systems, the quality can be measured by passing PESQ the clean original sample without noise, and the degraded sample. The PESQ algorithm compares the signals in the two samples, calculates the difference and finally gives an evaluation of the quality of degraded sample as an estimation of human perception. The PESQ score is mapped from 0.5 to 4.5, but the output range is mainly between 1.0 to 4.5 which is the normal range of MOS values that were found in listening quality experiment [1].

For convenience and cost reasons, PESQ analysis is used in place of subjective tests when a large number of tests are needed. For example, when testing call quality in VoIP products or in mobile networks a large set of tests can easily be conducted in a controlled manner by recording the input and output voice samples and passing them to a PESQ analysis tool.

1) Limitations of PESQ: PESQ requires the original sample and degraded sample for comparison and thus cannot be used in real time for active calls. Secondly, it has variances from subjective test results as described in [8], proving that the MOS values produced do vary significantly between subjective tests, especially for different languages.

2) Extension to MOS-LQ: MOS listening quality (MOS-LQ) was proposed and mainly aimed at giving results that are correlated to subjective tests by applying the 3rd order regression mapping function in (1), where x is the MOS from PESQ and y is the corresponding MOS-LQ:

$$y = \begin{cases} 1.0, x \leq 1.7 \\ -0.157268x^3 + 1.386609x^2 - 2.504699x \\ +2.023345, x > 1.7 \end{cases} \quad (1)$$

The mapping function shifts the MOS results from PESQ tool closer to the human perception, which is what E-Model tries to represent. Although it is designed to be applicable to a number of different languages, there is still variance from languages. In this paper, MOS-LQ is used for more accurate comparison with E-Model values as it is closer to subjective test results and is applicable to a wider range of network types (fixed, mobile, VoIP).

B. E-Model

The E-Model [12] is the most popular objective measurement method. It is a non-intrusive method that accepts network characteristics and codec information as inputs and outputs an estimated call quality score in real time. The output of E-Model is the "Rating Factor *R*" which can be mapped to MOS scale. E-Model was standardised in 2005 in [12] and further extended to wideband codecs in 2011 in [13].

Figure 1 shows the transmission parameters used as input to the computation model. Room noise of sender person (*Ps*) and room noise reviver person (*Pr*) representing environmental background noise and D-Factors, represent noise caused by the microphone and loudspeaker, which may vary from sender and receiver side and the values are handled separately in the algorithm. The parameters Sender Loudness Rating (SLR),

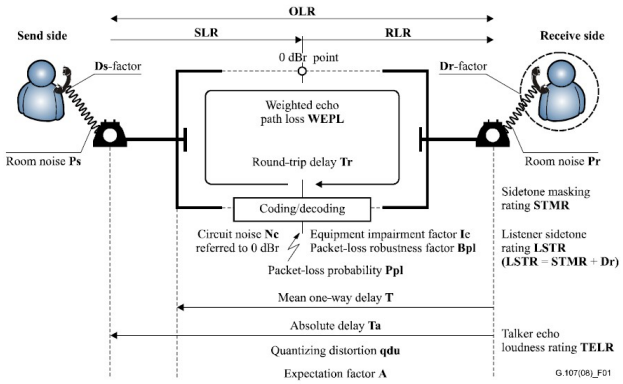

Fig. 1. E-Model algorithm parameters connections

Receiver Loudness Rating (*RLR*) and circuit noise (*Nc*) are referred to 0 *dBr* point by default. Other parameters including sum of SLR and RLR (Overall Loudness Rating, *OLR*), Quantizing Distortion (*qdu*), Equipment impairment (*Ie*), and advantage factor (A) are considered as values for the overall connection. The other parameters including Side-tone Masking Rating (*STMR*), Listener Sidetone Rating (*LSTR*), Weighted Echo Path Loss (*WEPL*) and Talker Echo Loudness Rating (*TELR*) are values considered only for receiver side.

