Say It As You Mean It – Analyzing Free User Comments in the VOICE Awards Corpus

Florian Gödde, Sebastian Möller

Quality and Usability Lab, Deutsche Telekom Labs, Technische Universität Berlin

florian.goedde@telekom.de, sebastian.moeller@telekom.de

Abstract

Usability questionnaires usually contain scales related to effectiveness, efficiency and overall satisfaction which provide a quantitative value for the user's opinion. However, analyzing quantitative data often does not show the reason underlying for a good or bad opinion. Simple questions like "What did you like about the system?" and "What did you not like about the system?" can shade light on the underlying reasons, but a lot of effort is needed for the analysis of such data. Nevertheless, the answers to these questions contain the users' opinion in their own words and hence often show high correlation with the overall rating of the system. In the frame of the SpeechEval [1] project we analyzed the German VOICE Awards corpus over three consecutive years, categorizing the answers to these two free text questions and analyzing correlations between the categories and the overall rating of the systems. We used the data to build a general linear model for predicting the overall rating.

Index Terms: usability, questionnaire, spoken dialog systems

10.21437/Interspeech.2010-561

1. Introduction

Usability evaluations for spoken dialog systems almost always include questionnaires to be filled out by the test participants. Normally, they consist of several questions asking for the experience with certain aspects of the system, like the performance of Automatic Speech Recognition (ASR), Text-To-Speech (TTS), or the dialog structure. Even on widely used questionnaires like the SASSI [2] or the proposed items in ITU-T Rec. P.851 [3], test participants have no option to express their experiences in free text, but have to map them to these questions and scales.

In order to get insight into what users consider as important aspects of a system, free text fields in questionnaires can be very helpful, since the user is able to express his opinion in his own words. In the German VOICE Awards - a competition, where commercial German-language spoken dialog systems are evaluated against each other – test participants had the option to write down remarkable positive and negative aspects of the systems tested. In the following sections, we describe a detailed look at what the users wrote, and how it is related to their overall rating and actual parameters from their dialogs. In Section 2, we introduce the data set we used, followed by the analysis of the corpus in Section 3, and the discussion of the results and a general linear model for rating prediction in Section 4.

2. Data Set

As a data basis we use dialogs and ratings collected during the German VOICE Awards. It consists of dialogs and questionnaires related to commercial German spoken dialog systems which have been tested in the frame of the VOICE Awards from 2005 to 2009. The systems are from different domains, for instance telephone banking, travel information, customer support or even telephone games. They differ much in complexity, ranging from systems that request only one information from the user, to complex systems made for several purposes with a large number of sub-dialogs like banking systems.

All systems were tested by expert and laypersons. For this paper we focus on the ratings of the laypersons in the years 2006-2008. We refer to [4] for a more detailed description of the corpus.

Table 1. *Number of systems and test participants in the Voice Awards corpus.*

	2006	2007	2008
Systems	42	25	30
Participants	10	11	12

The participants – differing each year - were asked to call every system as often as needed to fulfill a given task, and afterwards to rate the system using a non standardized questionnaire developed for the Voice Awards. The items are quantified on a 3- to 5-point Likert scale, depending on the question. Furthermore, the subjects could add free text comments to each item. The last two items were "What was remarkably positive about the system?" and "What was remarkably negative about the system?" to be answered with free text only. Hence, we use these items for the analysis.

To be able to statistically analyze the free text data, we categorized the answers. For nearly every statement we found a topic it fits in. For instance, the answers "I did not like the voice" and "The voice sounded very unnatural to me" would belong to the category "TTS negative". Comments like "Very straightforward handling" or "I always knew in which system state I was in" are both categorized as "dialog structure positive". A few statements that were too rare to have a statistical impact on the results were not categorized, for instance "I liked that the system supported more than one language". This procedure resulted in the categories shown in Table 2.

