Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 74 (2015) 12–36

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

# Interaction Quality: Assessing the quality of ongoing spoken dialog interaction by experts—And how it relates to user satisfaction

Alexander Schmitt [*], Stefan Ultes

*Dialog Systems Group/Institute of Communications Technology, University of Ulm, Albert-Einstein-Allee 43, D-89081 Ulm, Germany*

Received 18 October 2011; received in revised form 28 May 2015; accepted 2 June 2015
Available online 26 June 2015

## Abstract

This study presents a novel expert-based approach to assess the quality of ongoing Spoken Dialog System (SDS) interactions. We call this approach “Interaction Quality” (IQ). It is an objective measure which relies on statistical classification with Support Vector Machines (SVMs). We compare objective expert IQ annotations of ongoing SDS interactions with subjective User Satisfaction (US) ratings and show that IQ and US correlate ($\rho = .66$). Expert annotations obviously mirror the subjective user impression to a great extent while they are, above all, much easier to obtain. The IQ score that quantifies the quality of the interaction is generated using the median score of exchange annotations of several experts. US is tracked in a study with 38 users interacting with an SDS. A large, comprehensive set of domain-independent, automatic interaction parameters is introduced to quantify the interaction at arbitrary dialog exchanges. Furthermore, a manually annotated negative emotion feature is added to the parameter set in order to evaluate the contribution of emotions on the classification of IQ and US. For evaluation we use the CMU Let’s Go bus information system. The model yields a correlation of $\rho = .80$ when classifying IQ scores annotated in field data from the CMU system. Furthermore, the model achieves $\rho = .74$ for predicting US on lab data, and $\rho = .89$ for IQ on lab data. The presented approach outperforms related studies in the field. Only a marginal contribution of the emotion feature to the performance can be observed, implying that US is not influenced by visible emotions. We analyze causalities and correlations between the interaction parameters and the target variables US/IQ and identify relevant predictors. With the presented paradigm, critical dialogs can be found; once deployed as an online monitoring technique, this paradigm could render SDSs more user friendly and improve user acceptance.

*Keywords:* Spoken dialog systems; User satisfaction; Online monitoring; Paradise; Emotional state; Interaction parameters

## 1. Introduction

Let us assume a customer of a railroad company comes up to a counter to purchase a railroad ticket. The clerk enquires about the departure and the destination, the time of departure and further details required to issue the ticket. In this scenario, both the clerk and the customer can be friendly and polite, or emotionless, bored, harsh and stressed out. The clerk may be short-spoken or overly helpful. The interaction may be influenced by environmental factors, such as the noise level in the ticket hall or loudspeaker announcements. It may further be influenced by subjective factors, such as the personal attitude of the customer towards railroad clerks, or less obvious external factors, e.g., the fact that one of the dialog partners just had a strong argument with his significant other.

If we would ask the customer how he would judge his satisfaction with the service on a continuous scale—even during the interaction—he would certainly be able to do so. Furthermore, an ideal clerk would be capable of

[*] Corresponding author. Tel.: +49 731 5026253.
*E-mail addresses:* alexander.schmitt@uni-ulm.de (A. Schmitt), stefan.ultes@uni-ulm.de (S. Ultes).

http://dx.doi.org/10.1016/j.specom.2015.06.003
0167-6393/© 2015 Elsevier B.V. All rights reserved.

roughly estimating the customer's degree of satisfaction at all times and he would be intuitively able to adapt his strategy to improve his service. For his assessment of user satisfaction, the clerk could include a large number of information sources, such as:

- the emotional state of the user, derived from facial expressions together with acoustic and lexical information,
- discourse information, i.e., the verbal information exchanged by both dialog partners,
- the probability of task completion, i.e., how likely the concern of the customer can be handled,
- context information, e.g., the standing time the customer had to wait in line,
- the user's experience with the task,
- social information about the user, e.g., estimated personality, age, educational background, gender.