The Rating Factor *R* combines all transmission parameters for the connection and is calculated by:

$$R = Ro - Is - Id - Ie_{eff} + A \tag{2}$$

where *Ro* represents the basic signal-to-noise ratio, including noise caused by the circuit and background noise; the *Is* factor is the combination of all impairments that occur more or less simultaneously affecting the voice signal; *Id* represents the quality degradation caused by delay and *Ie-eff* represents the quality degradation caused by low bit-rate codecs and also includes the degradation due to packet losses; the advantage factor *A* is an adaptive value that in many cases is constant.

III. Related work

The E-Model assessment technique has been widely used and some research has been done in comparison to subjective tests. The off-line assessment PESQ metric has been a worldwide industry standard for objective voice quality testing since it was standardised as ITU-T recommendation P.862 [1] in 2001. The performance of PESQ has been studied by research groups regarding its accuracy to different contents and languages [6][8]. In [9], the authors presented the objective quality assessment methods as signal-based models and parametric models which include PESQ analysis and E-model respectively. They also discuss the limitations and advantages of both approaches, but the correlation of results from different models were not given. Similar to the work that was presented in [7] that proposed speech quality measurement using PESQ based on open source VoIP software in a lab environment, we conducted the entire set of experiments based on enterprise VoIP product in real network environment and also discussed about misleading points when comparing PESQ results with E-Model results. PESQ and E-Model, as the most commonly used QoE measurement in VoIP, are also presented in detail along with other popular measurements in [10].

A. Research about E-Model

Language variance in the Mean Opinion Score from subjective testing leads to a further variance of MOS when converted from E-Model to R score. A Japanese version of MOS was proposed in [14], in which a linear regression mapping function was applied to convert the E-Model MOS to Japanese E-Model MOS.

B. Research about PESQ

In [6], the authors investigated the accuracy of PESQ and a conclusion is given that it is a useful tool in helping identify potential system problems but not accurate enough to specify speech quality requirements in Service Level Agreements (SLAs).

Research from the inventor of PESQ [8] proves that PESQ-Listening Quality (PESQ-LQ) scale gives good results in varying network conditions and languages but however MOS itself does vary significantly between subjective tests, especially between different languages. Also, their work concludes that a good mapping that performs well on average may still give scores that are consistently too high for some languages and too low for others.

IV. Experimental Setup

In order to assess the audio call quality in VoIP, an extensive set of experiments have been completed to determine codec performance in varying network conditions and languages. These tests are able to simulate a large number of calls and each group of calls with different codecs were tested under varying network conditions. The parameters varied in these network conditions are: end to end delay, bandwidth, packet loss rate, and jitter.

The experiments were conducted in real enterprise network which is “clean” (only lightly weighted) with real enterprise Voice over IP product. In order to keep a clean environment, a strong and robust link with with 1 Gbps bandwidth between two end points in the network was chosen. Packet loss rate is 0% by default. Round trip delay in the network is measured as 14 ms on average, and jitter monitored from tests are within 3 to 7 ms and these values are also taken into account to achieve as much accuracy as possible.

A framework, shown in Figure 2, was developed as a plug-in to the product in order to run a large number of tests automatically. Given a selected input source and codec by starting a point to point audio call from the sender, the receiver answers the call and starts receiving incoming audio stream. Audio at both sender and receiver sides are recorded for PESQ analysis. The audio data follows one path from sender to receiver, which in Figure 2 is from left to right. The framework receives audio file as input and the selected codec is used to encode the audio data into a stream of packets. Prior to

