Multimodal dialogue system evaluation: a case study applying usability standards

Andrei Malchanau, Volha Petukhova and Harry Bunt

Abstract This paper presents an approach to the evaluation of multimodal dialogue systems, applying usability metrics defined in ISO standards. Users' perceptions of effectiveness, efficiency and satisfaction were correlated with various performance metrics and with interaction parameters derived from system logfiles and reference annotations. A preliminary 110-items questionnaire was designed and rated by respondents to measure their agreement on usability concepts. Eight main factors were observed to have impact on usability perception: task completion and quality, robustness, learnability, flexibility, likeability, ease of use and usefulness (value) of an application. As a result, an internally consistent and reliable questionnaire with 32 items (Cronbach's alpha of 0.87) was extracted. This questionnaire was used to evaluate the Virtual Negotiation Coaching system for metacognitive skills training in a multi-issue bargaining setting. The observed correlations between usability perception and derived performance metrics and interaction parameters suggest that the overall system usability is determined by the quality of agreements reached, by the robustness and flexibility of the interaction, and by the quality of system responses.

1 Introduction

Modern digital services and devices get more and more interconnected and integrated in everyday human activities. They often come each with their own interface, however, which users need to learn how to communicate with or to control. Multimodal natural-language based dialogue is increasingly becoming a feasible and attractive human-machine interface which can be used to provide a universal, accountable and personalized form of access to a variety of products and contents. Such interfaces offer a mode of interaction that has certain similarities with natural human communication, in using a range of input and output modalities that people normally employ in communication, achieving a certain level of 'digital immersion' which boosts user acceptance and enriches user experience. As a part of the inter-

Andrei Malchanau and Volha Petukhova
Spoken Language Systems Group, Saarland University e-mail: {andrei.malchanau;v.petukhova}@lsv.uni-saarland.de

Harry Bunt
Tilburg Center for Communication and Cognition, Tilburg University e-mail: harry.bunt@uvt.nl

active application design, evaluations are performed in order to assess the success of the developed solutions. Evaluation results serve to inform designers about the system's functional and non-functional deficiencies.

Dialogue systems are often exposed to user-based evaluation. This is commonly done by asking users to fill in a questionnaire after interacting with the system. It is still largely an open question which parameters should be taken into account when designing a satisfaction questionnaire, and which of these may correlate well with user satisfaction. Qualitative and quantitative measures are often automatically computed from test interactions with real or simulated users. Most existing evaluation metrics are designed for task-oriented information seeking spoken dialogue systems and do not apply well to complex multimodal interactions. In this paper we propose to assess multimodal dialogue system performance by relating various performance metrics, interaction parameters, and subjective perception of *usability* factors as defined by the ISO 9241-11 and ISO/IEC 9126-4 standards. This enables usability quantification in a meaningful and systematic way.

This paper is structured as follows. Section 2 discusses existing approaches to the evaluation of interactive conversational systems. Section 3 presents the ISO 9241-11 usability definition and metrics for effectiveness, efficiency and satisfaction. We adapt these metrics to the multimodal dialogue system evaluation task by specifying factors that impact usability perception by its users. In Section 4 we describe experiments and report results evaluating the Virtual Negotiation Coach application. Section 5 summarises our findings and outlines future research.

2 Related work

Several dialogue system evaluation approaches have been proposed in the past. PARADISE, one of the most widely-used evaluation models [1], aims at predicting user global satisfaction given a set of parameters related to task success and dialogue costs. Satisfaction is calculated as the arithmetic mean of nine user judgments on different quality aspects rated on 5-point Likert scales. Subsequently, the relation between task success and dialogue cost parameters and the mean human judgment is estimated by means of a multivariate linear regression analysis.

Another approach is to evaluate a dialogue system on the basis of test interactions substituting human users by computer agents that emulate user behaviour, see e.g. [2]. The various types of users and system factors can be systematically manipulated, e.g. using interactive, dialogue task and error recovery strategies.

As for system performance metrics and interaction parameters, several sets have been recommended for spoken dialogue system evaluation ranging from 7 parameters defined in [14] to 52 in [15] related to the entire dialogue (duration, response delay, number of turns), to meta-communication strategies (number of help requests, correction turns), to the systems cooperativity (contextual appropriateness of system utterances), to the task which can be carried out with the help of the system (task success, solution quality), as well as to the speech input performance of the system (word error rate, understanding error rate).

When evaluating an interactive application, real user judgments provide valuable insights into how well the application meets user expectations and needs. One of the methods to measure users' attitudes is to observe their behaviour and establish links between their emotions and actions [18]. Given the current technical possibilities, the tracking and analysis of large amounts of logged user-generated multimodal data has become feasible [17]. For instance, gaze re-direction, body movements, facial muscle contraction, skin conductivity and heart rate variance may serve as a source of information for analysing a user's affective state.

