Analysis of Problem Tokens to Rank Factors Impacting Quality in VoIP Applications

Jayant Gupchup, Yasaman Hosseinkashi, Martin Ellis, Sam Johnson and Ross Cutler
Microsoft Corporation
Redmond, WA, USA
{jayagup, yahossei, maellis, sajohnso, rcutler} @microsoft.com

***Abstract*—User-perceived quality-of-experience (QoE) in internet telephony systems is commonly evaluated using subjective ratings computed as a Mean Opinion Score (MOS). In such systems, while user MOS can be tracked on an ongoing basis, it does not give insight into which factors of a call induced any perceived degradation in QoE – it does not tell us what *caused* a user to have a sub-optimal experience. For effective planning of product improvements, we are interested in understanding the impact of each of these degrading factors, allowing the estimation of the return (i.e., the improvement in user QoE) for a given investment. To obtain such insights, we advocate the use of an end-of-call "problem token questionnaire" (PTQ) which probes the user about common call quality issues (e.g., distorted audio or frozen video) which they may have experienced. In this paper, we show the efficacy of this questionnaire using data gathered from over 700,000 end-of-call surveys gathered from Skype (a large commercial VoIP application). We present a method to rank call quality and reliability issues and address the challenge of isolating independent factors impacting the QoE. Finally, we present representative examples of how these problem tokens have proven to be useful in practice.**

Keywords—quality of experience; VoIP; data analysis

I. Introduction

The quality of experience (QoE) of VoIP and video-based communication services is commonly reported in terms of the Mean Opinion Score (or MOS) [1], [2]. A MOS value represents an average of subjective quality scores reported by end users and ranges from 1 to 5 – with 1 being the worst quality and 5 being perfect quality. While MOS ratings are useful in evaluating *overall* system quality, detailed ground truth on the *specific* quality degradations experienced by the user is often hard to obtain. Therefore, in addition to prompting for the opinion score, our application presents the user with a set of follow-up options to indicate the existence of commonly experienced quality degradation which may have occurred during the call. We refer to these additional options as problem tokens. The details of the call quality feedback dialog (CQF) used to gather the opinion score, and the problem token questionnaire (PTQ) for audio and video calls is shown in Figure 1. Note that we do not present the PTQ if the user gives an opinion score of 5 – indicating a perfect experience with "no problems".

The PTQ is a rich source of data that provides us with insights into the areas where the user felt that their QoE was degraded. In addition to providing us information about the system quality, it also allows us to collect ground truth for improving the performance of various components. For example, it is extremely challenging to detect if the user experienced any "echo" artifacts during the call using technical statistics. If the system was able to reliably detect echo using technical metrics, algorithms for echo cancellation would be applied to minimize echo artifacts. Not surprisingly, users are more likely to fill out the PTQ for calls with ratings of 1 or 2 (herein poor calls) compared to calls rated three or more (herein good calls); this bias in response rate is discussed further in Section IV. The share of poor calls expressed as