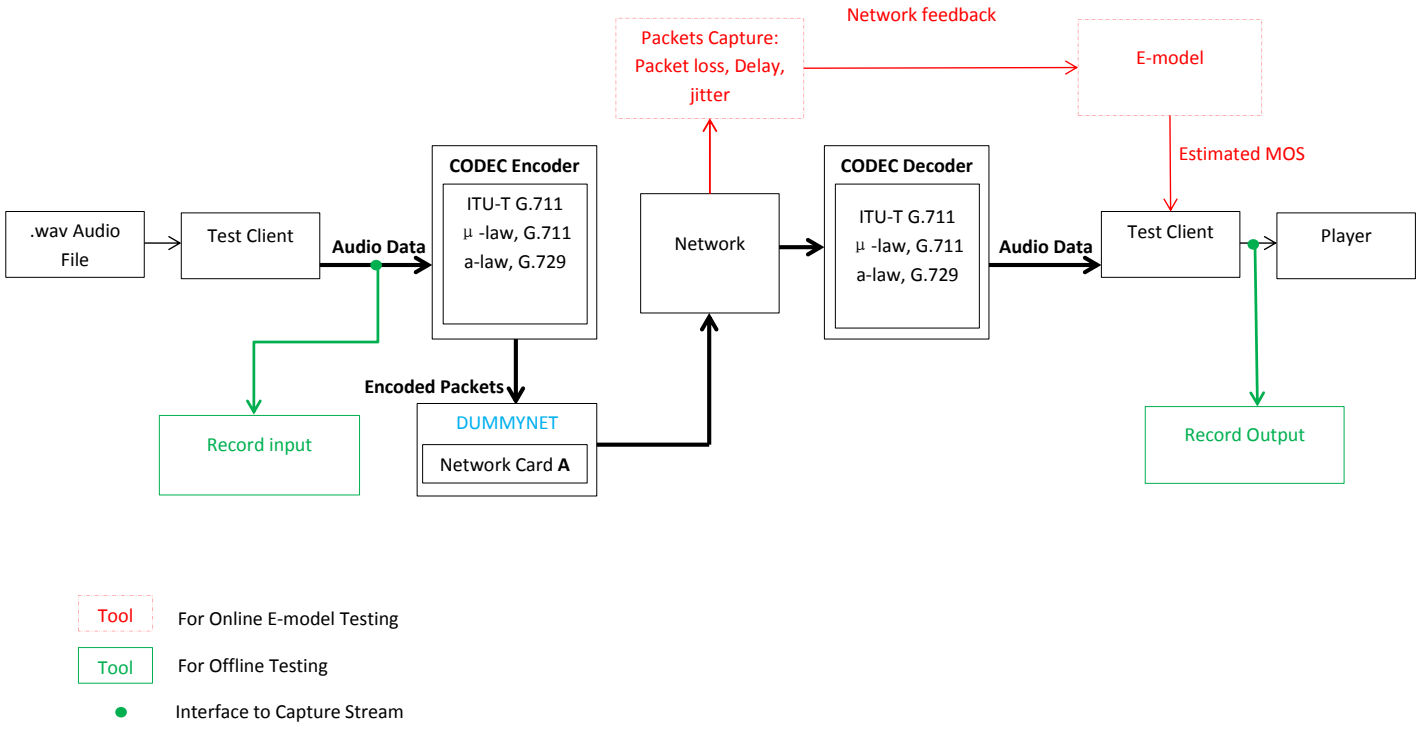

Fig. 2. Test Framework

the audio stream being placed into network, network packet loss may be added (if required) on the network card on the sender machine. After the audio stream passes the network and reaches receiver side, it is decoded by the same codec and becomes the output of receiver, which then is able to perform playback. Meanwhile, all packets going through the network are monitored at receiver side in order to monitor packet loss rate, delay and jitter by analysing the packet trace in *tshark*. These network characteristics are then fed into E-Model at the receiver side for computing MOS. For accurate analysis, the audio files at both sender and receiver sides are collected and compared by PESQ to get actual call quality. Thus, two ways of call quality assessment, the intrusive PESQ and non-intrusive E-, can be done.

The first thing needed in the implementation is a microphone emulator that accepts .wav audio file as input. The microphone emulator is a module that plays audio file as microphone. In this way, audio input can be controlled to be exact the same for each set of tests and also it is able to bypass the operating system and device impairments that may cause additional degradation of the call quality. The audio test samples are provided by Recommendation ITU-T P.501 as test signals for use in telephonometry. This recommendation describes that these test signals are applicable to various aspects of telephony products and proposes that the signals are used as test samples for the experiments in order to eliminate the risk of choosing inappropriate samples that lead to inaccurate results. The audio source samples are in PCM format with 256 kbps bit rate, 16 bit sample size and 16 kHz sample rate. The samples meet the recommended duration described in [1], which is about 8 to 12 seconds in duration containing pairs of sentences separated by silence.