Among many other fields of application, Spoken Dialog Systems (SDS) are deployed to serve customers over telephone and are being used specifically for such a task. Nevertheless, modern SDS are not able to self-monitor the discourse like a human would be. Instead, they are mainly static in terms of what to prompt to the user and static in terms of how to treat the user notwithstanding the course of the previous conversation. Moreover, speech recognition and language understanding are error-prone and both are—with few exceptions—not able to reliably cover free user input. This frequently leads to critical dialog situations where the interaction between system and user is about to fail and where task success is not achieved. Miscommunication caused by poor system performance and design as well as false user behavior and wrong user expectations leave behind dissatisfied users. In the worst case, this even leads to an abortion of the task.

The rising complexity of SDSs necessitates the development of innovative techniques to make future systems interaction-aware and enable them to detect critical dialog scenarios. With this knowledge, a dialog system would be capable of handling the situation just like a real, customer-friendly clerk would. Here, a SDS could react by *adapting the dialog strategy* when critical situations are automatically detected, cf. Langkilde et al. (1999), Levin et al. (2000), Walker et al. (2002), Hastie et al. (2002), Levin and Pieraccini (2006), Herm et al. (2008), Schmitt et al. (2008), Zgorzelski et al. (2010), Schmitt et al. (2010c), Ultes et al. (2011, 2012, 2014a,b). Another option would be the *escalation to a live agent* who finishes the task jointly with the user. This would spare the user a long-lasting, non-target aimed interaction with a failing SDS.

A solution has to be found that permits such a monitoring of the ongoing interaction between system and user which also enables an estimation of the quality covering the emotional state and interaction patterns. While, in our view, the primary intention of these models is the deployment for detecting low quality in dialogs online during the interaction, they could likewise be employed to spot poor dialog design and estimate the overall quality of a system, similarly to the PARADISE approach such as in Walker et al. (1997, 1998). In contrast to PARADISE, the models would be more detailed and fine-grained.

In Schmitt et al. (2011), we proposed the statistical modeling of Interaction Quality (IQ) on the exchange level, i.e., within ongoing dialogs. Objective expert annotators provided quality scores that served as input variables for the model. It could, however, not be shown how objective IQ scores correlate with actual User Satisfaction (US).

In this contribution, we approach the problem in its entirety. This includes a detailed and thorough discussion of the Interaction Quality paradigm. Expert ratings acquired with clear guidelines are used to mimic end user satisfaction at lower cost and ease of completion. Furthermore, the corresponding corpora are presented along with a detailed description of their exchange level features. Finally, the correlation between Interaction Quality and User Satisfaction are explicitly measured and analyzed.

This article is organized as follows. Initially, related work is addressed and discussed in Section 2. The term "Interaction Quality" is then introduced in Section 3, where also the choice of an appropriate target variable representing "quality" is described. A corpus based on dialogs from a real-life SDS with users interacting in the field annotated by human experts is presented in Section 4. For comparing the congruency and correlation of expert annotations with subjective user impressions, a user study has been conducted which is presented in Section 5. In Section 6, a comprehensive set of interaction parameters in the style of the ITU Supplement 24 to P Series "Parameters is introduced describing the interaction with spoken dialog systems" (ITU, 2005). The contribution of the input variables from the different system modules on predicting IQ and US is evaluated using Support Vector Machines (SVM) and discussed in Section 7. Finally, the results are discussed in Section 8. A conclusion is drawn in Section 9.

## 2. Related work in estimating user satisfaction for SDS

Previous work addressed in this article can be basically classified into two groups: "dialog level" and "exchange level" user satisfaction modeling. "Dialog level" refers to approaches that model user satisfaction with the aim to determine an average overall user satisfaction score at the end of a dialog. The intention of these approaches is to *evaluate a specific system* and to compare system versions after modifications. Input variables in that field are dialog-wide performance scores that are mapped to a specific user impression. "Exchange level" approaches, on the other hand, aim to estimate satisfaction at arbitrary points in a specific dialog. Input variables are parameters that represent the interaction at or up to a specific system-user-exchange. Their basic purpose is to understand which factors influence satisfaction. Furthermore, these