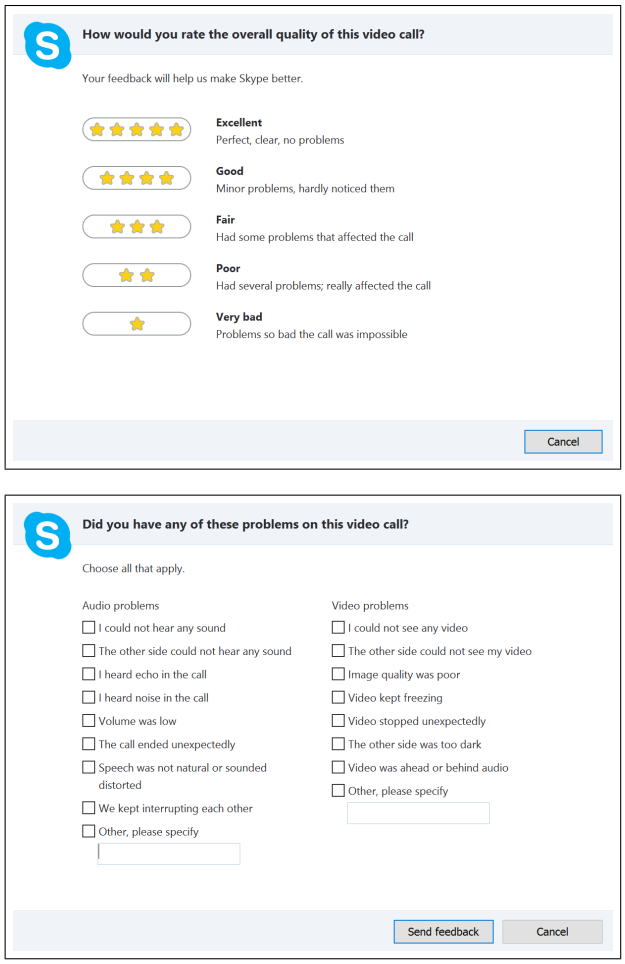

Fig. 1. The top panel shows the Skype call quality feedback (CQF) dialog shown at the end of a call. The CQF dialog allows a user to provide an overall subjective rating. The bottom panel shows the problem token questionnaire (PTQ) if the user gives an imperfect subjective rating.

a percentage of the total count of calls is referred to as the poor call rate (PCR). In this paper, we will use PCR as the metric of choice, however the methods we present can be applied to any other VoIP quality metric, average call duration (ACD) being one example.

In this paper, we focus on how we use the data gathered from the PTQ to gain actionable insights. In the course of analysis of PTQ data, we have obtained several results that we feel are useful to the community. Our main contributions are as follows:

1) We show that problem tokens are highly informative in explaining poor experiences. Problem tokens result in a 73% reduction in entropy (information gain) of the poor call label.
2) We present a method to estimate the impact to quality metrics and rank of impediments as measured by problem tokens. Note that this rank significantly differs from the rank of the overall token frequencies.
3) We improve the estimate of PCR impact on token areas by identifying factors that are relatively orthogonal using the correlation structure in the reported tokens.
4) We present practical applications of using problem tokens in decision making.

The rest of the paper is organized as follows: Section II provides a review of the related work. In Section III, we provide details of the data used for the analysis and results. Section IV presents the main contributions of our work, outlining our analysis methods and the results of said analysis. Based on our experience, Section V discusses some practical and real-world applications of using problem tokens. In Section VI, we summarize and outline possible future work.

II. Related Work

In VoIP applications, it is common practice to correlate subjective experience ratings with telemetry gathered from the various back-end system components for evaluation. Jiang et al. [3] studied the correlation and prevalence of poor networking conditions (network jitter, packet loss, etc.) on PCR. Pessemier et al. [4] combined subjective quality ratings with technical metrics using a decision tree to understand the technical features that best explain the subjective ratings. Their study found that user-perceived quality decreases as users get more familiar with the system while the average call duration increases over a period of 120 days. The analysis in this paper differs from the work of Pessemier et al. in the following ways: First, we correlate subjective ratings with problem data gathered from user feedback (as opposed to technical metrics), and second, the goal of our study is to breakdown quality metrics in terms of the rank and impact to the metric from the perspective of the user. The decision tree approach is highly suited for troubleshooting but it does not provide a breakdown of the top-level metric into its components in an uncorrelated manner. Moller et al. [5] outline a taxonomy structure, definitions of factors and their relationships to characterize the quality of experience. They advocate a questionnaire framework [6] for evaluating interaction quality of experience.

There is a body of work addressing the topic of subjective quality assessment. Methods for measuring subjective audio and video quality have been defined within ITU-T Rec. P.800 [2] and P.910 [7], and work continues within ITU-T's Study Group 12 to standardize new methods for objective quality assessment [8]. These methods include techniques for objective measurement of audio and video quality from technical factors, such as the ITU E-model [9], as well as full-reference metrics such as POLQA. These methods are generally intended for offline use, and therefore are of limited value in evaluating live systems. Weiss et al. [10] evaluated different approaches to predicting the overall subjective quality of speech using the quality of individual segments of calls. They found that most models (Weiss [10], Rosenbluth [11] and ETSI models [12]) outperformed simple averaging of MOS.

