CALLER EXPERIENCE: A METHOD FOR EVALUATING DIALOG SYSTEMS AND ITS AUTOMATIC PREDICTION*

K. Evanini, P. Hunter, J. Liscombe, D. Suendermann, K. Dayanidhi, R. Pieraccini

SpeechCycle, Inc., New York, USA

ABSTRACT

In this paper we introduce a subjective metric for evaluating the performance of spoken dialog systems, Caller Experience (CE). CE is a useful metric for tracking the overall performance of a system in deployment, as well as for isolating individual problematic calls in which the system under-performs. The proposed CE metric differs from most performance evaluation metrics proposed in the past in that it is a) a subjective, qualitative rating of the call, and b) provided by expert, external listeners, not the callers themselves. The results of an experiment in which a set of human experts listened to the same calls three times are presented. The fact that these results show a high level of agreement among different listeners, despite the subjective nature of the task, demonstrates the validity of using CE as a standard metric. Finally, an automated rating system using objective measures is shown to perform at the same high level as the humans. This is an important advance, since it provides a way to reduce the human labor costs associated with producing a reliable CE.

Index Terms— spoken dialog systems, performance evaluation, inter-rater agreement, classification

1. INTRODUCTION

Understanding and evaluating the Caller Experience (CE) provided by commercially-deployed automatic spoken dialog systems is crucial to measuring whether such a system is achieving its performance goals. Knowing how callers seem to experience a system can help guide business and design decisions. Specifically, a call's CE rating can indicate which interactions need to be streamlined, simplified, or made more robust. CE, though, is not a well-understood characteristic and has long been subject to speculation based on marketing survey results and anecdotal interactions with the system. As expected, these uncertain methods of gauging CE are often inaccurate and do not provide clear guidance as to what the true caller satisfaction is, or, just as importantly, what aspects of the interaction affect it.

In addition, it is important to separate CE from the caller's emotional state, which may be influenced by many things that are outside the bounds of the interaction between the caller and the system. Such factors may include external characteristics of the caller's environment or the disposition of the caller at the moment of the call. Caller Experience, though, is actually best defined as *the treatment of the caller by the system*. In other words, compared to an optimal hypothetical human-to-human interaction about the same subject with the same information available, did the automated system treat the caller as well as possible?

As a means of characterizing user satisfication with the system, we propose to measure CE by having expert listeners evaluate a large number of randomly selected recordings of human-computer interactions. The expert listener must understand the basic design of the system and be able to judge how the system is treating the caller. Basic elements of this treatment include questions such as:

- Does the system hear the caller when they say in-scope utterances?
- Does the system accurately recognize what the caller says?
- Are system responses as appropriate and helpful as possible?
- Does the system accurately identify and satisfy the reason for the call?

Having expert listening for several hundred appropriately selected calls can result in an accurate and helpful CE rating. The rating, on a scale of 1 to 5, can be used to make judgments about the usability and efficacy of the system. However, while expert listening is a reliable way to ascertain a CE rating, it has serious limitations. Namely, it requires trained experts and a large investment of time spent listening to calls.

This raises the question: can the rating of CE be automated? This paper postulates that it can. Using data from 1500 calls annotated by 15 expert listeners, we have implemented an algorithm to automatically provide the subjective CE rating from objective measures.

The proposed evaluation metric, CE, differs from previous approaches in that it is a single, subjective rating. The most widely used framework for evaluating spoken dialog systems, PARADISE [1], employs a combination of objective and subjective measures to determine a final overall evaluation. The

*PATENT PENDING

subjective measures most commonly used in the PARADISE framework require that the user respond to a survey about their experience interacting with the dialog system, and contain questions such as the following [2], [3]: *Did you complete the task?*, *Was the system easy to understand?*, *Did the system understand what you said?*, and *Did the system work the way you expected it to?*

Other experiments that have included subjective measures of performance in their evaluation framework have also relied on input from the users of the system [4], [5], [6]. However, this type of subjective data is not necessarily reliable, due to the fact that different users may interpret the questions differently; furthermore, little empirical research has been done into the selection of the questions for the survey [3]. Finally, such surveys are not practical in a real-time system in commercial deployment; since participation in the survey must be optional, any data collected from it would represent a skewed sample of callers. Due to these limitations, we propose to use expert human listeners to evaluate CE based on the treatment received by the caller from the system in comparison to an idealized human-to-human interaction, as described above.

