Capturing User Behavior in Subjective Quality Assessment of OTT Video Service

Weiwei Li*, Petros Spachos §, Mark Chignell*, Alberto Leon-Garcia*, Jie Jiang ‡, Leon Zucherman*
*University of Toronto, Toronto, ON, Canada
§School of Engineering, University of Guelph, Guelph, ON, Canada
‡Technology Strategy and Operations, TELUS communications Company, Toronto, Canada

***Abstract*—Customer satisfaction is an important factor governing adoption and retention of multimedia products and services, such as Over-The-Top(OTT) video transmission. Quality of Experience involves user-centric evaluation of various services. However, users differ in terms of their ratings of service quality. Some rating differences are due to unreliability (outlier users who are not motivated, or are not sensitive to differences in quality), but others are systematic differences in rating that may reflect different perspectives on quality. In this paper, we explore the use of outlier analysis and clustering as tools for interpreting QoE data. We report on experimental results demonstrating the use of outlier analysis and clustering. In interpreting the clusters, we examine users' opinions on different types of video disruption, and their ability to distinguish the different levels of impairments/ failures.**

I. Introduction

Mobile video service has grown explosively in the past two decades to a point where it requires a huge amount of network traffic. Service providers are struggling to maintain customer's satisfaction in the face of limited wireless spectrum, and measurement of the importance of Quality of Experience (QoE) for OTT video service evaluation has become increasingly important.

Traditional Quality of Service(QoS) focuses on the measurement of objective parameters measurement, such as packet loss, jitter, and delay. QoE proposes a user-centric evaluation for various services. However, due to the complexity of subjective parameters, it is hard to find a unique metric to assess QoE. Previous studies have proposed a number of possible QoE metrics, either measured directly or inferred indirectly [1], [2], [3]. The most common direct index involves user ratings that evaluate criteria such as content, technical quality, and satisfaction. Measuring a user's viewing time for a video is an example of an indirect measure. Given a particular QoE metric, the mean value of that metric is used to represent the level of QoE assessment.

We have found that users differ in term of how they perceive video quality in subjective video experiments[4]. Participants different in terms of their video technical quality (TQ) personality. Some participants are sensitive to interruptions, (e.g.,due to rebuffering), while other participants are tolerant of interruptions as long as video plays all the way through. However, other participants appear relatively disinterested in experiment and answer questions randomly or in ways that seem counterintuitive. These participants are generally statistical outliers and can be labeled as unreliable.

In this paper, Principal Component Analysis (PCA) of video disruptions ratings is used to identify outliers. After removing the statistical outliers, clustering of the PCA components is then used to identify different patterns of response to impairment/ failure types. Based on the findings presented below, we propose that this methodology can be used to detect abnormal user behaviors and to develop QoE personality models.

The rest of this paper is organized as follows: Section II reviews related works on QoE assessment. Section III briefly introduces the concept of QoE and the impairment/ failure types used in our subjective experiment. Section IV presents the analysis related to user behavior detection and clustering. Section V concludes with a discussion of the implications of this work.

II. Related Work

In [5], P. Reichl et al. pointed out that user behavior and user characteristics should be taken into consideration in QoE modeling. Personalized QoE models are proposed in [6] based on the user context.

PCA is widely used for QoE related analysis. In [7], the authors used Principal Components (PCs) on subjective parameters (such as content and sound quality,) to model QoE components by QoS parameters. In [8], [9], PCA was employed to find the impact of various QoS parameters. Videos were clustered based on content type, followed by PCA.

In QoE subjective measurement, mean ratings that deviate significantly from the average are identified as outliers.[10] A standard outlier screening methodology is proposed by the ITU, which checks the magnitude of the Kurtosis and standard deviation [11]. This general methodology for subjective QoE can be supplemented with techniques such as the k-Nearest Neighbor (kNN) outlier detection method proposed in [12]. Another potential approach is a recommendation system. In [13], the authors proposed a recommender system using PCA to filter the data, while in [7] matrix factorization models for recommendation were discussed, including PCA.