User studies have been used for decades to gather human feedback on audio/video quality for the purposes of performance evaluation. Traditionally, this has involved in-lab studies with a small number of participants, but more recently online crowdsourcing platforms (e.g., Amazon Mechanical Turk) have allowed sampling of wider population of users. This has been used for QoE evaluation in still image and audio/video scenarios [13], [14], [15]. By using crowdsourcing, it is possible to quickly obtain a very large number of evaluation samples, although there is additional variance in such experiments due to the lack of control compared with an in-lab study.

A number of data analysis techniques have been used to estimate the impact of predictors. The ideas outlined in [16] provide a good overview of the approaches used to estimate the importance of correlated predictors.

III. Dataset

The data and results reported in this paper were obtained from end-of-call surveys collected during real-world calls made using Skype. The details of the dataset are as follows:

- Calls were sampled uniformly at random from users during a two week period.
- Calls were one-to-one, rather than group or conference calls, and included both audio-only and video calls.
- 700,000 unique calls from in excess of 100,000 unique users.

If a user rates a call less than 5 on the CQF dialog then the PTQ is shown; however, since submission is optional, some ratings do not have corresponding problem tokens. The *representative* dataset has a significantly higher percentage of calls that are labeled good calls. For some results, we will resample the data at random such that the distribution of class labels (poor vs. good) is balanced. This secondary dataset is referred to as the *balanced* dataset. Unless otherwise specified, we will report results on the *representative* dataset. At this point, we would like to draw attention to our approach in presenting results in the rest of the paper. Since Skype is a commercial application, we are unable to provide absolute numbers of the quality metrics. However, we will provide relative ranks (scaled) to convey the relevant information.

IV. Analysis & Results

A. Informativeness of Problem Tokens

The percentage of the problem token selection for all rated calls and calls with poor ratings is shown in Figure 2. The following observations can be made based on the figure:

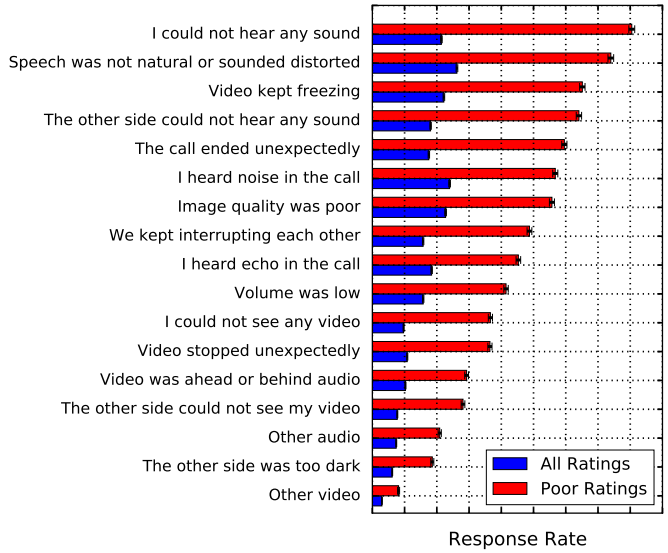

Fig. 2. Problem token response rates for all rated calls and poor calls[1].

- Users are significantly more likely to respond to PTQ when the call is rated as poor compared to when the call is rated as good. For example, the response rate for "I could not hear any sound" token is about three times higher for poor calls compared to the overall rated population.
- The response rate sort order is different for overall rated calls and poor calls. This indicates that some problem areas are more likely to result in a poor call compared to others.

While it is clear that users are more likely to respond to the PTQ questionnaire when they have a poor experience, the response rate (user selecting any token) is about 54% among the poor call population, which can dilute some of the results. In order to mitigate this bias, from here onwards our analysis considers only poor calls where the token feedback is provided. Note that we resample the data such that the original PCR is preserved.

Computing information gain [17] is another approach to measuring the information content present in the problem tokens. The information gain of two uncorrelated variables is 0. At the other extreme, the maximum value of information gain is 1; in other words, it represents the reduction in uncertainty achieved in one variable when we know the value of the other variable. We compute the information gain on the balanced dataset between the $poor\_call$ indicator variable and $any\_token\_reported$ indicator variable – a Boolean vector set to 1 if a user selected any problem token; else 0. The information gain for the dataset was found to be 0.73. Since this is computed on a balanced set, the information gain also represents the fractional reduction in entropy for the $poor\_call$ label if we know $any\_token\_reported$.

