Improved E-model for Monitoring Quality of Multi-Party VoIP communications

Mohamed Adel[†], Haytham Assem[‡], Brendan Jennings[†], David Malone[*], Jonathan Dunne[‡] and Pat O'Sullivan[‡]
[†]TSSG, Waterford Institute of Technology, Ireland
Email: madel@tssg.org, bjennings@ieee.org
[*]Hamilton Institute, National University of Ireland Maynooth, Ireland
Email: david.malone@nuim.ie
[‡] IBM Software Lab, Dublin, Ireland
Email: {haythama, jonathan_dunne, patosullivan}@ie.ibm.com

***Abstract*—Maintaining good Quality-of-Experience (QoE) is crucial for Voice-over-IP (VoIP) applications, particularly those operating across the public Internet. Accurate online estimation of QoE as perceived by end users allows VoIP applications take steps to improve QoE when it falls below acceptable levels. ITU-T recommendation G.107 introduced the E-model, which provides a means to assess QoE levels for two-party VoIP sessions. In this paper we provide an analysis of the accuracy of the E-model for multi-party VoIP sessions when all audio is processed by a centralised focus node. We analyse the impact of what we term the "Focus Transcoding Effect (FTE)," the "Focus Forwarding Effect (FFE)," and the number of end-points participating in the session. Through comparison to QoE metrics produced by the offline PESQ method for three common audio codecs, we show that the standard E-model does not provide accurate QoE assessment for multi-party VoIP sessions. We then introduce an improved E-model for these codecs for multi-party VoIP sessions. We describe the implementation of the improved E-model in a QoE monitoring application, showing that it produces results similar to actual PESQ scores.**

Keywords*—*VoIP, QoE, PESQ, E-model, Multi-party.

I. Introduction

Voice-over-IP (VoIP) performance depends on a number of network-related factors, including available bandwidth, end-to-end delay, packet loss and jitter. Variance in these parameters often leads to degradation of VoIP performance and the Quality-of-Experience (QoE) perceived by end users. Moreover, other than network issues, application specific factors like the choice of codec, codec parameters, and jitter buffer sizing also impact QoE. It is important for implementers of VoIP applications to assess QoE as perceived by the end user and take mitigating actions when it degrades to unacceptable levels. Mean Opinion Score (MOS) is the commonly accepted metric to measure the QoE of a call as perceived directly by the end user—it encapsulates the effects of both network and implementation specific issues.

In recent years VoIP has become an extremely important application class, with VoIP clients being very widely used by businesses and individuals. The success achieved by the basic two-party VoIP communications in terms of reliability and the cutting of costs has encouraged the emergence of multi-party VoIP conferencing facilities. Intuitively, it is more difficult to ensure QoE in multi-party sessions since, at different times during a sessions different people, connecting via different network paths, will be speaking. In this paper we examine whether the E-model for online QoE estimation model which was developed for two party VoIP sessions are applicable to multi-party sessions. We find that it is not—for three commonly used audio codecs we find that it consistently over-estimates MOS values for a range of network-path packet loss conditions. We specify an enhanced E-model, describes its realisation in an online VoIP QoE monitoring tool and show results that indicate that it provides a more accurate QoE estimation.

The paper is structured as follows. §II introduces the most commonly applied QoE metrics and provides a brief survey of related work. §III introduces the main architectures used in multi-party VoIP systems. §IV analyzes the QoE of multi-party VoIP, identifying and analysing the main effects that lead to QoE degradation in comparison to two-party calls. §V specifies our corrected E-model for multi-party calls, whist §VI describes its realisation in a call quality monitoring system. Finally, §VII concludes the paper and outlines areas for further work.

II. Related Work

As measuring voice quality is important to the service providers and end users, ITU-T provides two test methods subjective and objective testing. Subjective testing represents the earliest attempts on this issue to evaluate the speech quality by giving Mean Opinion Scores (MOS). ITU-T Rec. P.800 [1] presents the MOS test procedures as users can rate the speech quality from 1 (Poor) to 5 (Excellent) scale. Of course, the numbers of the listeners are an important factor in estimating accurate scores. Thus, subjective testing using MOS is time consuming, expensive and does not allow real time measurement. Consequently, in recent years, new methods were developed for measuring MOS scores in an objective way (without human perception): notably PESQ [2] and the E-model [3].

