Evaluation of a Speech-driven Telephone Information Service Using the PARADISE framework: a closer Look at Subjective Measures

Paula M.T. Smeele and Juliette A.J.S. Waals

Department of Perception, Speech and Hearing Group
TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
{smeele, waals}@tm.tno.nl

Abstract

For the evaluation of a speech-driven telephone flight information service we applied the PARADISE model developed by Walker and colleagues [1] in order to gain insight into the factors affecting the user satisfaction of this service. We conducted an experiment in which participants were asked to call the service and book a flight. During the telephone conversations quantitative measures (e.g. total elapsed time, the number of system errors) were logged. After completion of the telephone calls, the participants judged some quality related aspects such as dialogue presentation and accessability of the system. These subjective measures together represent a value for user satisfaction. Using multivariate linear regression, it was possible to derive a performance function with user satisfaction as the dependent variable and a combination of objective measures as independent variables. The results of the regression analysis also indicated that an extended definition of user satisfaction including a subjective measure 'Grade' provides a better prediction than the analysis based on the narrow definition used by Walker et al. Further, we investigated the correlation between the subjective measures by conducting a principal components analysis. The results showed that these measures fell into two groups. Implications are discussed.

1. Introduction

Most service providers enable customers to gather information or order products by using the telephone. To this purpose, call centers have been set up where service agents try to answer the questions of the caller. Nowadays, in case of standard questions, automated interactive voice response (IVR) services are being used. IVR services guide callers through a set of questions (a dialogue) using pre-recorded speech and recognition of the telephone keys pressed by the caller (DTMF). When the number of options the caller can choose from becomes too large or complex, IVR is no longer suitable, simply because the caller cannot memorize all the information presented. However, adding speech recognition, the application domain of the IVR services can be extended. For the evaluation of speech-driven information services, or spoken dialogue systems, one should not revert to evaluations of the individual system components but consider the entire system as the test object. Results of an evaluation of the dialogue component, for example, might indicate longer dialogue durations and fewer system errors for systems using an explicit confirmation strategy compared to systems with an implicit confirmation strategy. However, such an evaluation based on the dialogue structure does not reveal how the users judge the system as a whole. An evaluation method should include measures at different levels: objective measures describing certain measurable characteristics of the system, and subjective measures referring to quality features perceived by the user. Many studies on spoken dialogue systems are focussed on the design process and offer structured protocols for evaluation in order to check whether the system or system component (still) meets the user requirements [2, 3, 4]. These protocols usually include defining and collecting quantitative measures. Other studies concentrate on users' perceptions and try to uncover the underlying features [5, 6]. Only a few attempts are made to describe quality ratings in terms of both system properties and user judgements. Möller [7] proposed a taxonomy that allows quality dimensions to be classified, and methods for their measurement to be developed. The PARADISE evaluation framework developed by Walker and colleagues [1] is an instrument for defining and collecting objective measures concerning dialogue quality, dialogue efficiency, task success, and a number of subjective measures. The model proposes that the primary objective of a dialogue system is to maximize "user satisfaction". It posits that a performance function can be derived by applying multivariate linear regression with user satisfaction as the dependent variable and task success, dialogue quality, and dialogue efficiency measures as independent variables. Although Möllers taxonomy is very extensive, it is not yet clear how the various quality dimensions fit into a performance function describing the overall "quality of service" of a dialogue system. We decided to use the PARADISE model (i) to gain insight into the factors affecting the user satisfaction of a speech-driven telephone flight information service, and (ii) to enable a quantitative description that allows the prediction of user satisfaction from results of objective measurements.

2. Description of the dialogue system

"Irene" is a telephone-based spoken dialogue system that provides callers with information on the availability of flights between some airports in the Netherlands and Spain or France [8]. It consists of the following dialogue parts for which user input is required:

- Departure city
- Arrival city
- Ticket type: single, return
- Departure date
- Return date
- Verification: *"Irene"* uses explicit confirmation strategy

- Book another flight

Other system functionalities are:

- Speech recognition: based on single word recognition
- Error handling and help: depending on dialogue part
- Speech output: pre-recorded speech
- Flight information
- Logging of data

3. Evaluation

3.1. Participants

Sixty-three TNO employees with a representative variation in sex, age, and speech accent participated in the evaluation. The participants were asked to call *"Irene"* and book a flight during a predefined period of 3 weeks in October/November 2002.