[1]To preserve commercial confidentiality, absolute values are hidden in figures throughout the paper.

Algorithm 1 TIMU – Token impact on metric univariate

procedure TIMU($df, problem\_set, metric, fix\_value$)
  $df\_fix \leftarrow$ COPY(df)
  $df\_fix[problem\_set, metric] \leftarrow fix\_value$
  $metric\_original \leftarrow$ MEAN($df[metric]$)
  $metric\_fix \leftarrow$ MEAN($df\_fix[metric]$)
  $mean\_impact \leftarrow$ ABS($metric\_original - metric\_fix$)

  ▷ Use propogation of errors to estimate ...
  ▷ uncertainty of the impact of the metric
  $metric\_var \leftarrow$ VAR($df[metric]$)
  $metric\_fix\_var \leftarrow$ VAR($df\_fix[metric]$)
  $metric\_fix\_cov \leftarrow$ COVARIANCE($(df[metric], df\_fix[metric])$)
  $combined\_std \leftarrow \sqrt{metric\_var + metric\_fix\_var - metric\_fix\_cov}$

  $combined\_se \leftarrow \frac{combined\_std}{\sqrt{df.rows}}$
  $mean\_impact\_95\_ci \leftarrow 1.96 * combined\_se$

  return $mean\_impact, mean\_impact\_95\_ci$

B. Impact of Problem Areas on Metrics

The token frequencies (Figure 2) provide us with a ranking for prioritizing product improvement areas. It is worth noting that the response rate of tokens is quite different for all rated calls versus poor calls. For example, the percentage of users reporting "I could not hear any sound" is lower than those reporting "I heard noise in the call" for all calls. While the former represents a catastrophic situation where users cannot proceed with completing the desired task, the latter might be an annoyance but would not prevent completion of the desired task. This is the intuition behind why we see a higher rank for the token "I could not hear any sound" compared to "I heard noise in the call" when only considering poor calls.

The above intuition points to the fact that the impact to the PCR metric for each problem token is related non-linearly with the overall token frequencies. Therefore, the ranking provided by the impact to the metric is a more natural way to prioritize product improvements than considering raw token frequencies.

In order to map the token frequencies to the impact on quality metrics, we use two approaches. The first approach relies on two assumptions:

- Independence: A problem token is set independently of other problem tokens.
- Mutual exclusion: Users selecting a particular token would not have had a poor experience if they had not encountered this impediment.

This approach is referred to as the token impact on metric univariate (TIMU). The TIMU method is suitable for ranking impairments. The second solution, token impact on metric multivariate (TIMM), addresses the independence and mutual exclusion assumptions. TIMM provides a logical grouping of problem areas, and an estimate of the impact of those areas in terms of the quality metric. We advocate using TIMU and TIMM in conjunction – while TIMM identifies groups and provides an estimate of the impact between groups, TIMU allows us to rank areas within those groups. Next, we will go into the details of the two approaches.

C. TIMU Approach

The TIMU approach is outlined in Algorithm 1. The idea is to estimate the impact of problem tokens on a quality metric

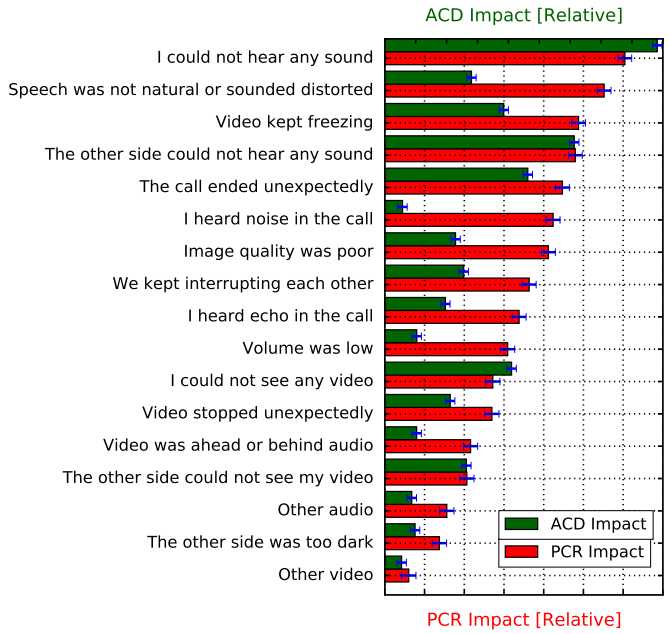

Fig. 3. Estimated impact of problem tokens on PCR and ACD using TIMU.