On the other hand, QoE research needs use behavior analysis. The appearance of Crowdsourcing enables collect experimental subjective evaluation over the Internet [14], [15]. However, the reliability of subjects is doubted. Doing QoE

evaluation through Internet, subjects might not pay attention on the video or click the answer randomly. Multiple algorithms and methodologies are proposed to screening unreliable user in [15].

In this paper, we propose the usage of PCA to distinguish user behavior. As shown in the following discussion, PCA is a promising method for distinguishing between QoE rating reliability and QoE rating personality.

III. Experiment description

A. Session-based QoE and QoS

Our experiment collected subjective evaluation ratings of OTT video service. Previous research proposed that the whole life-cycle of a video session was involved in QoE assessment[16]. The concept of a life-cycle contains three parts:

1) request a video by user,
2) wait for the video to start, watched the video along with some possible impairments, and
3) quit the video service normally or abnormally.

When we discuss the life-cycle, there are two different issues impacting viewing experience: Integrity impairment, and failure. Integrity reflects the degree to which the video session unfolds without impairment [11]. Integrity impairments are well studied in previous research and their relationship to QoE has been extensively discussed [17], [18]. However, the issue of failures, where the video didn't start, or didn't play through to the end, hasn't been sufficiently addressed in QoE evaluation. In our previous research, we have found that failures have a strong impact on user satisfaction [19], [20], [21]. We have defined two types of failure (Accessibility and Retainability), to account for QoE throughout the session. Accessibility is concerned with whether the video starts to play or not, and Retainability is concerned with whether or not the session continues (with/ without impairments) until there is a normal end (the video completes or the user terminates the playback)[22].

Based on the life-cycle of video service, we have designed eight types of impairment/ failure:

- Pristine: $I0$.
- Impairment: $I1$, $I2$, $I3$.
- Retainability failure: $R0\_70$, $R2\_50$, $R1\_30$.
- Accessibility failure: $A10$

$I0$ indicates that the participant can watch the whole video without any interruption. $I1$, $I2$ and $I3$ indicates that the video encounters 1, 2 or 3 instances (respectively) of temporary interruption during display ($10s$/ interruption). However, the participant is able to view the whole video. $R0\_70/R2\_50/R1\_30$ indicates that the participant encounters 0, 2 or 1 instances of temporary interruption and only the first 70, 50 or 30 seconds (respectively) of the content is displayed.

B. Dataset and data collection methodology

There were 60 participants in total. All participants were students from the University of Toronto and were over 18-years old. Each participant watched 31 short video clips. The average length of videos was $96.7s$. The viewing order of these videos was randomized. however, the numbers of videos with each impairment/ failure type were the same for each participant.

The Absolute Category Rating (ACR) method was employed for QoE assessment [23], and the rating scale followed the ITU standard recommendation, i.e., MOS [11]. Each video clip presented one type of the designed impairment/ failure. Four questions were asked after each video, relating to Technical Quality (TQ), Content Quality (CQ), Overall Experience (OX), and Acceptability. For more details regarding the questions and the experiment procedure please refer to our previous work [21]. However, the impairment/ failure types used in the present study were different from those used in that previous study.

IV. Subjects Behavior Analysis

In this work, we focus on user evaluations of TQ. The relationship between TQ and other QoE factors (CQ, OX, and Acceptability) is discussed in [21], [24]. In our data analysis, we used 5 to 1 to represent the scale of MOS, i.e. Excellent = 5, Good = 4, Fair= 3, Poor= 2, and Bad= 1.

A. User behavior analysis

Users can have various responses to the same impairment/ failure type due to multiple factors. We illustrate this through an example. Figure 1 shows the average ratings of impairment/ failure types of two participants, *P073* and *P074*. In the experiment, each participant viewed 6 samples of $I0$, 6 samples of a specific impairment type ($I1$, $I2$, and $I3$), 2 samples of each Retainability failure ($R2\_50$ and $R1\_30$), and 1 sample of $A10$. The average rating is the mean value of samples of one impairment/ failure type rated by one subject. As shown in Fig. 1, *P073* gave lower ratings for impairment types ($I1$-$I3$) compared to *P074*, but showed a higher tolerance (in terms of ratings) for failure types. On the other hand, *P074* seemed to care about the completeness of the video more (with lower ratings for the retainability failures). As long as the video had a failure, this subject evaluated TQ as *bad*, i.e. the lowest score in the scale.