3.2. Telephone conversations and data logging

During the mentioned period 110 telephone calls have been made resulting in 905 recorded user utterances. The following (objective) data of the user-system interaction have been automatically logged and measured:

- total duration of the conversation, task duration
- # system turns, # user turns, # task turns
- # recognition errors
- # completed tasks (flight database has been consulted)

3.3. Questionnaire

After completion of the telephone calls, the participants filled out a questionnaire, distributed via Internet, where they were asked to judge the following quality related aspects: Text-To-Speech (TTS) Performance, ASR Performance, Task Ease, Interaction Pace, User Expertise, Expected Behavior, Future Use. In agreement with Walker et al. [9], the questions were all stated in terms of positive dimensions of the system; the participants were asked to indicate the degree to which they agree with these statements on a 5-point Likert scale. In addition, they reported their perceptions as to whether they had completed the task (obtaining flight information) via yes/no answers. Finally, the participants rated the overall quality of the system *"Irene"*, in the form of a 'Grade' (10-point scale). This measure is an extension to the measures proposed by Walker et al. [1]. Sixty-three of the 110 callers completed the questionnaire. The results of the evaluation are based on the data of these 63 participants.

4. Analysis of the measurements and Results

4.1. Objective measures

In the analysis of the data, a distinction has been made between dialogue and task duration. 'Dialogue Duration' refers to the total duration of the telephone conversation, 'Task Duration' to the duration of the conversation part crucial for obtaining flight information. The boolean variable 'Task Success' represents the number of times the flight information database has been successfully consulted (successful = '1' else '0'). The data have been analyzed per participant and per telephone conversation. Table 1 presents the results for the objective measures.

Table 1: Mean results for the objective measures

N	dialogue duration	task duration	task success	# task turns
63	2min:38s	2min:17s	0.7	9.2

# system turns	# user turns	# recognition errors	# utterance errors	utterance error rate
11.1	10.8	2.0	2.6	23.7%

In 42 out of 63 conversations, *"Irene"* provided the caller with flight information (task success = 0.7). In 15 out of the 21 not successfully completed calls the error occurred in the first dialogue part where participants were required to provide the system with a departure city. The number of recognition errors equals the number of times the system had no user input or did not understand the input. These errors have been logged automatically. However, there are cases in which *"Irene"* recognized user input that would not be confirmed by the user later on. We included these cases in the recognition errors resulting in the number of utterance errors (see Table 1). The utterance error rate equals the number of utterance errors divided by the number of user turns.

4.2. Subjective measures

Table 2 presents the mean scores for the subjective measures. As mentioned earlier, the value for 'TTS performance' through 'future use' lie between 1 and 5, for 'grade' between 1 and 10. The value for User Satisfaction is obtained by summing the values for 'TTS performance' through 'grade'. The variable 'Task Success' represents the reports of the participants' perception of task completion. Their reports have been converted to '1' (yes) or '0' (no).

Table 2: Mean results for the subjective measures

TTS performance	ASR performance	task ease	interaction pace	user expertise
4.3	2.9	3.2	3.3	4.1

expected behavior	future use	grade	user satisfaction	subj. task success
3.0	2.7	5.5	28.9	0.7

5. Applying the PARADISE model

As mentioned earlier, the PARADISE posits that a performance function for a dialogue system can be derived by applying multivariate linear regression with user satisfaction as the dependent variable and task success, dialogue efficiency, and dialogue quality measures as independent variables. In agreement to this, we modeled our data using stepwise multivariate linear regression with User Satisfaction as the dependent variable and Task Success, Dialogue Duration, Task Duration, # Task Turns, # System Turns, # User Turns, # Recognition Errors, Utterance Error Rate as independent variables. In Table 3 the results of the analysis are depicted.

Table 3: Results of stepwise regression analysis

$R^2 = 0.63$, Adjusted $R^2 = 0.61$, $F(3,59) = 33.29$, $p < 0.000$

Multiple regression N=63	Beta	St. Error of Beta	p-level
task success	0.39	0.11	0.00
# user turns	0.32	0.10	0.00
# recognition errors	-0.33	0.10	0.00

Task Success, # User Turns and # Recognition Errors appeared to be signifant factors. The other measures did not significantly affect User Satisfaction. The column Beta of Table 3 presents the weights assigned to the factors. The variable Task Success is the subjective measure in stead of the objective counterpart. Results of a study by Walker et al. [10] indicated that perceived task success more accurately predicts User Satisfaction than the objective measure. To investigate this, we performed a stepwise linear regression analysis using the objective Task Success measure as one of the independent variables. The results of this analysis is as follows: $R^2 = 0.60$, Adjusted $R^2 = 0.59$, $F(2,60) = 45.42$, $p < 0.000$. Comparing the results of both analyses it seems that the subjective and objective Task Success measures contribute to User Satisfaction to a similar extent.

In our calculation of User Satisfaction we used the extended definition that included a subjective measure 'grade'. In the definition used by Walker and colleagues this measure was not included. To compare the results, we also analyzed the data using the narrow definition. Table 4 presents the results of the stepwise multivariate regression analysis.