As PESQ measurement is language and content dependent and E-Model is network and codec dependent, the set of parameters used for comparison are: sample contents (language, male or female, speech content), packet loss rate (jitter is considered as part of this variance). The samples selected from ITU-T P.501 are in American English, French, Chinese and Japanese with male and female respectively with different content. Thus, content and network variance can be tested under a designed experiment that keeps one variable fixed and changes the other one.

In the experiment, 21 sets of tests were conducted and in each set of test, network packet loss rate was set from 0% to 8% in 0.5% increments. This overall process is reproduced for each one of the codecs (ITU-T G.729, G.711 μ-law and G.711 a-law).

Packet losses were generated using *dummy-net* [4], where for non congestion-related drops, probabilistic match option is used to emulate links with uniform random loss patterns. This indicates that the packet losses are not content dependent which means the lost packets can happen in either speech

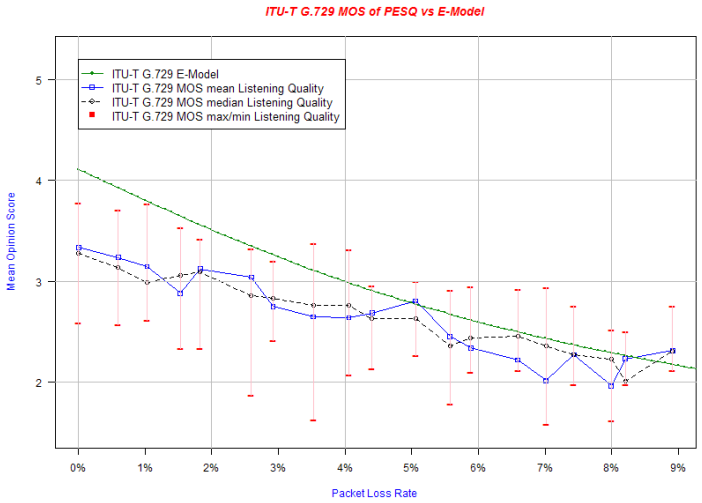

Fig. 3. Average MOS of ITU-T G.729 PESQ and E-Model

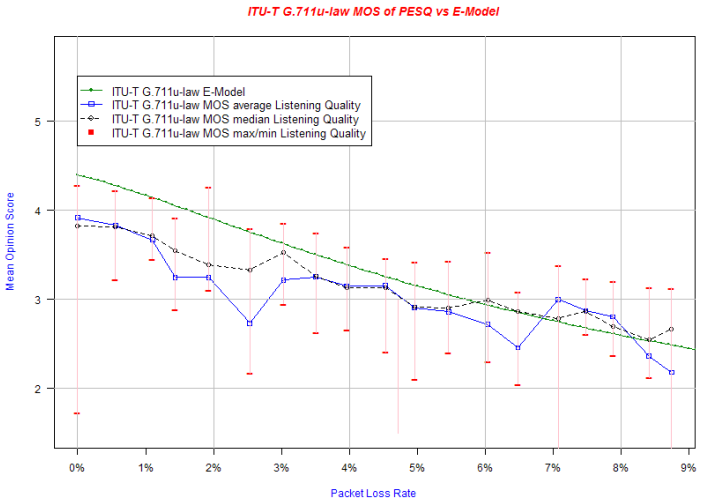

Fig. 4. Average MOS of ITU-T G.711μ PESQ and E-Model

period or silence period in an audio stream and it also makes a difference whether it happens in a spike or smooth of a speech period.

Due to the design of the framework, there is the possibility to conduct PESQ analysis by comparing the recorded received voice file with the original voice file rather than the recorded sending voice file, which means it the experiment can be done on the fly even while the call is going on. But the advantage of our implementation is that due to the delay (approximately 2.65 seconds) introduced during the period of requests forwarding going through the remote server, using the two recorded files at both sides can avoid content inconsistency in the recorded files and have exact same speeches of a specific period.