PESQ, Perceptual Evaluation of Speech Quality, is an *intrusive* testing method which takes into account two audio signals: one is the reference signal while the other one is the actual degraded signal. Both signals are sent through the PESQ algorithm and the result is a PESQ score. Given that the full signals are required in advance, this approach cannot be used to monitor real time calls.

978-1-4799-2851-4/13/$31.00 ©2013IEEE

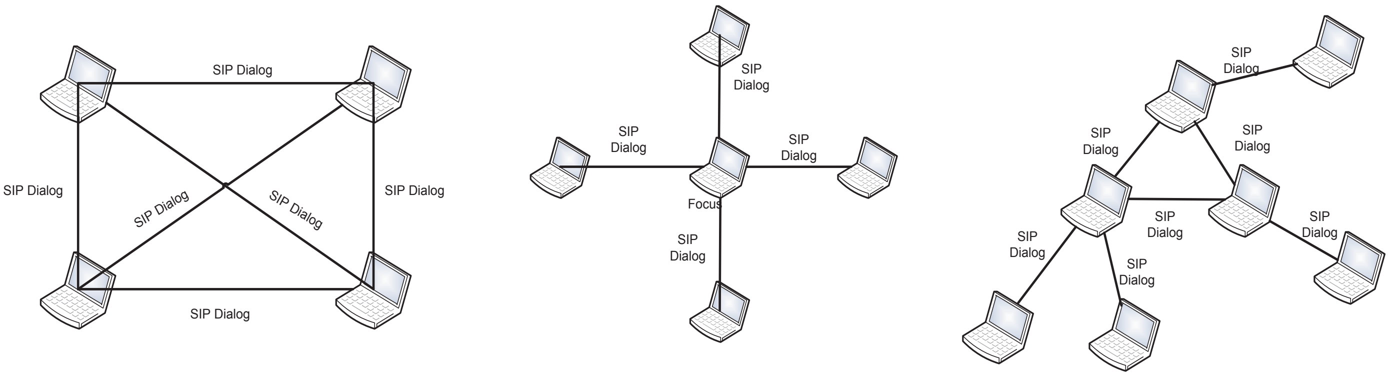

Fig. 1. Decentralized, centralized, and hybrid models for multi-party VoIP sessions. In the decentralised model there is no focus; in the centralised model this exactly one focus, which may itself be a participant in the session; in the hybrid model more than one node may act as the focus for a subset of the other nodes.

On the other hand the E-model is a *non-intrusive* testing method that can be applied in real-time. It is a mathematical model that combines all the impairment factors that affect the voice quality in a single metric called the R value that can then be mapped to the MOS scale (Table I summarises this mapping). R values are in the range 0 to 100, where $R = 0$ represents the worst quality and $R = 100$ represents the best quality. R is calculated as:

$$R = R_0 - I_s - I_d - I_{e,eff} + A \qquad (1)$$

where R_0 is the Signal to Noise ratio (S/N) at 0 dBR point, I_s represents the speech voice impairments, I_d is the impairments occurred due to the delay, $I_{e,eff}$ is the impairment due to the equipment (e.g.: codecs and packet loss) and A is the advantage factor (e.g.: $A = 0$ for wireline). As outlined in [4], [5], [6] the E-model can be utilized to be used in the speech quality evaluation over VoIP-Based Communication Systems. However, E-Model is only valid for ITU Codecs, in [7] we derived E-Model for SILK and iLBC which are widely used non-ITU codecs.