One difficulty of mapping QoS parameters to subjective evaluation is that there is no gold standard for what the ratings should be. We can not say *P074* made a mistake because all failure types were rated as *bad*. Perception of video quality is inherently subjective. QoE is typically assessed using MOS, which provides a standard way to characterize the relationship between QoS and QoE. However, this approach disregards important variability in the data.

B. PCA on single impairment/ failure types

We examined whether use's average ratings on impairment/ failure types are similar or not. In our previous studies [20], [19], we found that most TQ scores of Accessibility failure were rated as 1 on the MOS scale. Therefore, the present study focused on Integrity impairments and Retainability failures. We generated a $m \times n_1$ matrix, $X1$. Each row represented

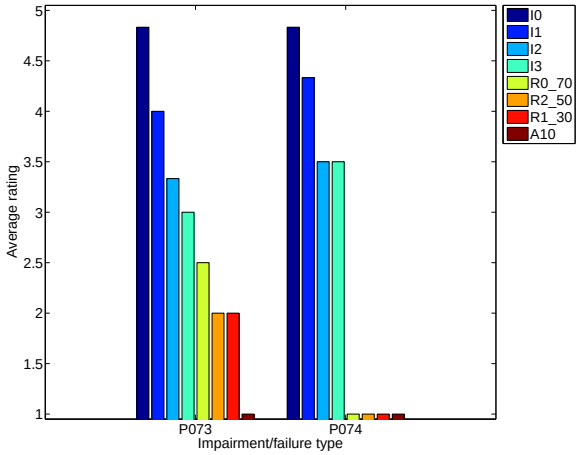


Fig. 1. Average ratings of two subjects

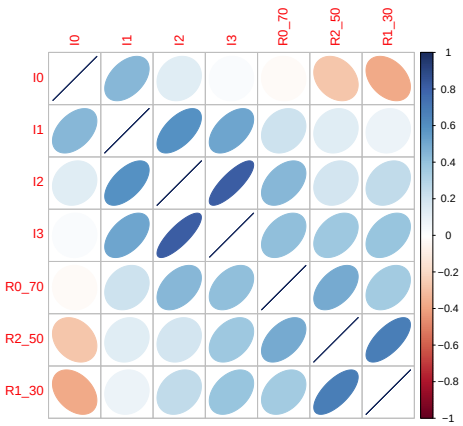


Fig. 2. Correlation matrix of $X1$

the average rating of each impairment/ failure type by one participant, where m was the number of participants. Each column represented the average ratings of a specific impairment/ failures, where n_1 is equal to the number of impairment/ failure types. Note that this analysis excluded the accessibility failure $A10$, i.e. $n_1 = 7$.

Figure 2 shows the correlation matrix based $X1$. These ellipses are shaped to be contours of corresponding Pearson correlation coefficients listed on the names of rows and columns. All ellipses are colored from dark blue (positive correlation) to dark red (negative correlation). If a ellipse leans towards the right, it indicates a positive correlation, and if a ellipse leans towards left, it is negative correlation. High positive correlations between neighboring impairment/ failure types are observed in Figure 2, such as $corr(I0, I1)$, $corr(I1, I2)$, $corr(I2, I3)$, and $corr(R1\_30, R2\_50)$. This indicates that user opinions on impairment/ failure types are correlated, especially within impairment types and within failure types. Weak negative correlation coefficients are found between $I0$ and $R2\_50/R1\_30$.

Applying PCA on $X1$, Table I shows the eigenvalues, the corresponding proportions, and the cumulative proportions of PCA. 2 PCs explain 69.91% of the total variance of the 7 types.