Table 2. *Categories and sample statements*

Aspect, both pos. and neg.	Sample
ASR	"The system understood me well / bad."
TTS	"I liked / disliked the voice."
Confirmation	"The system did (not) repeat my answers for confirmation."
Efficiency	"The dialog was fast / too long."

Copyright © 2010 ISCA

Elaborateness	"I liked the detailed information given." vs. "Too much unnecessary information was presented."
Dialog structure	"It was easy/hard to navigate through the system."
Learnability	"The system gave a lot of / no instructions."
Kindness	"The system reacted friendly / unfriendly."
Task completion	"I did (not) reach my goal."
Aspect, positive only	Sample
Idea	"Very innovative idea"
Human operator	"I liked being transferred to a human operator."
Aspect, negative only	Sample
System reaction time	"The system reacted too slowly after I said something."

Using these 12 categories, we found 630 statements in the data of 2006,478 statements for 2007, and 562 statements for 2008.

To be able to correctly interpret the results, it is useful to know how elaborate the users were, i.e. how many comments they gave per system. Table 3 shows the number of comments per user and system for each year. Users mostly gave one or two comments on a system. The subjects made zero or three comments nearly equally often, differing slightly more in 2006. Giving 4 or 5 comments about one system was rarely the case.

In the next section, we present more details about the statements and their relation to the overall rating and two dialog parameters, namely dialog length and number of errors.

Table 3. *Number of statements the users made for each system.*

Number of Statements	2006	2007	2008
0	85	50	65
1	121	59	115
2	146	97	110
3	56	50	55
4	11	18	13
5	1	1	2

3. Results

In this section, we present the results of the analysis of the Voice Awards corpus. First, we have a look at the occurrences of the aspects. Then we try to statistically relate them to the overall system rating of the users, and then show relations of some of the categories to actual dialog parameters.

3.1. Quantitative Analysis

Table 4. *Number of statements per aspect and year.*

Aspect, positive	2006	2007	2008
ASR	77	15	37
TTS	33	27	29
Efficiency	50	48	54
Idea	44	34	21
Dialog structure	57	33	52
Kindness	13	21	27
Task completion	6	14	11
Confirmation	4	7	7
Elaborateness	16	7	4
Human operator	2	7	16
Learnability	27	20	39
Aspect negative	2006	2007	2008
ASR	70	53	61
TTS	83	56	58
Efficiency	13	21	24
Sys. reaction time	6	8	13
Dialog structure	52	35	24
Kindness	11	13	9
Task completion	23	16	28
Confirmation	10	3	11
Elaborateness	16	16	24
Learnability	17	24	13

In 2006, the most often mentioned positive aspect of the systems was good speech recognition. In the other two years it is outperformed by dialog efficiency, which is placed third in 2006. Although a good dialog structure plays a major role in all three years, it is mentioned only third most in 2007. In 2007, the innovative idea of a system was the second most often mentioned factor. Important positive factors are also (although the ranking differs between the years) a high quality TTS, the kindness of the system, and the learnability.

ASR performance is not only the most often mentioned positive aspect of the systems, but also one of the most often named negative aspect. Together with often mentioned poor TTS quality, it belongs directly to input/output and is therefore very present in the users perception during the interaction. Another major problem was an unclear dialog structure, bad efficiency and task failure. Also noticeably often mentioned were poor learnability and too elaborate (in most cases long) system prompts.

3.2. Relation to Overall Rating

It is not only interesting to see, what aspects of a spoken dialog system are noteworthy for the users, but also which aspects have the most positive or negative influence on the overall rating.

The mean rating over all systems is shown in Table 5. The users rated the systems on a German school grade system, where 1 is the best rating, and 6 worst. In Table 5 one can see that the highest average rating was achieved in 2006 with 2.64, declining the following to years to 3.08 in 2008.