Corrected versions of the E-model have been proposed to simplify the calculations and focus on the most important factors required for monitoring the call quality [8]. Paulsen et al. [9] introduced a new parametrized QoS measurement method for VoIP applications. Their proposed improved E-model use the “glass box” principle. They took into account typical IP-Environmental parameters whilst the original E-model is designed for circuit-switching networks and can not really take such factors into account. They compared their results to the PESQ and the original E-model results and they show higher accuracy in measuring the MOS compared to the original E-model. Ren et al. [10] studied how the jitter affects the VoIP quality and how to model such effect into the E-model. They used the PESQ algorithm to measure such effect and as a result, they introduced a new Ij formula which is added to the original E-model representing jitter impairment factor. Zhang et al. [11] came afterwards to use the prior extended E-model in order to compare the performances of the original and extended E-model (including the jitter impairment factor) by applying them both on different VoIP systems (Skype, Google Talk and Windows live messenger). They concluded that Windows live messenger outperforms in terms of listening, Skype has the largest MOS, and Google Talk generally has the least MOS. Obafemi et al. [12] studied the E-model with a focus on the effect of the ignorant parameter jitter playout buffer on the accuracy of the call quality resulted from the E-model. Their results shows that the adaptive play out buffer should not be ignored when evaluating the perceived call quality. They suggest modifying the original E-model to include measurements of an adaptive playout buffering. Zhang et al. [13] proposes a new algorithm to measure the packet loss burstiness to be included in the E-model as a replacement of the random probability of the packet loss to calculate the MOS value. They show that their improved E-model have a higher accuracy under bursty packet loss conditions.

TABLE I. RELATIONSHIP BETWEEN R AND MEAN OPINION SCORE.

R	**Satisfaction Level**	**MOS**
90-100	Very satisfied	4.3+
80-90	Satisfied	4.0-4.3
70-80	Some users dissatisfied	3.6-4.0
60-70	Many users dissatisfied	3.1-3.6
50-60	Nearly all users dissatisfied	2.6-3.1
0-50	Not recommended	1.0-2.6

The main difference between our work and those reviewed above that we noticed that the accuracy of the E-model has not been tested before in the VoIP conferencing system. There is surprisingly limited work in the multi-party QoE area, compared to the wide literature on QoE for person-to-person calls. Thus, in this paper we perform a detailed analysis for the accuracy of monitoring the call quality using the original E-model. Based on our analysis, we proposed a correction for the current original E-model in order to be used in the VoIP conferencing systems.

III. MULTI-PARTY VOIP

The Session Initiation Protocol (SIP) [14], used by the majority of VoIP applications, supports establishment of communications sessions with multiple participants. Nevertheless, VoIP applications have considerable flexibility in how VoIP sessions are realised. Currently, VoIP conferences are implemented through three possible connection topologies [15]: Decentralized, Centralized and Hybrid. These models as shown in Figure 1; they are described as follows:

- *Decentralized Model*

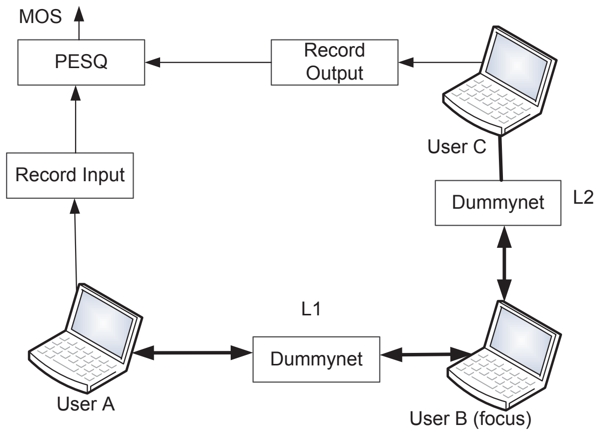

Fig. 2. Centralized Multi-party VoIP Setup.

In the decentralized model all conference endpoints are connected to each other via unicasts or multicasts. Each endpoint interacts with the rest of the endpoints using SIP. There is no focal point or centre for the conference, so the flow of the data is distributed among all the clients;

- *Centralized Model*
  A centralized model is based on a central point of control called a *focus*. The focus can be a dedicated Media Server (such as that as used in IBM Sametime Unified Telephony [16]), or one of the conference endpoints can perform this task (as used in Skype [17] and Jitsi [18]). The focus is typically responsible for SIP signaling between all the conference endpoints. Moreover, all the transmitted audio data in the conference call must pass first through the focus to be decoded, mixed (if more than one user is speaking) and finally re-encoded and sent to the rest of endpoints;
- *Hybrid Model*
  A hybrid model is based on a combination of centralized and decentralized architectures. It relies on an underlying overlay network, where some of the nodes act as a parent nodes which are fully connected to each other; others are child nodes which are only connected to one parent node.