TABLE I
LOADING, PCA ON X1

Component	Eigenvalue	Proportion	Cumulative
1	3.035	0.434	0.434
2	1.859	0.266	0.699
3	0.727	0.104	0.803
4	0.624	0.089	0.892
5	0.330	0.047	0.939
6	0.280	0.040	0.979
7	0.146	0.021	1

Table II lists the first two PCs (accounting for almost 70% of the variance in the ratings), i.e, the two eigenvectors with the largest two eigenvalues. $I0$ had no significant contribution to 1^{st} PC, which appeared to represent general attitude towards video impairment/ failure. The 2^{nd} PC appeared to indicate different attitude towards the relative importance of completeness of video service versus impairments, since the corresponding eigenvector values of $I0-I3$ were negative and the values corresponding to failures were positive.

TABLE II
FIRST TWO PCS, PCA ON X1

	Component 1	Component 2
$I0$	0.012	-0.598
$I1$	-0.337	-0.457
$I2$	-0.462	-0.274
$I3$	-0.492	-0.136
$R0\_70$	-0.392	0.097
$R2\_50$	-0.379	0.392
$R1\_30$	-0.365	0.422

As discussed above, PCA can be used for outlier detection. In Fig. 3, three participants are located far from others: *P001*, *P023*, and *P062*, while Figure 4 shows the average ratings of these participants.

The score on the 1^{st} PC of *P001* is much higher than the corresponding scores for other participants, with this participant showing low tolerance to any video disruption. The average rating of $I0$ is above 4, while the ratings of the three other impairment types are between 1.5 and 2.5. All failure types got 1, the lowest score in the scale.

On the other hand, $P023$ had the lowest score on the 1^{st} PC. *P023* ratings on impairment types and $R0\_70$ were above 4.5, and the average ratings of $R2\_50$ and $R1\_30$ were around 2, still a high score for failure types.

The last selected participant, *P062*, had a high score on the 2^{nd} PC. The average ratings for *P062* showed little variation between disruption types, with the exception of $R0\_70$. *P062* showed evidence of being a random clicker and a possible outlier of a type discussed in [14].

C. Outlier Detection

Figure 4 indicates that the scores on two PCs can detect possible outliers among subjects. Therefore, we employ the kNN algorithm to detect outliers and then compare the

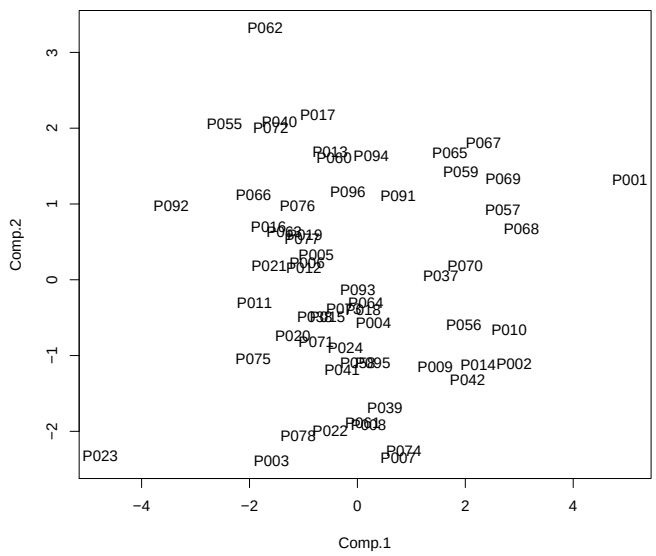


Fig. 3. PCA scores on 1^{st} and 2^{nd} PC of X1

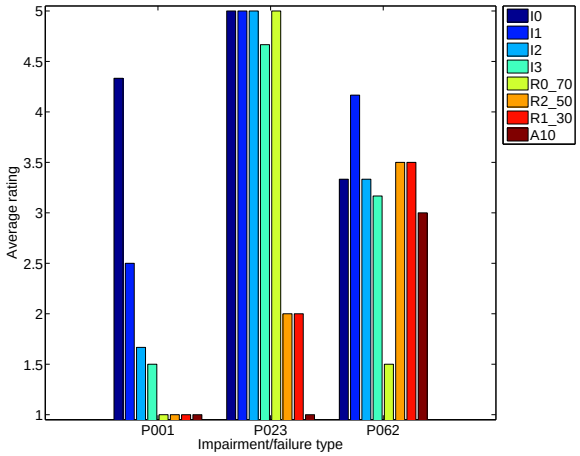


Fig. 4. Average ratings of selected subjects

corresponding MOS. As a distance-based outlier detection algorithm, we define the average distance of each participant to his/her k nearest neighbors as the outlier score for that participant. Two assumptions are the basis of the distance-based outlier detection:

- A normal subject should have a dense neighborhood
- The distance from a normal subject to the neighbors should not be long.

Thus the outlier score, as defined above, should be much higher if a participant is an outlier.

Applying kNN on $X1$ with $k = 4$, we identified participants with the top 20% highest outlier scores to , i.e., 12 participants, as possible outliers. Meanwhile, we selected another 12 subjects with the lowest outlier scores as a comparison group.

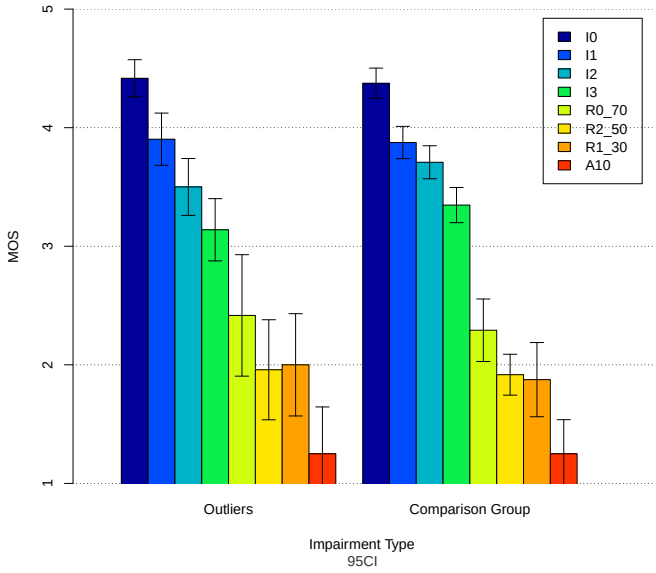


Fig. 5. MOS comparison between high outlier scores and low outlier scores

As a check on how the outlier score works, we calculated Standard deviation of Opinion Scores (SOS) in each group. In [25], it is proposed that the user ratings in a properly executed QoE test should share only a limited amount of variation. The corresponding MOS and standard deviation of the two groups (12 largest outlier scores and 12 smallest outlier scores) are listed in Fig. 5 and Table III.

TABLE III
STANDARD DEVIATION, APPLYING KNN ON SCORES OF X1

	Outliers	Comparison Group
$I0$	0.666	0.542
$I1$	0.937	0.580
$I2$	1.021	0.592
$I3$	1.117	0.632
$R0\_70$	1.213	0.624
$R2\_50$	0.999	0.408
$R1\_30$	1.022	0.741

As expected, SOS in the comparison group were much lower than SOS in the group of 12 participants identified as outliers. This demonstrates that PCA scores can capture the user behavior in a way that facilitates outlier analysis.

D. Clustering Based on PCA Scores

After finding that PCA component scores could identify outliers, we then employed clustering on the PCA components to see further evaluate the QoE assessment across samples. K-means analysis was used for clustering the 48 participants who remained after screening out the 12 subjects with the highest outlier scores.

Before clustering

Figure 6 shows MOS of impairment/ failure types after screening outliers. Generally speaking, MOS values of 48 subjects have the following characteristics:

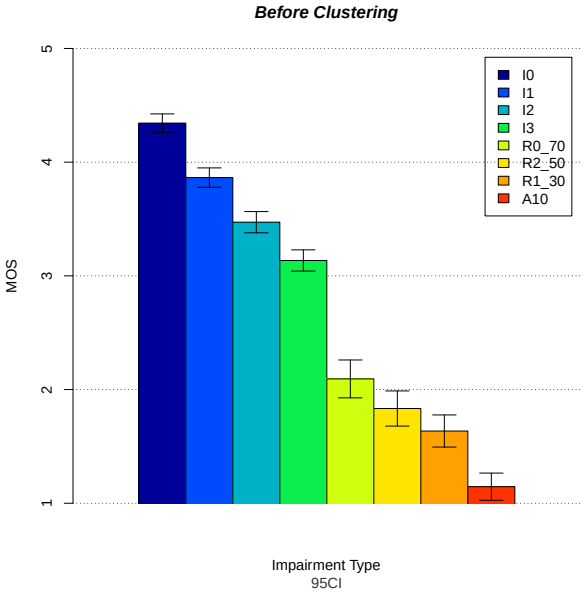
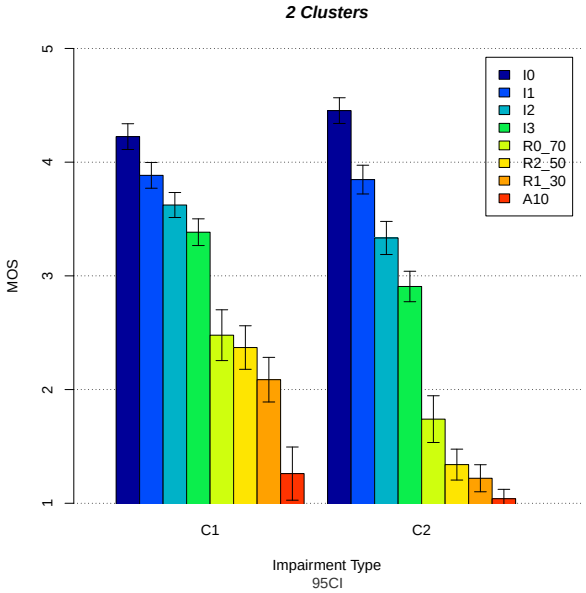
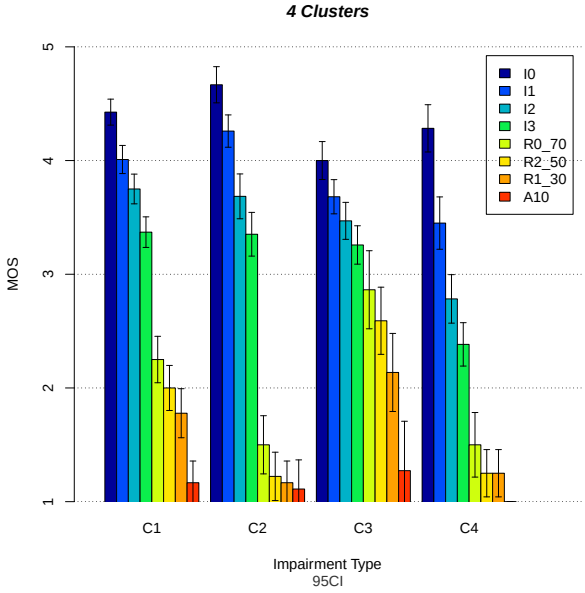


Fig. 6. MOS, before clustering



(a) MOS: 2 clusters



(b) MOS: 4 clusters

Fig. 7. MOS, after clustering

- MOS of all impairment types are above 3.
- MOS of all Retainability failure types are between 1.5 and 2.1.

Two and more clusters

We divided participants into two, and four, clusters respectively. Figure 7(a) shows MOS scores for each cluster in the 2 cluster solution, and Figure 7(b) shows MOS scores for each cluster in the 4 cluster solution. Comparing Fig. 6 and Fig. 7, it can be seen that there are clear differences in rating behavior for the various impairment/ failure types, across the different clusters.

Comparing C1 and C2 in the two cluster solution (Figure 7(a)), it can be seen that C1 participants are less sensitive to differences within impairments, and differences within failures. C1 individuals are also more tolerant of the non-retainability failures, but not the accessibility failure (A10).