Table 5. *Mean overall ratings by year*

Year	2006	2007	2008
Mean rating ± std. dev.	2.64 ± 1.357	2.88 ± 1.4	3.08 ± 1.5

In all three years the highest positively influencing factors on the system rating were ASR, kindness and efficiency. ASR and kindness both had nearly the same influence on the ratings, slightly lower influence is observable for efficiency. For instance, users who had written in the field for "remarkable positive aspects", that the system understood them very well (which is category "ASR positive") rated the

system in 2007 on average 0.88 points better than the mean rating of that year, in 2008 even 1.11 points, and 2006 the users still rated the systems 0.77 points better. Table 6 also shows the ANOVA calculations for the user groups who mentioned the aspects compared to those who did not. In all years and for all the three aspects a significant difference is observed. The ANOVA for kindness in 2006 did not show such a high significance, because kindness is mentioned only 13 times in 2006 as a positive aspect of a system, which is not often enough to make a strong difference. But still the significance is on the 0.05 level.

Table 6. *Mean overall ratings and ANOVA results by aspect mentioned and year*

Aspect	2006	2007	2008
ASR *pos.*	**1.9** ± 1.071 F=30.758, p<0.001	**2.0** ± 0.655 F=7.221, p<0.01	**1.97** ± 0.957 F=24.213, p<0.001
Kindness *pos.*	**1.85** ± 1.214 F=4.621, p<0.05	**2.05** ± 1.071 F=9.423, p<0.01	**1.96** ± 1.255 F=17.069, p<0.001
Efficiency *pos.*	**2.02** ± 1.02 F=12.215, p<0.01	**2.23** ± 0.973 F=15.812, p<0.001	**2.20** ± 1.155 F=23.407, p<0.001
Efficiency *neg.*	Not significant	**4.2** ± 1.322 F=19.079, p<0.001	**4.5** ± 1.391 F=24.846, p<0.001
Kindness *neg.*	**3.82** ± 1.25 F=8.73, p<0.01	**4.08** ± 1.038 F=9.456, p<0.01	**4.56** ± 1.59 F=9.188, p<0.01
Dialog structure *neg.*	Not significant	**3.91** ± 1.579 F=21.851, p<0.01	**3.88** ± 1.361 F=7.416, p<0.01
ASR *neg.*	**3.67** ± 1.422 F=56.292, p<0.001	**3.54** ± 1.448 F=12.89, p<0.001	**4.23** ± 1.454 F=50.18, p<0.001
Task completion *neg.*	**4.13** ± 1.359 F=31.869, p<0.001	**3.94** ± 1.692 F=9.079, p<0.01	**4.39** ± 1.397 F=25.116, p<0.001

On the contrary, the highest negatively influencing aspects differ through the years. Whereas in 2006, mentioning bad efficiency and dialog structure had no significant influence on the overall rating, at least efficiency is a very strong factor in 2007 and 2008. Perceived task failure (or negative task completion) and unkindness were major factors in all three years. Bad ASR performance was also significantly influencing the overall rating, less strong in 2007 than the other years.

Another interesting factor is the difference between positive and negative aspects mentioned by the test participants in relation to the overall rating. How does the rating change with more positive or negative aspects noticed? Hence we calculated the difference between the number of positive aspects and negative aspects mentioned per case. The results for all three years are shown in Table 7.

Most of the participants in all years had a difference of 0 between positive and negative aspects mentioned. As seen in Table 3, in 2006 85 of the 140 cases with a positive-negative difference of 0 indeed had written nothing. In the remaining 55 cases the subjects had written an equal number of positive and negative aspects.