We focus on the Centralized model since it is common in designing VoIP multi-party conferencing systems. Each endpoint is connected directly to the focus, and it has no current knowledge of other connections between other endpoints and the focus. Multiple links to the focus are often subjected to different degradation factors.

IV. QoE Analysis of Multi-Party Calls

In this section, we describe an analysis of QoE of multi-party VoIP calls initiated using a centralized multi-party VoIP application. We estimate MOS scores using both PESQ and the E-model. PESQ is an intrusive method, requiring both the original and the degraded signal, so we take it as a benchmark for the E-model estimates as it should achieve a high degree of estimation accuracy. In our analysis, we study the performance of three commonly used codecs: G711, SILK and ILBC, under different network conditions. Figure 2 shows the testbed used in our experiments; we establish different VoIP multi-party calls between three users. In the figure the labels L1 and L2 indicate the links between user B, which acts as the central focus, with user A and user C respectively. We use Dummynet [19] to emulate different packet loss rates in L1 and L2 in the range from 0-5%, with 0.5% increments. In our analysis and to unify our comparison's parameters, we consider user A (speaking), user B (focus) and user C (listening).

We record the original and degraded audio signals from user A and user C respectively; these signals were then used as input to the PESQ algorithm, which produces MOS values in the range 1-5. In order to have accurate measurements and scores, we have taken more than 200 PESQ MOS values under different network condition for each codec. We also developed an online monitoring module that employs the E-model to estimate the MOS score but we excluded the delay factor from our calculation since the PESQ does not take it into account when estimating MOS. For the E-model, we do include the packet loss rate from both links when we repeat such experiment for the three codecs using Jitsi as a VoIP application.

From Figure 3 we see that the PESQ score under different packet loss rate of L1 and L2 with focus transcoding differs from the PESQ score of the call without focus transcoding. Also these PESQ scores differ compared to the PESQ score that is resulted from a single link with the sum of the packet loss rate of the two links. For instance, the PESQ score of the conference call between A and C passing by focus B having packet loss rate of L1 and L2 equals to 1% and 3% respectively differs when compared to a single link between 2 users having a packet loss rate of 4%. Specifically, using the G.711 codec, the PESQ MOS score was 2.526 when having the 2 links while it was 2.81 when having single link. When using ILBC codec, the PESQ score was 1.82 using the 2 links while it was 2.37 when having single link with the sum of the packet loss rate of L1 and L2. We also observe that changing the order of introducing packet loss to conference links produces very similar results; in other words, adding 1% packet loss rate to L1 and 3% to L2 would give almost the same result as 3% to L1 and 1% to L2.

These observations leads us to study the effect of the presence of the central focus in order to be able to model its effect and to develop a corrected E-model that can be used in monitoring the call quality of multi-party calls. We address the following three effects of introducing a central focus: the *Focus Transcoding Effect*, the *Focus Forwarding Effect*, and the *Number of Users*.

A. Focus Transcoding Effect (FTE)

In the centralized model, all of packets are forwarded to the node that acts as a central focus. In order to understand the signal, this node decodes the packets back in to an audio signal. Then this signal is re-encoded and forwarded to the rest of the users in the conference call after re-negotiation of the used codec. The process of the decoding/re-encoding of the packet is called the transcoding process. This process has an influence on the QoE perceived the end user. We have studied the FTE