in a univariate fashion (i.e., without considering the correlation among tokens). The procedure accepts the following arguments: the dataset, set of problem calls, name of the metric, and a value for the quality metric that would reflect a good experience – for PCR, we pick a value of 0 indicating that call would not have been rated as poor. For ACD, we pick the average of the call duration for calls where no problem is reported. The idea is to apply the "fix value" on the problem set. The difference in the original metric and the fixed metric is the impact of the problem set on a given metric. Lines 8-14 show the computation of the uncertainty of the estimate using propagation of error technique [17], [18]. This is done by combining the estimate of the variance of the original metric, the fixed metric, and the covariance among the two.

The outcome of applying the TIMU approach for PCR and ACD is shown in Figure 3. It is interesting to note that the rank of the problem areas is different for PCR and ACD. The media reliability metrics ("I could not hear any sound", "Call ended unexpectedly", "I could not see any video") have the highest impact for ACD. A number of quality areas such as "unnatural or distorted speech" and "freezing video" have more impact on PCR then on ACD. We have found this approach to be very useful to rank areas that need improvement (or investment). One shortcoming that needs to be mentioned here is that the impact on the metric is overestimated due to the correlation and mutual exclusion assumptions. However, the results can be used to estimate the rank of the problem areas. This shortcoming is addressed by the TIMM approach.

Before proceeding to the TIMM approach, we further motivate the need to improve on the TIMU approach by looking at the correlation of the problem tokens. We use the Jaccard similarity score [19] to measure the degree of overlap between the tokens. Perfect overlap between two Boolean vectors results in a Jaccard similarity score of 1, whereas no overlap leads to a Jaccard similarity score of 0. The token correlations

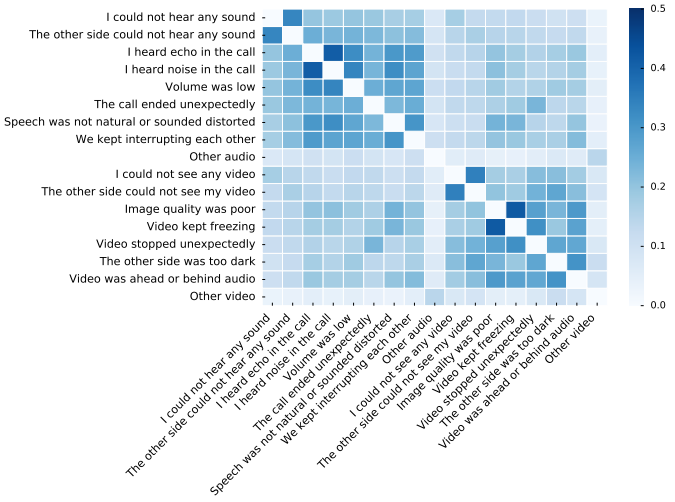

Fig. 4. Jaccard similarity scores for problem tokens (diagonals set to zero).

Algorithm 2 TIMM – Token impact on metric multivariate

1: **procedure** TIMM(df, $metric$, $loading\_threshold$)
2: **Clean Data:**
3: ▷ Remove uninformative variables from df to avoid singularity
4: **Estimate PCorr:**
5: ▷ Estimate polychoric correlation matrix from cleaned problem token matrix
6: **Tune NumLatentFactors:**
7: ▷ Use Parallel Analysis on PCorr to fix the number of latent factors
8: **Estimate Dimension Loadings:**
9: ▷ Suppress weak loadings to zero using $loading\_threshold$
10: ▷ Generate new dimensions from the factor loading of dominant contribution
11: **Build Predictive Model for** $metric$ **Using Estimated Dimensions:**
12: ▷ Fit generalized linear model (GLM)
13: ▷ $metricNoChange \leftarrow$ Predict $metric$ value for mean dimension values
14: **Estimate Impact on Metric:**
15: **for each** $dim \in Dimension$ **do**
16: ▷ $metricChange \leftarrow$ Predict metric when the dim is reduced
17: ▷ Use $metricChange$ and $metricNoChange$ to estimate the improvement in metric