For the four cluster solution (Figure 7(b)), C3 shows less sensitivity across all disruption types, although none of the clusters shows obvious bad behavior. This suggests that the outlier score was effective in removing outliers. C2 and C4 are both intolerant of failures, but C4 participants are much more sensitive to differences in impairments. Participants in C1 are similar to C2 in their handling of impairments, but they are relatively more tolerant of failures than the C2 participants.

It can be seen in Table IV that SOS scores tend to be lower after clustering, but this effect is much less than the corresponding differences between participants with high vs. low outlier scores (Table III). The relatively stable SOS scores across the four clusters suggests that these clusters reflect differences in QoE personality, rather than differences in reliability.

Another interesting phenomenon is that the SOS of $R0\_70$ before clustering is 0.822, which is the highest among impairment/ failure types. After clustering, the SOS values of $R0\_70$ decreased within the clusters for both the 2 cluster and 4 cluster cases. The SOS of $R0\_70$ in C2 of the 4 cluster solution decreased to 0.515. Participants in C1 and C3 judge $R0\_70$ less harshly, likely due to the relative completeness of the video since $R0\_70$ provided about 80% of the video content (the average length of video was $96.7s$).

Subjects agreed that failures were worse than impairments. Some subjects rated $R0\_70$ as "Bad" (around 2) while others rated that type of failure as "Fair" (around 3). SOS values of $R2\_50$ and $R1\_30$ decreased after clustering too. This demonstrates that if we want to understand more about user's behavior, clustering is a promising method to find more personalized characteristics.

TABLE IV
SOS BEFORE AND AFTER K-MEANS CLUSTERING

Case	Cluster No.	Subject No.	$I0$	$I1$	$I2$	$I3$	$R0\_70$	$R2\_50$	$R1\_30$
Before clustering	C1	48	0.696	0.732	0.805	0.804	0.822	0.763	0.698
2 Clusters	C1	23	0.673	0.674	0.653	0.698	0.753	0.645	0.661
	C2	25	0.701	0.784	0.902	0.830	0.723	0.479	0.419
4 Clusters	C1	18	0.599	0.648	0.685	0.705	0.604	0.586	0.638
	C2	9	0.583	0.521	0.723	0.705	0.515	0.428	0.384
	C3	11	0.679	0.612	0.662	0.686	0.774	0.666	0.774
	C4	10	0.805	0.891	0.825	0.739	0.607	0.444	0.444

V. CONCLUSION

In this paper, we presented our findings on measures of reliability and accessibility in subjective QoE ratings. Applying PCA on ratings of impairment/failure types allowed us to detect possible outliers. Clustering PCA results divided user behavior into different patterns, which led to four interpretable groupings of participants that differed in terms of their QoE personality.

In this paper we developed an outlier score but we did not solve the problem of where the cutoff in that score should be. We arbitrarily chose a 20% cutoff and it seemed to work since none of the clusters that we obtained showed evidence of unreliable behavior both in terms of their mean ratings (Figure 7) and their SOS scores (Table IV). An alternative way of assessing outliers is to look for visual outliers in plots of the PCA components (e.g., Figure IV-B). While more research on this topic is needed, we suggest that the methods of PCA followed by clustering may be useful in disentangling the issues of reliability and personality in QoE assessment.

ACKNOWLEDGMENT

This research was supported by a grant from TELUS and a matching grant from NSERC/CRD.

REFERENCES

[1] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *Image Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1427–1441, 2010.

[2] Y. Chen, F. Zhang, F. Zhang, K. Wu, and Z. Q., "Qoe-aware dynamic video rate adaptation," in *IEEE Global Communications Conference (GLOBECOM) 2015*, December 2015.

[3] R. Imran, M. Odeh, N. Zorba, and C. Verikoukis, "Spatial opportunistic transmission for quality of experience satisfaction," *J. Vis. Comun. Image Represent.*, vol. 25, no. 3, pp. 578–585, Apr. 2014.

[4] M. Chignell, W. Li, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Enhancing reliability through screening and segmentation: An online video subjective quality of experience case study," *Procedia Computer Science*, vol. 69, pp. 55 – 65, 2015, the 7th International Conference on Advances in Information Technology.