Table 4: Results of regression analysis using Walker's definition of User Satisfaction

$R^2 = 0.48$, Adjusted $R^2 = 0.47$, $F(2,60) = 28.22$, $p < 0.000$

Multiple regression N=63	Beta	St. Error of Beta	p-level
# user turns	0.48	0.09	0.00
# recognition errors	-0.51	0.09	0.00

Using Walker's definition of User satisfaction, only # User Turns and # Recognition Errors are significant factors. Comparing R^2 and Adjusted R^2, it appears that the description of the data is more accurate using the extended definition of User Satisfaction.

Turning back to Table 3, it was found that Task Success, # User Turns, and # Recognition Errors significantly influenced User Satisfaction. The combination of these three factors explained 61% of the variance. Using the weights assigned to the individual factors (Table 3, column Beta), we can now derive the following performance function:

User Satisfaction = 0.39 * N (Task Success) + 0.32 * N (# User Turns) – 0.33 * N (# Recognition Errors) (1)

Where N is a normalization function that allows the weights to be independent of the scale of the individual factors. The performance function enables a quantitative description of the system *"Irene"*. It can predict the User Satisfaction on the basis of objective measures.

6. Discussion and Conclusions

Using the PARADISE framework, we were able to derive a performance function that describes the user satisfaction of the spoken dialogue system *"Irene"* as a weighted linear combination of the factors Task Success, # User Turns, and # Recognition Errors. By improving these factors, the user satisfaction can be effectively increased. We extended the definition of User Satisfaction by including a subjective measure 'grade'. The regression analysis using the extended definition provided a better prediction of User Satisfaction than the analysis based on the narrow definition. We propose to use the extended definition in future evaluations of dialogue systems.

The measure 'grade' may reflect other user experiences that cannot be expressed by TTS Performance, ASR Performance, Task Ease, Interaction Pace, User Expertise, Expected Behavior, or Future Use, but do contribute to user satisfaction to some extent. The definition of User Satisfaction may be further improved by including new measures and fine-tuning existing ones, and/or by changing the relative weights of the measures. This needs further attention.

To take a closer look at the subjective measures, we performed a principal components analysis on these measures. The results indicated that ASR Performance, Task Ease, Expected Behavior, Future Use, and Grade are predominantly described by dimension 1 (factor loadings > 0.80), while TTS Performance and User Expertise are best described by dimension 2 (factor loadings > 0.67). Interaction Pace does not fit very well in either of the two dimensions (factor loadings < 0.52). What the dimensions mean is difficult to say, but it is possible that TTS Performance, User Expertise, and maybe also Interaction Pace have a comparison-related value, i.e. users judge a dialogue system in comparison to another dialogue system. The results of the stepwise multivariate linear regression analyses on the individual subjective measures also indicated a division into two groups. ASR Performance, Task Ease, Expected Behavior, Future Use, and Grade are adequately predicted by one or more

objective measures, whereas TTS Performance, User Expertise, and Interaction Pace cannot be predicted (the analyses did not reveal any significant factors for these three subjective measures).

The derived user satisfaction function in this study concerns a spoken dialogue system using a form of explicit confirmation strategy. Further research should indicate whether the user satisfaction of other spoken dialogue systems using explicit confirmation depends on the same factors as was found for the system *"Irene"*.

7. References

[1] Walker, M., Hindle, D., Fromer, J., Di Fabbrizio, G., & C. Mestel (1997). "Evaluating competing agent strategies for a voice email agent". In *Proceedings of Eurospeech 1997.*

[2] Bernsen, N., & L. Dybkjaer (2000). "A methodology for evaluating spoken language dialogue systems and their components". In *Proceedings of LREC 2000 Athens.*

[3] Levin, L., Bartlog, B., Llitjos, A., Gates, D., Lavie, A., Wallace, D., Watanabe, T., & M. Woszczyna (2000). "Lessons learned from a task-based evaluation of speech-to-speech machine translation". In *Proceedings of LREC 2000 Athens.*

[4] Danieli, M., & E. Gerbino (1995). "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*, p. 34-39.

[5] Hone, K., & R. Graham (2001). "Subjective assessment of speech-system interface usability". In *Proceedings of Eurospeech 2001 Scandinavia.*

[6] Polifroni, J., & S. Seneff (2000). "Galaxy-II as an architecture for spoken dialogue evaluation". In *Proceedings of LREC 2000 Athens.*

[7] Möller, S. (2002). "A new taxonomy for the quality of telephone services based on spoken dialogue systems". In *Proceedings of the 3rd SIGdial Workshop on Discourse and Dialogue, USA-Philadelphia*, p. 142-153.

[8] "Irene", speech-driven flight information service (2002). Vocalibur Language & Speech Technology BV. http://www.vocalibur.com.

[9] Walker, M., Hirschman, L., & J. Aberdeen (2000). "Evaluation for DARPA Communicator spoken dialogue systems". In *Proceedings of LREC 2000 Athens.*

[10] Walker, M., Kamm, C., & J. Boland (2000). "Developing and testing general models of spoken dialogue system performance". In *Proceedings of LREC 2000 Athens.*