Table 7. *Mean ratings by difference of positive and negative aspects and year. Number of occurences in brackets.*

#Pos-#Neg	2006	2007	2008
-3	6.0 (1)	4.88 (8)	5.2 (10)
-2	4.0 (32)	4.33 (24)	4.7 (20)
-1	3.51 (91)	3.91 (43)	3.94 (72)
0	2.64 (140)	2.87 (91)	3.19 (110)
1	1.86 (77)	1.88 (51)	2.29 (79)
2	1.27 (45)	1.7 (20)	1.6 (35)
3	1.63 (8)	1.43 (7)	1.36 (11)

The same applies for 50 participants in 2007 and 45 participants in 2008 that also had written an equal number of positive and negative aspects. Expectedly the average rating is rising with more positive aspects, and declining with more negative. Only in 2006 we observe a more negative mean rating for a difference of +3 than for +2. This is possibly due to the small amount of cases for a difference of +3. The greatest difference between the average ratings is observable for 0 to +1 and 0 to -1. A closer look at these cases reveals another interesting fact: ratings with the same difference between positive and negative aspects are very similar. For instance, the mean rating where this difference was +1, and *exactly one positive aspect* was mentioned, was 1.82 ± 0.583 in 2007 (31 cases, see also Table 8). The remaining 20 cases, where the difference was +1 and *more then one aspect was mentioned*, showed a mean rating of 1.95 ± 0.51, which is no significant difference. The cases where the difference between positive and negative aspects was -1, but more than one negative aspect was mentioned, show a mean rating of 3.87 ± 1.506, which is also not significantly different of the ratings with exactly one negative aspect.

Table 8. *Mean ratings for cases, where exactly one positive (one negative respectively) aspect was mentioned.*

	2006	2007	2008
One positive aspect	**1.76** ± 0.764 N=51	**1.84** ± 0.583 N=31	**2.09** ± 0.908 N=55
One negative aspect	**3.47** ± 1.188 N=70	**3.93** ± 1.215 N=28	**4.02** ± 1.242 N=60

3.3. Relation to Dialog Parameters

Another interesting question to observe is, whether the factors mentioned by the users are measurable in the dialogs. For instance, kindness is obviously hard to measure and quality assessment of TTS systems for finding evidence for positive or negative TTS statements is not in the scope of this paper. But we can have a look at perceived efficiency and perceived ASR performance. We have to limit this to the years 2007 and 2008, since for 2006 we do not have transcribed dialogs for most of the subjects.

Table 9 shows the result of the analysis. There is a strong connection between perceived and actual dialog length. Users who mentioned efficiency as a negative aspect of a system had indeed in both years significantly longer dialogs than users

who did not mention it. On the other hand, users who complimented on the good efficiency of a system had significantly shorter dialogs then the others.

Table 9. *Relation between statements and dialog length and average number of errors per dialog.*

	2007		2008	
	Mentioned	Not mentioned	Mentioned	Not mentioned
Efficiency pos.	175.68 s	212.28 s	146.34 s	190.17 s
Efficiency neg.	269.01 s	199.92 s	211.13 s	180.48 s
ASR pos.	1.55 errors	1.81 errors	0.6 errors	1.33 errors
ASR neg.	1.85 errors	1.78 errors	1.9 errors	1.11 errors

In 2008, we find a significant connection between average errors per dialog and positive as well as negative comments on the ASR quality. Users who mentioned the good speech recognition had an average of 0.6 errors per dialog; users that criticized the bad ASR performance had 1.9 errors per dialog. Average errors over all dialogs were 1.26. However, we cannot observe that difference in the data of 2007. Although it shows the same tendency, the differences in error rates between the users are still too small to be statistically significant.