2. EXPERIMENTAL DESIGN

1500 calls were selected to be listened to from an interactive voice-based telephony system currently in deployment. The dialog system is a top-level call router with over 250 distinct call categories [7]. A set of 15 expert raters listened to approximately 100 calls each, and provided a CE rating for each call. Calls in which the caller did not interact with the automated system (e.g., by providing no speech input) were excluded from the CE rating. In total, 1390 calls with a valid CE rating were selected for analysis. Of these, 1188 calls (85%) were randomly selected and set aside as the training set for the automated rater (see Section 2.2). The remaining 202 calls (15%) were selected for repeat listening by expert human listeners. This smaller set was then used to compare both how well the human listeners perform when compared with each other and how well the automated rater performs when compared to human listeners.

2.1. Human Listeners

In order to be able to compare the consistency between different individual human listeners, each of the 202 calls in the test set was listened to two additional times, for a total of three listenings per call. This number was settled on as a compromise between breadth (total number of distinct calls listened to) and depth (number of repeat listeners per call).

For each repeated listening of any given call, a new human listener was selected randomly from the initial set of 15 listeners. When listening to a call for a second or third time, the listeners were not aware of what CE rating was given to the call by the previous listener(s), so that they would not be influenced by the prior ratings. Thus, each of the 202 calls was listened to by three distinct listeners; these three sets of listening tasks will be referred to as *human1*, *human2*, and *human3* below.

2.2. Automated Rater

The automated rater was created by constructing a statstical classifier from the set of 1188 training calls, using the CE values from 1 – 5 provided by the human listeners as the target classes. For each call the feature vector used for training consisted of objective measures that can be automatically extracted from the speech logs that are generated routinely for all calls to the system. Specifically, four of these measures which were considered to be most informative in determining the CE were used for training the automated rater: the classification status of the call (how well the system determined the reason for the call), the number of speech recognition errors during the call, the number of operator requests from the caller, and the exit status of the call (whether the caller's task was completed, or where the caller was subsequently transferred).

A decision tree [8] was chosen for the statistical classifier, since its model is easy to interpret and can provide useful information about the relative importance of the features in the feature set. The classifier was constructed from the training set by iterating over all possible splits of values (y) for all possible features (f) to determine which split produced the highest information gain (IG). IG is defined as the difference in entropy (H) between the distribution (D) before the split and the weighted sum of the entropies of the nodes after the split (for a split that has K possible outcomes) as shown in 1. The decision tree was implemented with a 25% confidence threshold for pruning, and the resulting model contained 31 leaves.

$$IG(f,y) = H(D) - \sum_{k=1}^{K} \frac{|D_k|}{|D|} \times H(D_k) \qquad (1)$$

For each of the 202 test calls, the automated rater chose the most likely class (CE rating) by following the nodes of the decision tree model corresponding to the feature values for that call. The set of CE ratings predicted by the automatic rater are referred to as *auto* below.

3. RESULTS

3.1. Agreement Metric

After the three rounds of call listening were conducted on the test set, the ratings from the three sets of human listeners were compared with each other as well as with the predictions made by the automatic rater. In order to determine how well the different sets of listeners agreed in their subjective evaluation of CE for each call, inter-rater agreement was

measured using Cohen's κ [9]. This metric is a more robust way of comparing agreement than simply using the percent agreement, since it takes into account the amount of agreement expected due to chance based on the distribution of the classes. Cohen's κ is defined in 2, where $P(a)$ is the relative observed agreement between two raters, and $P(e)$ is their hypothetical agreement due to chance.

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)} \tag{2}$$

The simplest κ measurement treats all instances of disagreement between raters identically. However, in a task such as CE rating, where the classes represent a ranked continuum—such that a rating of 5 is closer to a rating of 4 than any other rating, a rating of 2 is closer to 1 or 3 than 4, etc.—it makes more sense to calculate κ by taking into account these inherent distances between the classes. For our comparison, we used a linearly weighted κ, in which each disagreement between two raters is assigned a weight, w, using the formula in 3, where d represents the numerical difference between the classes, and k represents the number of classes.

$$w = 1 - \frac{d}{k - 1} \tag{3}$$

So, in the task at hand with 5 levels of CE (and, thus, a maximum numerical difference of 4 between ratings), an exact agreement between two raters receives a weight of 1, a difference of 1 point receives a weight of 0.75, etc.