V. Results Analysis

The experimental results contain MOS values obtained from an extensive set of point to point VoIP calls under varying controlled packet loss rates, codecs and languages. The packet loss rates set for each test ranged from 0% to 8%. The tool used to generate the packet loss results in a normal distribution, centred at the requested loss rate value. For example, a 10 sec audio stream using ITU-T G.711 μ-law has 600 packets and loss rate is set to 5%, the theoretical number of lost packets should be 30, while in reality, the number could vary from 25 to 35 making the resulting packet loss 4.2% to 5.8%, or in some extreme cases, the deviation may be larger. For this reason, the value specified in the tool used to apply the packet loss is not used for computing the call quality, rather the real packet loss rate is calculated by analysing the audio RTP stream.

Comparison results for MOS of PESQ and E-Model are shown in Figures 3, 4 & 5, which show results for ITU-T G.729, ITU-T G.711 μ-law and ITU-T G.711 a-law respectively. The loss ranges in the graphs goes up to 9% which is beyond the theoretical maximum 8% loss rate. This is

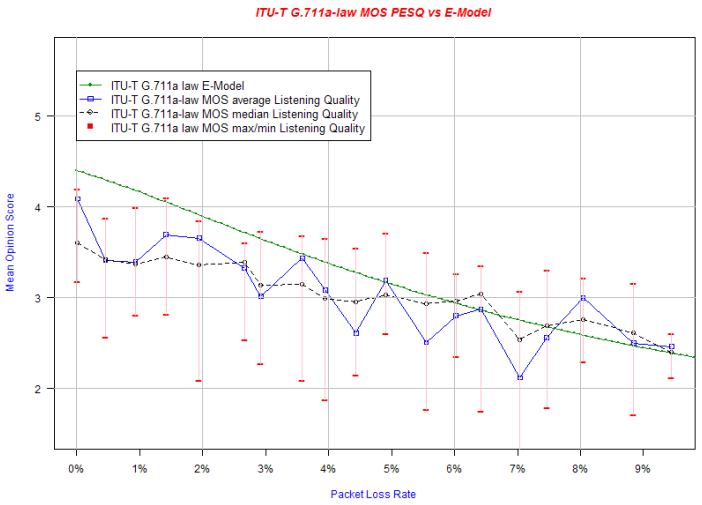

Fig. 5. Average MOS of ITU-T G.711a PESQ and E-Model

because the randomness of loss distribution that may lead to more packets lost in tests with 7%, 7.5% and 8% loss rates. These results highlight the difference between the call quality measurements obtained from PESQ comparing to those obtained when using the E-Model in real time. The aim is to show that although there are variations in the results, the E-Model is still reliable enough to be used as an on-line tool for enterprise product call quality evaluation and for making dynamic decisions such as codec selection and adaptation, and also PESQ analysis can be a good call quality indicator only under large number of tests.

The Mean Opinion Scores from PESQ are distributed close to the E-Model results. From the distributions of all the results, it is clear to see the trend of MOS from PESQ analysis is generally lower than E-Model score while it gets higher than E-Model score as packet loss rate increases. The threshold

TABLE II
MEAN/MEDIAN VALUE OF STANDARD DEVIATION OF CODEC PERFORMANCE OF PESQ FOR DIFFERENT LANGUAGES

Language	ITU-T G.729	ITU-T G.711μ	ITU-T G.711a
All	0.187/0.180	0.196/0.200	0.228/0.227
American English	0.214/0.188	0.210/0.213	0.232/0.226
French	0.168/0.161	0.168/0.166	0.200/0.187
Chinese	0.157/0.172	0.180/0.158	0.240/0.224
Japanese	0.209/0.197	0.228/0.264	0.238/0.270

here is around 5% packet loss rate.

Each evaluated MOS from the PESQ tool has a deviation from the E-Model result. In Figures 3, 4 & 5, mean and median values for a particular packet loss range are calculated respectively. The range is grouped from 0% - 0.25%, 0.25% - 0.75%, 0.75% - 1.25%, 1.25% - 1.75% and so on. To keep most accurate distribution, the average loss rates are calculated within each loss range. However, there is a difference in these two value sets. Mean value gives an average score for all the MOS values in its range, including outliers that have extreme low/high values resulting from packet loss being concentrated in voice activity regions. While on the other hand, median value gives a more comprehensive value representing the average value in a set, but also is able to eliminate or ease the influence caused by those outliers, at least at some extent.