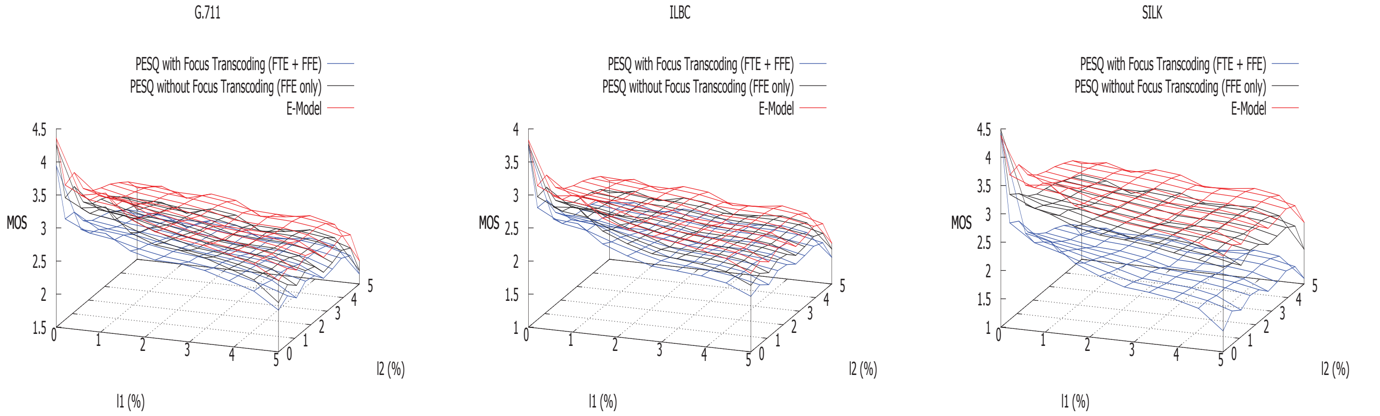

Fig. 3. QoE for Multi-party call for G.711, ILBC, and SILK. The x and y axis indicates the percentage of the packet loss of the links L1 and L2 respectively, whilst the z axis indicates the MOS score.

effect by measuring the PESQ MOS score of a conference call with the setup shown in Figure 2 using the three different codecs under different packet loss rates. The resulting PESQ scores are shown in Figure 3 for the G711, ILBC and SILK codecs respectively labelled as PESQ with transcoding effect.

B. *Focus Forwarding Effect (FFE)*

We have studied the forwarding process of the packets from the focus to the rest of users and its influence on the QoE at the end user. In order to study such effect only, a typical peer-to-peer call is established between A and C, which are connected to Internet through the same gateway. We have used the testbed shown in Fig 2 by adding another node between user A and user C so that all of the packets forwarded from A to C are forced to pass by a gateway node first. This process emulates the forwarding effect only without the transcoding effect. We have measured the PESQ scores of different established calls under different packet loss rates using 3 different codec. These can be shown in Figure 3 for G711, ILBC and SILK codecs respectively, labelled as PESQ without focus transcoding.

C. *Number of Users Effect*

In order to study the impact of the number of users in the call we start a conference call with 3 users, and we increase the number of users from 3 to 6, adding a single user each time. At each time of increasing one more user, we have measured the PESQ MOS score at a certain user. First, the call was initiated with user A, user B, and user C using G.711 codec where user B is acting as the central focus in the call. In our experiment, we measure and track the call quality of user A using PESQ algorithm, so user A is considered as a speaker whilst we consider user C is the listener. We ensured that there is no network losses when measuring the call quality at user A. When 3 users were participating in the call, the MOS PESQ was 3.98. We found that this score is constant when increasing the users every time from 3 to 6 users. This shows that, at least for a reasonably small number of users, the number of users in the centralized model of the multi-party call has no appreciable effect on the end user perceived call quality.

D. *Accuracy of the E-model*

In order to study the accuracy of the original E-model in assessing QoE of multi-party audio calls, we have employed the original E-model in our monitoring system at the end user C, using the same testbed shown in Figure 2. The measured MOS resulted from the E-model is shown in Figure 3 for the G711, ILBC and SILK codecs respectively labelled as E-model. We clearly see that the standard E-model consistently overestimates call quality. It is therefore clear that the E-model needs to be correct to take the FTE and FFE into account. Moreover, we have noticed that such gap between the PESQ score with FFE and FEE with the E-model is codec dependent. Thus, codec dependent coefficients need to be derived for such correction of the E-model.