are shown in Figure 4. Note that the diagonal elements have been made zero as those would always represent 1. We see very strong correlations among the tokens. For example, when users complain about "echo", more than 40% of the time they also complain about experiencing "noise" in the call. In 34% of cases, "video stopped unexpectedly" complaints are accompanied by "poor video quality" complaints. While it is our goal to make these tokens as unambiguous as possible during the design phase of these tokens (out of scope of this paper), it is clear that users perceive quality problems as a collection of problem groups rather than a single problem. Therefore, we need an approach that computes the impact to PCR by considering these correlations.

D. TIMM Approach

The TIMM approach is based on projecting the observed data into a lower dimensional space of meaningful factors and carrying on the estimation of impact on metrics in the lower dimensional space. This is achieved through Exploratory Factor Analysis (EFA) [20], [21] and Generalized Linear

TABLE I. FACTORS EXTRACTED FROM PROBLEM TOKENS

Problem Groups (%Variance explained)	**Problem Tokens**
Audio Quality (26%)	We kept interrupting each other Speech was not natural or sounded distorted Volume was low I heard echo in the call I heard noise in the call
Video Quality (25%)	The other side was too dark Video stopped unexpectedly Video was ahead or behind audio Image quality is poor Video kept freezing
One-way Video (12%)	I could not see any video The other side could not see my video
One-way Audio (11%)	I could not see any sound The other side could not see my sound
Reliability (7%)	The call ended unexpectedly

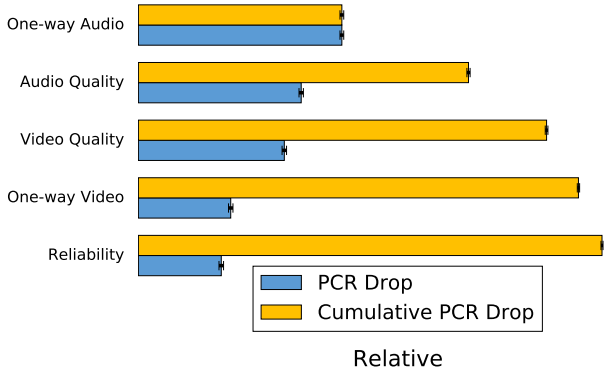

Fig. 5. Predicted maximum relative reduction in PCR using TIMM.

Model (GLM) techniques [22]. We skip the details of EFA and GLM methodology and instead briefly discuss the key characteristics that make these standard frameworks work so well for problem token data. Since the problem tokens are ratings that indicate users' satisfaction (i.e., the token is set to 1 if problem is encountered and 0 if not encountered), these tokens can be modeled as dichotomous observations of a continuous trait, say "satisfaction level". If satisfaction level dips lower than a certain threshold, the user rates 1, otherwise 0. This way, the observed variable is binomial while the latent variable is continuous. The correlation structure between latent continuous variables is estimated from binary observations using the Polychoric correlation coefficient [23], [24]. We compared EFA on Polychoric correlation to Principal Components Analysis (PCA) on Pearson correlation. In addition to the theoretical incompetence of Pearson correlation coefficient for binomial data, this approach does not preserve class separability (i.e., separation between good and poor calls), nor provided interpretable results. An overview of the TIMM procedure is provided in Algorithm 2.

The Polychoric correlation coefficient proved to be highly effective in revealing meaningful groupings of problem tokens through EFA (with varimax rotation [21]). A 5-dimensional subspace of rotated factors with dominant loadings accounts for 81% of total variability in the 15-dimensional space of problem tokens. These factors are not orthogonal as in PCA [20], but provide a reasonable trade-off between interpretability and dimensionality reduction. The weak remaining correlation between factors is captured in the GLM model through interaction effect terms. By dropping the tokens with small loading from each factor (we used a threshold of 0.5), the problem groups (PGs) shown in Table I are uncovered.