The difference between highest and lowest MOS from PESQ results can be 1.0 or more. The variance of the scores is not only caused by different input samples but also because of the random distribution of packet loss. Comparing to a video stream where key frame loss could cause huge quality degradation, audio streams have speech and silence periods where packets for the speech periods are obviously much more important than the ones for silence period due to the conversational voice information they carry. It is reasonable that G.729 codec performs worse than G.711 μ-law and G.711 a-law. This is because that G.711 μ/a-law operate at 64 Kbps with no compression while G.729 is more suitable for mobile networks at 8 Kbps but still stays competitive with comparable quality to G.711. It is worth mentioning that lower rate codecs tend to degrade more rapidly in the presence of loss that higher rate codecs do; this is primarily due to the inter-packet dependencies that these codecs use to achieve their lower bit rates.

Table II illustrates the difference between PESQ scores for each test that is distributed within each of the loss ranges. There are standard deviations calculated specifically for each loss range, and Table II shows the mean and median values of these values. The overall analysis for tests including all languages as shown in first row "All". The analysis combines variances for all loss ranges so it it a indicator of the stability of a certain codec. It indicates that G.729 has better stability than G.711μ-law and G.711a-law with respect to PESQ MOS results. However, differences between results obtained for each of the languages are not sufficiently statistically significant to allow any hard conclusions to be drawn, for example that a specific language has lower deviation than others.

TABLE III
PEARSON CORRELATION

Pearson Correlation	Range
high correlation	0.5 to 1.0 or -0.5 to 1.0
medium correlation	0.3 to 0.5 or -0.3 to 0.5
low correlation	0.1 to 0.3 or -0.1 to -0.3

TABLE IV
PEARSON CORRELATION OF PESQ AND E-MODEL

Language	G.729	G.711μ-law	G.711a-law
All	0.726	0.727	0.637
American English	0.738	0.798	0.591
French	0.810	0.722	0.648
Chinese	0.811	0.802	0.734
Japanese	0.779	0.453	0.761
All	0.726	0.727	0.637

Pearson Correlation is a helpful statistic tool that calculates the correlation between variables measuring how well they are related. Its possible results are between -1 and 1, and -1 means a perfect negative correlation between the two variables while 1 means perfect positive correlation and 0 means no linear relationship between them. The closer the value gets to -1 or 1, the more correlated the two set of variables are. The relation of the value and level of correlation is shown in Table III.

The Pearson Correlation Coefficient of PESQ and E-Model scores are shown in Table IV. Each single test MOS result from PESQ is compared with E-Model result grouped by language and codecs. We see that the overall correlations for 3 codecs are around 0.7 which indicates even though the correlation is not perfect, there is still a high correlation between the results of PESQ and E-Model. There is no direct evidence showing the correlation is language dependent.

From Table II and IV, we can see that even though PESQ and E-Model is language dependent respectively at some extent, the deviation of PESQ and the correlation between PESQ and E-Model is independent from language. However, from Figures 3,4&5, PESQ is highly dependent on the speech content and the effect of packet loss. The results of the two call quality assessments can be highly correlated when large number of PESQ analysis is done. Thus, it is possible to perform extensive set of PESQ analysis experiment using the proposed framework and the mean / median values of the results can be good indication of the call quality in certain network environments.

VI. CONCLUSION & FUTURE WORK

In this paper we investigated two widely used call quality assessment methods by performing a set of experiments using a commercial VoIP product. The limitations and advantages of each one of the methods are discussed and compared. Due to the high cost of subjective testing, PESQ and E-Model provide the possibility of large scale automation call quality testing. Rather than taking only one of the methods, taking advantage of both off-line intrusive PESQ analysis and on-line non-intrusive E-Model can give a much reasonable quality indicator that will play an important role in product development and

testing. This paper analyses the accuracy of PESQ and E-Model and shows the correlation between both methods. Even though each of the methods is not a perfect indicator of exact real call quality and each has its own limitations, the overall results of the two correlate well in terms of indicating the call quality for a certain period of time during the call. This gives us confidence when performing quality tests using the two assessment methods.