V. Corrected E-Model for Multi-Party Calls

In this section, we derive a correction function to the ITU standard E-model in order to make it suitable for evaluating QoE of multi-party calls. In Figure 4, we mapped the values of the original E-Model (x-axis) to the actual quality estimated by PESQ algorithm (y-axis). For example for SILK, at 4% packet loss, the audio MOS value estimated by E-model is equal to 2.87, while the actual MOS perceived as calculated by PESQ algorithm should be equal to 1.73. By applying curve fitting by using the least squares method, the points fit well with a third degree function MOS_C. It indicates the actual QoE in a multi-party session as perceived by end-users, considering the focus degradation factors FTE and FFE. It is derived for each of the three codecs with parameters x1, x2, x3 and x4 as shown in Table II. In Equation 2, MOS is the standard E-model function as explained.

$$MOS_C = x1 \cdot MOS^3 + x2 \cdot MOS^2 + x3 \cdot MOS + x4 \quad (2)$$

In order to estimate QoE online, the corrected E-model can be employed by first capturing the network characteristics (packet loss rate is the sum of loss rates of L1 and L2), acquiring the codec robustness factor then calculating the R which is then

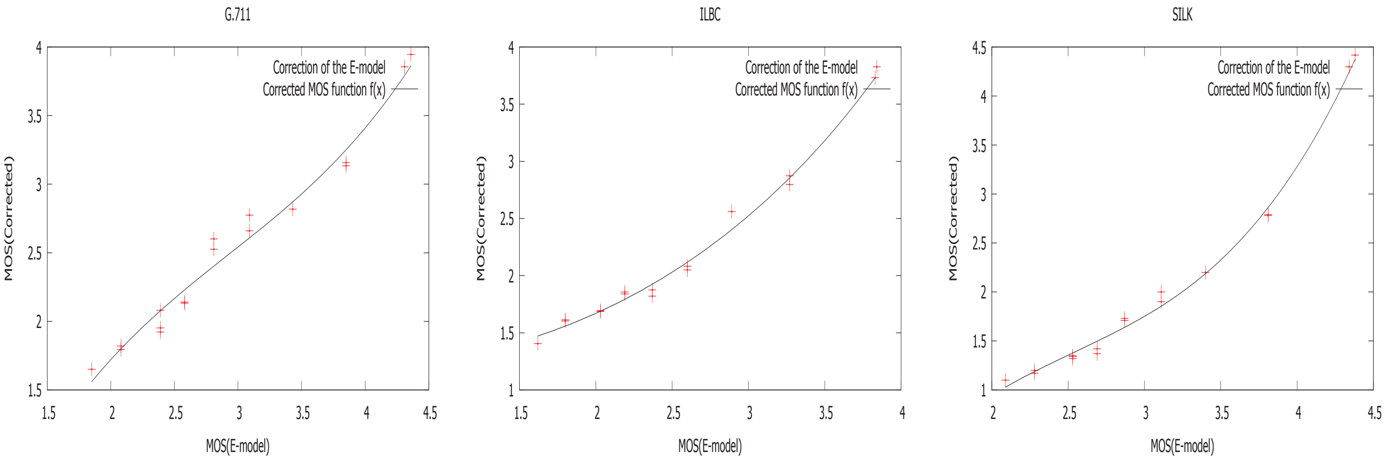

Fig. 4. Correction Function for G.711, ILBC, and SILK

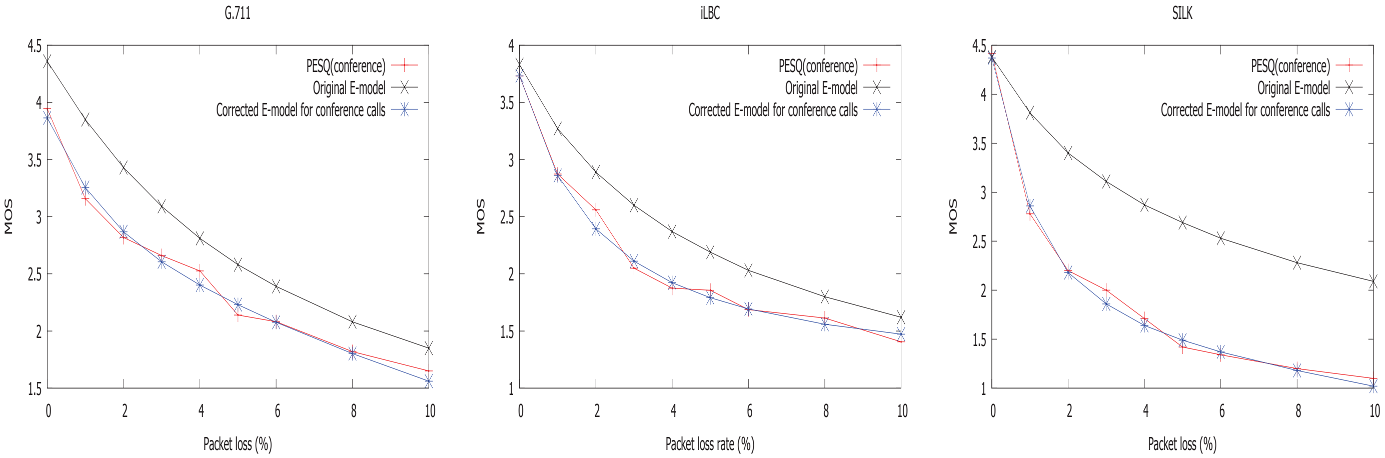

Fig. 5. Corrected E-model against Packet Loss Rate for G.711, ILBC, and SILK

mapped to MOS. This MOS value resulted from the standard E-model is then used in Equation 2 to calculate MOS_C, the estimated MOS perceived by the end-users of multi-party VoIP session.

VI. Monitoring System Design and Results

We have developed a monitoring system based on our corrected improved E-model for monitoring the VoIP call quality for the multi-party calls. Our monitoring system targets specific number of RTP packets to capture and perform an effective MOS value calculation based on our corrected E-model. Our system uses a coefficient database according to the codec used in the call, see Table II. It is based at the network terminals, and the environment could be a personal or family network with voice quality monitoring. Our monitoring system works as follows. First, the system uses a network capturing module to capture a certain number of packets to certain IP and port. The non-RTP packets will be filtered. After this process is finished, the system will then starts to anlyze the data, delay and packet loss rate. Finally, the measured network conditions is converted into the MOS_C to indicate the call quality at the end user in the multi-party call. We took our results online by introducing random packer loss rates in the network in the range from 0-10% using Dummynet. For comparisons our system also computes MOS values using PESQ and the standard E-model.