3.2. Human-to-human agreement

Table 1 presents the κ value for comparisons between the three sets of human listeners on the test set. All three κ values are quite close, meaning that the level of agreement among the three sets of listeners is consistent. Furthermore, the κ values are relatively high, indicating that different expert human listeners were able to provide similar subjective CE ratings to the same calls.

Tasks Compared	κ
human1 vs. human2	0.77
human1 vs. human3	0.78
human2 vs. human3	0.80

Table 1. Comparison of agreement among human raters

Figure 1 shows a more detailed analysis of the CE rating task. It presents the frequencies of different levels of CE rating differences for each human-to-human comparison. The percentages of calls in which the two human listeners agreed completely (i.e., provided the exact same CE rating) are 54.0%, 56.9%, and 59.4% for the three human-to-human comparisons. Similarly, the combined percentages of calls in which the two human listeners differed by at most one CE point were 88.7%, 87.6%, and 91.6%, respectively.

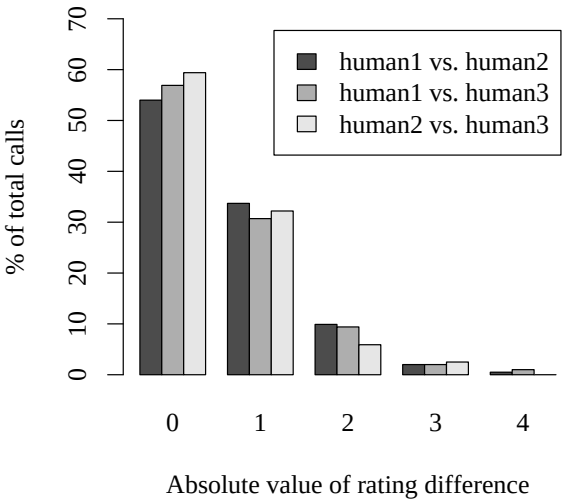

Fig. 1. Comparison of agreement among human raters

3.3. Human-to-automatic rater agreement

The CE predictions from the automatic rater for each call were compared to the CE ratings provided by the three sets of human listeners. The κ values for these three comparisons are provided in Table 2.

Tasks Compared	κ
human1 vs. auto	0.75
human2 vs. auto	0.85
human3 vs. auto	0.80

Table 2. Comparison of agreement between human raters and automatic rater

A comparison of the results in Table 2 with the human-to-human results in Table 1 shows that the automatic rater agrees with human listeners to about the same degree as the humans agree with each other: the average κ for the three human-to-automated rater comparisons, 0.80, is similar to the average κ for the three human-to-human comparisons, 0.78.

Figure 2 shows the number of each degree of difference in CE ratings for the three human-to-autmated rater comparisons. Again, a high percentage in each set achieved a rating that was either identical or within one point: 88.1%, 95.5%, and 92.1%, respectively.

4. DISCUSSION

Comparisons between the results in Tables 1 and 2, on the one hand, and Figures 1 and 2, on the other, show that the automatic rating system is able to provide CE ratings as consistently as humans: the average κ values and the average classification performance of the automatic rater vs. the human listeners are similar to the average values obtained by compar-

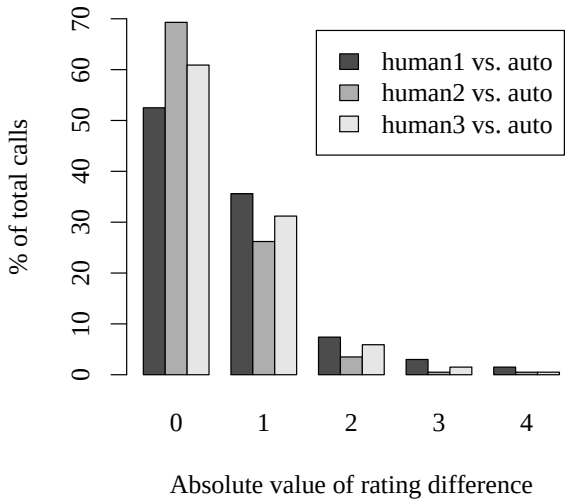

Fig. 2. Comparison of agreement between human and automatic raters

ing the different sets of human listeners. However, there is a larger range of variation among the three human-to-automatic rater comparisons than among the human-to-human comparisons. For example, when compared with the second set of listeners (*human2*) the automatic rater showed the highest κ, and had by far the largest number of exact matches 69.3% vs. 60.9% and 52.5%. However, the fact that the average values for the three comparisons using the automated system are nearly identical to the three human-to-human comparisons suggests that the variation would level out with a larger set of training data demonstrates that the automatic rating process successfully emulates human behavior.