Furthermore, due to the analysis of variance of PESQ results from E-Model, it is highly recommended that large number of test samples with different contents and preferably different languages from both male and female are needed for PESQ analysis. Only the trend of large distribution of results can represent the call quality regardless of the effect of sample contents. E-model is also a useful tool that can tell a possible call quality it could be with certain codec and network conditions.

According to the feature of the E-Model, continuous monitoring of on-going call quality can enable more intelligent codec selection and enable quality based codec adaptation to maintain higher voice call quality. For future work, the experiment could be enhanced to enable support for multi-party voice calls which can adapt to end users having different network conditions.

The support for wideband codecs from both approaches are in progress for next phase of research. For off-line assessment, Perceptual Objective Listening Quality Analysis (POLQA) is considered as next generation voice quality testing technology. It has been standardized by ITU-T as new Recommendation P.863, and is suitable for voice quality analysis under mobile networks, such as 3G and 4G/LTE networks.

ACKNOWLEDGMENT

This research was partially funded by Science Foundation Ireland (SFI) via grant 08/SRC/I1403 FAME SRC (Federated, Autonomic Management of End-to-End Communications Services Strategic Research Cluster). This work was also supported by the Irish Research Council for Science, Engineering and Technology, co-funded by Marie Curie Actions under FP7; and by the Telecommunications Graduate Initiative (TGI) which is funded by the Higher Education Authority under the Programme for Research in Third-Level Institutions (PRTLI) Cycle 5 and co-funded under the European Regional Development Fund (ERDF).

REFERENCES

[1] ITU-T Recommendation 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", 2001.

[2] iLocus. "The 10 that Established VoIP (Part 1: VocalTec)" Internet: http://www.ilocus.com/2007/07/the_10_that_established_voip_p_2.html, July. 16, 2007 [December, 29, 2012].

[3] ITU-T Recommendation P.835, "Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm", 2003.

[4] M. Carbone and L. Rizzo, Dummynet revisited, ACM SIGCOMM Computer Communication Review, vol. 40, no. 2, pp. 12-20, 2010.

[5] Carvalho, Leandro, et al. "An E-model implementation for speech quality evaluation in VoIP systems." Computers and Communications, ISCC 2005. Proceedings. 10th IEEE Symposium on. IEEE.

[6] S. Pennock, Accuracy of the perceptual evaluation of speech quality (pesq) algorithm, Proc. Of MESAQIN, 2002.

[7] B. Huntgeburth, S. Schumann, and J. Londak, Voice over ip (voip) speech quality measurement with open-source software components, in ELMAR, 2010 PROCEEDINGS. IEEE, pp. 215-218.

[8] Rix, Antony W. "Comparison between subjective listening quality and P. 862 PESQ score." Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN03), Prague, Czech Republic, 2003.

[9] S. Moller, W.-Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Waltermann, Speech quality estimation: Models and trends, Signal Processing Magazine, IEEE, vol. 28, no. 6, pp. 18-28, 2011.

[10] L. Sun, I.-H. Mkwawa, E. Jammeh, and E. Ifeachor, Voip quality of experience (QoE), in Guide to Voice and Video over IP. Springer, 2013, pp. 123-162.

[11] De Rango, Floriano, et al. "Overview on VoIP: Subjective and Objective Measurement Methods." International Journal of Computer Science and Network Security 6.1, 2006: 140-153.

[12] ITU-T Recommendation G.107, "The E-model, a computational model for use in transmission planning", Mar. 2005;

[13] ITU-T Recommendation G.107.1, "Wideband E-model", Dec. 2011;

[14] A. Takahashi, A. Kurashima, and H. Yoshino, Objective assessment methodology for estimating conversational quality in voip, Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, no. 6, pp. 1984-1993, 2006.