Logistic regression is used to predict the reduction in PCR by fixing each of the problem groups shown in Table I. The most accurate model consists of all the main effect terms (the PGs) and two interaction effect terms; specifically between two pairs of PGs: Audio Quality (PG1) and Video Quality (PG2), and Audio Quality (PG1) and One-way Audio (PG4). In practical terms, this means that when PG1 and PG2 (and similarly PG1 and PG4) are reported *together*, they have an impact different to the sum of their individual contributions. The Area Under Curve (AUC) using this approach is 95%; this is a significant improvement over the baseline approach of using $any\_token\_reported$. The baseline method has a false positive rate (FPR) of 10.8% and a true positive rate (TPR) of 48%. At the same FPR, the logistic regression model has a TPR of 93% resulting in a significant improvement in performance.

Figure 5 shows the maximum reduction that can be achieved by fixing a single PG at a time. The blue bars indicate the reduction in PCR if a single PG is fixed while all other PGs still occur at their current level. In the population we studied, the data indicates that fixing One-way Audio has the highest return on investment (RoI) while Reliability shows the smallest RoI in terms of user satisfaction. This provides the priority in problem groups and helps formulate efforts to fix them within our study population. Note that the values shown in blue are not additive since they represent the drop in PCR assuming only one problem group is fixed. However, the interaction terms in the model help to predict the combined effect. Yellow bars in Figure 5 demonstrate the expected cumulative drop in PCR. It is worth mentioning that we see TIMU and TIMM methods providing complementary information. While TIMM provides an estimate of RoI in fixing problem groups, TIMU provides a relative ranking withing the problem group.

V. DISCUSSION

In our experience, problem tokens have served as a useful source of data in solving many practical decision making challenges. Here, we outline some representative examples.

A. Analysis of quality for new releases/versions

When new versions of Skype are released, engineering teams are keen to track the user-perceived QoE. This is usually done by comparing the quality metrics of the new release to previous releases; typically, regressions in quality attract more attention than improvements. Upon discovering that a quality metric has regressed, the natural response is to ask which changes in the product have caused this regression. However, this is not always an easy question to answer. A typical release contains a number of changes that can interact with each other in complex ways. These changes may not be detected in component, integration or end-to-end regression tests, but once released may interact under certain hardware or network conditions previously unknown – resulting in poor experiences for potentially millions of users. On numerous occasions, we

have used the problem token data as a first response to reduce the search space of the quality regression. For example, one release contained a change to bandwidth allocation logic, a corner case resulted in a sharp uptick in PCR and the response rate of the "I could not hear any sound" token. This allowed us to narrow down the underlying problem resulting in a faster turnaround time for the fix.

B. Unbiased comparisons when updating system components

Problem token data has been useful in evaluating the user experience when making systemic changes in components. The problem in evaluating systemic changes is that the technical metrics are often not comparable between the two systems. For example, when making a major overhaul in the jitter control component, we were unable to use the technical metrics to compare the two systems, since the definitions of the metrics themselves had changed. However, the associated problem tokens ("We kept interrupting each other", "Speech was not natural or sounded distorted") are based on user feedback, and can therefore be used to compare the two components.

VI. Summary

In this paper, we analyze the value of the end-of-call "problem token questionnaire" in Skype calls. Using a dataset collected from over $700,000$ calls, we show that problem tokens give useful insights in understanding the areas where our users perceive a quality degradation. We show that instead of relying on the raw token frequencies of problem tokens, these data can be used more effectively by estimating the impact on quality metrics. Towards this goal, two approaches are presented with the requirement that results are easy to interpret and take action on.

The TIMU method is used to rank the problem areas that are impacting quality metrics experienced by users. The TIMM method exploits the correlation structure of the problem tokens to learn categories, and estimates of impact to the quality metrics within those categories. The goal of these two methods is to provide the next level of detail by breaking down a quality metric, this is then primarily used to estimate areas that require improvement. We also share some practical examples of how problem tokens can be employed by engineering teams for effective decision-making in situations where technical metrics are not easily available.

We note that the design of the PTQ (as with any questionnaire) is a key factor for the effectiveness and response rate of these tokens. Techniques for effective design include keeping the question set small, using clear and unambiguous text, and randomizing presentation order to minimize priming bias [25]. However, we defer discussion of these issues to future work.

To conclude, we would like to emphasize that understanding the overall impact of the problem tokens provides us with a very natural way to measure user-perceived QoE, and has allowed us to make investments to improve it.