TABLE II. Derived 3rd Degree Parameters for Different Codecs for Multi-party calls.

Parameters	G.711	ILBC	SILK
x1	0.111	0.045	0.26
x2	-0.978	-0.068	-1.982
x3	3.597	0.326	5.769
x4	-2.451	0.929	-4.748

We established conference calls using Jitsi, then applied the modified E-model under various packet loss rates; the results are shown in Figure 5. We have tested three codecs G.711, iLBC, and SILK. The correction made for the E-model has resulted in accurate results, very similar to what PESQ estimates. Crucially, our model can be used online to estimate QoE, unlike PESQ which has required to be performed offline by recording on both sides and then comparing both original and degraded signals.

VII. Conclusion and Future Work

ITU-Recommendation G.107 introduces the E-model which supports an approach to estimate the VoIP call quality of the person-to-person calls. The main advantage of the E-model is that it can be applied in real-time which enables monitoring call quality during the call. Due to the increased demand of the communications between more than one party in different locations, conferencing VoIP systems were introduced and became more mature. In this paper, we have studied the QoE of the VoIP conferencing systems that use the centralized model, where all audio is processed by a single focus node. Our results quantified the negative impact of this for three commonly used audio codecs. We identified two significant effects we termed the Focus Transcoding Effect (FTE) and the Focus Forwarding Effect (FFE). These effects are not taken into account in the standard E-model which will lead to estimating inaccurate multi-party call quality. Consequently, we have corrected the original E-model in order to allow it be used for online monitoring of multi-party call quality. We described how we derived the coefficients used for 3 commonly used codecs (G.711, ILBC and SILK) for our corrected E-model. We demonstrated its efficacy by implementing it in a monitoring system—which analyzes the impact of voice quality encoding factors under various network conditions and uses our corrected E-model to assess the multi-party voice call quality in real-time.