[5] P. Reichl, S. Egger, S. M?ller, K. Kilkki, M. Fiedler, T. Hossfeld, C. Tsiaras, and A. Asrese, "Towards a comprehensive framework for qoe and user behavior modelling," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, May 2015, pp. 1–6.

[6] Y. Chen, Q. Chen, F. Zhang, Q. Zhang, K. Wu, R. Huang, and L. Zhou, "Understanding viewer engagement of video service in wi-fi network," *Comput. Netw.*, vol. 91, no. C, pp. 101–116, Nov. 2015.

[7] I. Ketyk, K. D. Moor, W. Joseph, L. Martens, and L. D. Marez, "Performing qoe-measurements in an actual 3g network," in *Broadband Multimedia Systems and Broadcasting (BMSB), 2010 IEEE International Symposium on*, March 2010, pp. 1–6.

[8] A. Khan, L. Sun, E. Jammeh, and E. Ifeachor, "Quality of experience-driven adaptation scheme for video applications over wireless networks," *IET Communications*, vol. 4, no. 11, pp. 1337–1347, July 2010.

[9] A. Khan, L. Sun, and E. Ifeachor, "Content clustering based video quality prediction model for mpeg4 video streaming over wireless networks," in *Communications, 2009. ICC '09. IEEE International Conference on*, June 2009, pp. 1–5.

[10] Y. Chen, K. Wu, and Q. Zhang, "From qos to qoe: A tutorial on video quality assessment," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 1126–1165, Secondquarter 2015.

[11] ITU-T, "Subjective video quality assessment methods for multimedia applications," *Recommendation P.910, Telecommuniation standardization sector of ITU*, Sep, 2009.

[12] H.-P. Kriegel, P. Kroger, and A. Zimek, *Outlier detection techniques*. Tutorial at PAKDD, 2009.

[13] M. G. Vozalis and K. G. Margaritis, "A recommender system using principal component analysis," in *Published in 11th Panhellenic Conference in Informatics*. Citeseer, 2007, pp. 271–283.

[14] T. Hoβfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of youtube qoe via crowdsourcing," in *Multimedia (ISM), 2011 IEEE International Symposium on*, Dec 2011, pp. 494–499.

[15] T. Hoβfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, Feb 2014.

[16] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," in *Proceedings of the ACM SIGCOMM 2011 Conference*, ser. SIGCOMM '11. New York, NY, USA: ACM, 2011, pp. 362–373. [Online]. Available: http://doi.acm.org/10.1145/2018436.2018478

[17] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36–41, March 2010.

[18] A. Khan, L. Sun, E. Jammeh, and E. Ifeachor, "Quality of experience-driven adaptation scheme for video applications over wireless networks," *IET Communications*, vol. 4, no. 11, pp. 1337–1347, July 2010.

[19] W. Li, H. Ur-Rehman, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Impact of retainability failures on video quality of experience," in *Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference on*, Nov 2014, pp. 524–531.

[20] W. Li, H.-U. Rehman, D. Kaya, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Video quality of experience in the presence of accessibility and retainability failures," in *Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), 2014 10th International Conference on*, Aug 2014, pp. 1–7.

[21] P. Spachos, W. Li, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Acceptability and quality of experience in over the top video," in *Communication Workshop (ICCW), 2015 IEEE International Conference on*, June 2015, pp. 1693–1698.

[22] A. Leon-Garcia and L. Zucherman, "Generalizing mos to assess technical quality for end-to-end telecom session," in *Globecom Workshops (GC Wkshps), 2014*, Dec 2014, pp. 681–687.

[23] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation BT. 500-13, Recommendations of the ITU, Radiocommunicaion Sector*, 2012.

[24] W. Li, P. Spachos, M. Chignell, A. Leon-Garcia, L. Zucherman, and J. Jiang, "Impact of technical and content quality of overall experience of OTT video," in *IEEE Consumer Communications and Networking Conference (CCNC) 2016*, Aug 2016.

[25] T. Hofeld, R. Schatz, and S. Egger, "Sos: The mos is not enough!" in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, Sept 2011, pp. 131–136.