An examination of the decision tree model produced by the training process gives some insight into the criteria being used by the human raters when providing their subjective CE ratings. The first feature that the model splits on is the number of utterances within a call that are not recognized correctly by the system, and the value that it splits on is 1 (i.e., whether the entire call had 0 misrecognitions vs. 1 or more misrecognitions). This finding coincides well with other experiments that have shown that the recognition score is the most reliable predictor of a dialogue system's performance [10]. This fact has also led some to criticize the PARADISE evaluation framework, since they claim that all other dialogue-quality costs are correlated with the recognition score to such an extent that they are no longer meaningful performance metrics [11]. However, we would not go as far as this conclusion; the other three features used as input to train the automated predictor (the call's classification status, its exit status, and the number of operator requests) were all selected as nodes in the pruned decision tree, meaning that they did provide useful information, at least for some calls, in predicting the CE.

5. CONCLUSIONS

In this paper we propose the use of a single, subjective numerical rating to evaluate the performance of a telephone-based spoken dialog system. This Caller Experience rating is provided by expert human listeners who have been rigorously trained and who have knowledge of the design of the dialog system. We demonstrate that different human raters can be trained to achieve a satisfactory level of agreement. Furthermore, we show that a statistical classifier trained on ratings by human experts can reproduce the human ratings with the same degree of consistency. A procedure for reliable automatic rating of CE will prove beneficial to ongoing monitoring of spoken dialog systems, since it enables an increase in both breadth and depth of CE ratings. On the one hand, more calls can be given a CE rating than would be possible with limited human resources; on the other hand, more information can be provided about individual calls, e.g., to help decide between two disparate ratings by different human experts.

6. REFERENCES

[1] Marilyn Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella, "PARADISE: A general framework for evaluating spoken dialogue agents," in *Proceedings of the 35th annual meeting of ACL*, 1997, pp. 271–280.

[2] Candace A. Kamm, Diane J. Litman, and Marilyn A. Walker, "From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems," in *Proceedings of the 5th International Conference on Spoken Language Processing*, 1998, pp. 1211–1214.

[3] Kate S. Hone and Robert Graham, "Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI)," *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, vol. 6, no. 3–4, pp. 287–303, 2000.

[4] Elizabeth Shriberg, Elizabeth Wade, and Patti Prince, "Human-machine problem solving using Spoken Language Systems (SLS): Factors affecting performance and user satisfaction," in *Proceedings of the DARPA speech and NL workshop*, 1992, pp. 49–54.

[5] Morena Danieli and Elisabetta Gerbino, "Metrics for evaluating dialogue strategies in a spoken language system," in *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995, pp. 34–39.

[6] Joseph Polifroni, Lynette Hirschman, Stephanie Seneff, and Victor Zue, "Experiments in evaluating interactive spoken language systems," in *Proceedings of the DARPA speech and NL workshop*, 1992, pp. 28–33.

[7] David Suendermann, Phillip Hunter, and Roberto Pieraccini, "Call classification with hundreds of classes and hundred thousands of training utterances ... and no target domain data," in *Proceedings of the PIT*, Kloster Irsee, Germany, 2008.

[8] J. Ross Quinlan, *C4.5: Programs for machine learning*, Morgan Kaufmann, 1992.

[9] Jacob Cohen, "A coefficient of agreement for nominal scales," *Educational and Psycological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[10] Marilyn A. Walker, Diane J. Litman, and Candace A. Kamm, "Evaluating spoken dialogue agents with PARADISE: Two case studies," *Computer Speech and Language*, vol. 12, no. 3, pp. 317–347, 1998.

[11] Melita Hajdinjak and France Mihelič, "The PARADISE evaluation framework: Issues and findings," *Computational Linguistics*, vol. 32, no. 2, pp. 263–272, 2006.