For future work, we intend to measure and quantify the degradation factors for video communications when using the centralized multi-party architecture with the common video codecs H.263 and H.264. Furthermore, we can extend our work to take certain decisions at the end-users' side based on the actual quality perceived in order to minimize the degradation effect caused by the focus. Codec switching could be a solution—re-negotiating a new codec at certain links could lead to minimizing the Focus Transcoding Effect and improving the QoE for end-user at that link.

Acknowledgment

The authors were supported by Science Foundation Ireland (SFI) grants 07/SK/I1216a and 08/SRC/I1403.

References

[1] I. Rec, "P. 800: Methods for subjective determination of transmission quality," *International Telecommunication Union*, 1996.

[2] ——, "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.

[3] ——, "G. 107," *The E-model, a computational model for use in transmission planning*, 2009.

[4] A. Clark, P. Iee, and et al., "Modeling the effects of burst packet loss and recency on subjective voice quality," 2001.

[5] R. Cole and J. Rosenbluth, "Voice over IP performance monitoring," *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 2, pp. 9–24, 2001.

[6] L. Lustosa, L. Carvalho, P. Rodrigues, and E. Mota, "E-model utilization for speech quality evaluation over VoIP-based communication systems," *Proc. 22nd SBRC*, 2004.

[7] H. Assem, M. Adel, B. Jennings, D. Malone, J. Dunne, and P. O'Sullivan, "A generic algorithm for mid-call audio codec switching," in *1st IFIP/IEEE International Workshop on QoE Centric Management (QCMan 2013)*.

[8] H. Assem, D. Malone, J. Dunne, and P. O'Sullivan, "Monitoring VoIP call quality using improved simplified E-model," in *International Conference on Computing, Networking and Communications(ICNC 2013)*, 2013, pp. 927–931.

[9] S. Paulsen and T. Uhl, "Adjustments for QoS of VoIP in the e-model," in *Telecommunications: The Infrastructure for the 21st Century (WTC), 2010*. VDE, 2010, pp. 1–6.

[10] J. Ren, C. Zhang, W. Huang, and D. Mao, "Enhancement to E-model on standard deviation of packet delay," in *Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference on*. IEEE, 2010, pp. 256–259.

[11] H. Zhang, Z. Gu, and Z. Tian, "QoS evaluation based on extend E-model in VoIP," in *Advanced Communication Technology (ICACT), 2011 13th International Conference on*. IEEE, 2011, pp. 852–854.

[12] O. Obafemi, T. Gyires, and Y. Tang, "An analytic and experimental study on the impact of jitter playout buffer on the E-model in VoIP quality measurement," in *ICN 2011, The Tenth International Conference on Networks*, 2011, pp. 151–156.

[13] H. Zhang, L. Xie, J. Byun, P. Flynn, and C. Shim, "Packet loss burstiness and enhancement to the E-model," in *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2005 and First ACIS International Workshop on Self-Assembling Wireless Networks. SNPD/SAWN 2005. Sixth International Conference on*. IEEE, 2005, pp. 214–219.

[14] J. Rosenberg, "RFC 4353A framework for conferencing with the session initiation protocol," 2006.

[15] B. Sat, Z. Huang, and B. Wah, "The design of a multi-party VoIP conferencing system over the internet," in *Multimedia, 2007. ISM 2007. Ninth IEEE International Symposium on*. IEEE, 2007, pp. 3–10.

[16] "IBM sametime unified telephony," 2013. [Online]. Available: http://www.ibm.com/

[17] "Skype," 2013. [Online]. Available: http://www.skype.com

[18] "JITSI," 2013. [Online]. Available: http://www.jitsi.org

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