Acknowledgments

We would like to thank Mu Han, Robert Aichner, and the Skype call quality data science team for useful discussions on problem token analysis.

References

[1] F. De Rango, M. Tropea *et al.*, "Overview on VoIP: Subjective and objective measurement methods," *International Journal of Computer Science and Network Security*, vol. 6, no. 1, pp. 140–153, 2006.

[2] ITU-T, "Methods for subjective determination of transmission quality," International Telecommunication Union, Telecommunication Standardization Sector, August 1996, Rec. ITU-T P.800.

[3] J. Jiang, R. Das *et al.*, "Via: Improving internet telephony call quality using predictive relay selection," in *SIGCOMM '16: Proceedings of the 2016 ACM SIGCOMM Conference*, 2016, pp. 286–299.

[4] T. Pessemier, I. Stevens *et al.*, "Analysis of the quality of experience of a commercial voice-over-IP service," *Multimedia Tools and Applications*, vol. 74, no. 15, 2015.

[5] S. Möller, K.-P. Engelbrecht *et al.*, "A taxonomy of quality of service and quality of experience of multimodal human-machine interaction," in *QoMEX 2009: Proceedings of the International Workshop on Quality of Multimedia Experience*, 2009, pp. 7–12.

[6] ITU-T, "Subjective quality evaluation of telephone services based on spoken dialogue systems," November 2003, Rec. ITU-T P.851.

[7] ——, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, Telecommunication Standardization Sector, April 2008, Rec. ITU-T P.910.

[8] P. Coverdale, S. Möller *et al.*, "Multimedia Quality Assessment Standards in ITU-T SG12," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 91–97, 2011.

[9] ITU-T, "The E-model: a computational model for use in transmission planning," International Telecommunication Union, Telecommunication Standardization Sector, June 2015, Rec. ITU-T G.107.

[10] B. Weiss, S. Möller *et al.*, "Analysis of call-quality prediction performance for speech-only and audio-visual telephony," in *QoMEX 2014: Proceedings of the 6th International Workshop on Quality of Multimedia Experience*, 2014.

[11] J. Rosenbluth, "ITU-T Delayed Contribution D. 064: Testing the quality of connections having time varying impairments," *Source: AT&T*, 1998.

[12] European Telecommunications Standards Institute, "Speech and multimedia Transmission Quality (STQ); Estimating Speech Quality per Call," ETSI TR 102 506 v. 1.1.1, Tech. Rep., 2006.

[13] K.-T. Chen, C.-C. Wu *et al.*, "A Crowdsourceable QoE Evaluation Framework for Multimedia Content," in *MM '09: Proceedings of the 17th ACM international conference on Multimedia*, 2009.

[14] F. Ribeiro, D. Florencio, , and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *ICIP 2011: Proceedings of the 18th IEEE International Conference on Image Processing*, 2011.

[15] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd – A framework for crowd-based quality evaluation," in *Proceedings of the 2012 Picture Coding Symposium*, 2012.

[16] S. Tonidandel and J. M. LeBreton, "Relative importance analysis: A useful supplement to regression analysis," *Journal of Business and Psychology*, vol. 26, no. 1, 2011.

[17] P. R. Bevington and D. K. Robinson, *Data reduction and error analysis for the physical sciences*. McGraw-Hill, 2003.

[18] K. O. Arras, "An Introduction to Error Propagation: Derivation, Meaning and Examples of Cy = Fx Cx Fx*ı*," École Polytechnique Fédérale de Lausanne, techreport LSA-REPORT-1998-001, September 1998.

[19] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2006.

[20] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, 2007.

[21] B. Everitt and T. Hothorn, *An Introduction to Applied Multivariate Analysis with R (Use R!)*. Springer, 2011.

[22] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. Taylor & Francis, 1989.

[23] K. Pearson, *Mathematical contributions to the theory of evolution*. Dulau and co., 1904, vol. 13.

[24] U. Olsson, "Maximum likelihood estimation of the polychoric correlation coefficient," *Psychometrika*, vol. 44(4), pp. 443–460, 1979.

[25] J. A. Krosnick and S. Presser, "Question and questionnaire design," *Handbook of survey research*, vol. 2, no. 3, pp. 263–314, 2010.