4. Discussion

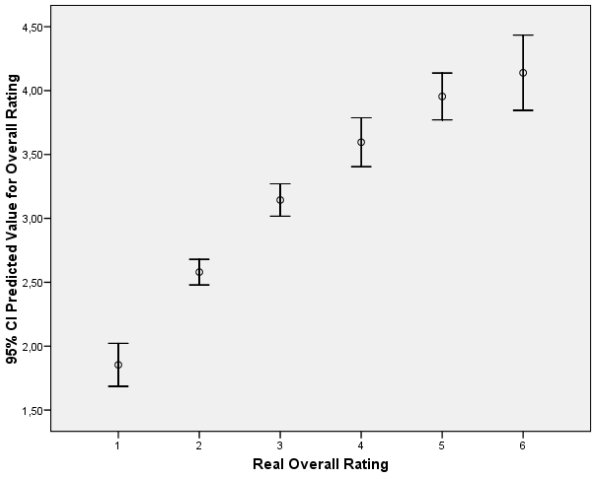
Analyzing the data of 97 systems in the Voice Awards corpus we discovered, that the most noteworthy positive aspects of spoken dialog systems are high efficiency, kindness, a good ASR and dialog structure. Remarkable negative aspects are poor ASR and TTS quality, as well as an unclear dialog structure. But this not necessarily means a high influence on the rating of such a system. Although bad ASR and TTS quality are the most stated flaws of the systems, they do not have such a high influence on the rating like unkindness or bad efficiency. They are just the most obvious aspects of a system, since they mark the interfaces of the system. Given that, it is interesting, that they do not have such a strong influence on the overall opinion.

Furthermore, a clear dialog structure has a significant influence on the user rating. Users seem to like clear questions to answer, since a lot of users stated, that they did not now what to say to get to their goal. Not knowing what to say to the system had a stronger influence on the overall rating than being poorly understood or a poor system voice quality.

We showed that users seem to make a tradeoff between good and bad system aspects, and adjust their rating accordingly. Users with the same difference between the number of good aspects and bad aspects mentioned also gave similar ratings. We plan to use the results of the analysis to improve models for predicting user satisfaction. For instance, we could weight certain dialog parameters like dialog length or number of errors by their perceived importance for the overall rating, to be able to account for more variance in ratings then currently used models.

Figure 1 shows the predicted ratings of a General Linear Model (GLM) trained with the data from the years 2007-2008 against the real ratings. Features included all positive and negative categories plus the number of positive aspects, the number of negative aspects and the difference for all cases. Additionally we included the dialog parameters dialog length in seconds and number of errors. We do not have enough transcribed dialogs for the data of 2006, hence we did not use this data to build the model. Using these features, the GLM produces slightly too negative ratings for systems that are rated high (i.e. 1, 2 or 3), and too positive ratings for systems that were rated 4-6. Extreme ratings like 1 or 6 seem still hard to predict. This could maybe be improved by weighting the statements of the subjects. For instance, in the current analysis the statements “The system did not understand me a few times” and “The system never understood me” are both negative ASR aspects. But in the overall rating, the latter statement presumably has a more negative effect. In future analyses of the data, this could be taken into account by giving the latter statement a higher weight.

Figure 1. *Overall Rating of the subjects vs. mean of the predicted ratings of the GLM.*



5. Acknowledgements

The authors would like to thank our SpeechEval project partners Tatjana Scheffler, Rolland Roller and Norbert Reithinger from the DFKI, as well as Prof. Anthony Jameson and Oliver Jacobs for kindly providing the Voice Awards data. SpeechEval is funded by the Investitionsbank Berlin through the ProFIT framework, grant #10140648. This project is being co-financed by the European Union (European Regional Development Fund). Furthermore we would like to thank our anonymous reviewers for their helpful comments.

6. References

[1] Möller,S., Schleicher,R., Butenkov,D., Engelbrecht,K.P. Gödde,F., Scheffler,T., Roller,R. and Reithinger, N. „Usability Engineering for Spoken Dialogue Systems Via Statistical User Models”, Proceedings of the First International Workshop on Spoken Dialogue Systems Technology (IWSDS 2009)

[2] Hone, K.S. and Graham, R., “Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI)”, Natural Language Engineering, 6(3-4):287-303, 2000.

[3] ITU-T Rec. P.851, “Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems”, International Telecommunication Union, Geneva, 2003.

[4] Scheffler, T., Roller, R. and Reithinger, N., “SpeechEval – Evaluating Spoken Dialog Systems by User Simulation”, KI 2009: Advances in Artificial Intelligence, pp.209-216, Springer Berlin/Heidelberg, 2009.