The most common practice is to solicit user judgments on different system quality aspects with the help of a questionnaire. The absence of standard questionnaires for dialogue systems evaluation makes it difficult to compare the results from different studies, and the various existing questionnaires exhibit great differences:

- The PARADISE questionnaire has nine user satisfaction related questions [12].
- The Subjective Assessment of Speech System Interfaces (SASSI) questionnaire contains 44 statements rated by respondents on 7-point Likert scales [13].
- The Godspeed questionnaire comprising 24 bipolar adjective pairs (e.g. fake-natural, inert-interactive, etc.) related to (1) anthropomorphism, (2) animacy, (3) likeability, (4) perceived intelligence and (5) perceived safety to evaluate human-robot interactions on 5-point Likert scales [16].
- The REVU (Report on the Enjoyment, Value, and Usability) questionnaire was developed to evaluate interactive tutoring applications and comprises 53 statements rated on 5-point Likert scales divided into three parts: OVERALL, NL (Natural Language), and IT (Intelligent Tutor) [3].
- The Questionnaire for User Interface Satisfaction (QUIS[1], [8]) measures satisfaction related (1) overall user reaction, (2) screen, (3) terminology and system information, (4) learnability, (5) system capabilities, (6) technical manuals and on-line help, (7) on-line tutorials, (8) multimedia, (9) teleconferencing, and (10) software installation. A short 6-dimensional form contains 41 statements rated on 9-point Likert scales, a long one has 122 ratings used for diagnostic situations.

The QUIS questionnaire is widely used and is considered as de-facto standard for user satisfaction assessment when performing usability studies. The QUIS forms can be customized by selecting evaluation aspects relevant for a specific application and use case, as we will show in the next sections when evaluating a multimodal dialogue system.

3 Usability definition

It is a common practice to evaluate an interactive system and its interface using a number of observable and quantifiable metrics for effectiveness, efficiency and satisfaction - see the ISO 9241-11 and ISO/IEC 9126-4 standards.

Task completion and the accuracy with which users achieve their goals are associated with the system's *effectiveness*. Task completion is calculated as the proportion

[1] Version 7.0 is available http://www.lap.umd.edu/QUIS/index.html

of successfully completed tasks given the total number of tasks. To measure success of information retrieval tasks in information seeking dialogues, Attribute Value Matrix (AVM) metrics are used as proposed in PARADISE. In tutoring interactive applications, the task completion rate will depend on the system's ability to provide meaningful feedback [3]. In the next section we will define effectiveness metrics for our negotiation training use case.

Efficiency is associated with the effort that users spend to perform specified tasks and is often correlated with temporal and duration properties of the interaction, e.g. number of turns, pace, reaction times, etc. Measures of efficiency associated with user's cognitive costs relate to [19]:

- *robustness*, referring to the level of support provided to the user in determining achievement and assessment of goals; is related to observability, recoverability, responsiveness and task conformance;
- *learnability*, referring to the ease with which new users can begin effective interaction and then to attain a maximal level of performance; is related to predictability, familiarity and consistency; and
- *flexibility*, referring to the multiplicity of ways in which the user and the system can communicate; is related to initiative, task substitutivity and customisability.

Satisfaction is concerned with user attitudes associated with the product use, and is often assessed with the help of questionnaires. Satisfaction is measured at the task and test levels. Popular post-task questionnaires are After-Scenario Questionnaire (ASQ, [4]), NASA Task Load Index (TLX)[2] and Single Ease Question (SEQ)[3]. Satisfaction at the test level serves to measure users' impression of the overall ease of use of the system being tested.

In order to develop a reliable questionnaire for assessing user perception of a multimodal dialogue system usability we conducted an online study. QUIS 7.0 served as the basis for respondents to make their selection of aspects they think are important for them when evaluating a multimodal dialogue system. QUIS provides a useful decomposition of the usability concept into several dimensions (factors), enabling a clear mapping of system performance to distinctive usability perception aspects, with the advantage of being able to assess the impact of different items on usability perception instead of simply summing up or averaging to compute an overall satisfaction score (as e.g. in PARADISE or with the System Usability Scale, SUS [5]). Adapting the QUIS questionnaire for the purposes of multimodal dialogue system evaluation, we considered factors assessed by the SAASI and Godspeed questionnaires. Previous studies showed that evaluative adjectives, bipolar adjective pairs and specific evaluative statements appeared to be more accurate than global satisfaction questions and were the most preferred forms for respondents [8, 20]. In our study, 36 evaluative adjectives, 40 bipolar adjective pairs, and 34 evaluative statements were ranked on 5-point Likert scales by 73 respondents, from which 69.6% considered themselves as dialogue researchers or related, and all respondents used dialogue systems at least once in their life. The study showed that important aspects

[2] https://humansystems.arc.nasa.gov/groups/TLX/

[3] A 7-point rated question to assess how difficult users find a task, see https://measuringu.com/single-question/

related to user satisfaction are concerned with *task completion, task quality, robustness, learnability, flexibility, likeability, ease of use* and *usefulness/value* of the application. We adopted the QUIS 7.0 structure and populated it with 32 selected items rated the highest (> 4.0 points with standard deviation < 1.0) in the online study. The resulting questionnaire[4] has six dimensions measuring (1) overall reaction, (2) perceived effectiveness, (3)system capabilities, (4) learnability, (5) visuals/displays and animacy, (6) real-time feedback. The questionnaire allows to evaluate a system's functionality related to multimodality (items in dimension 3 and 5) and tutoring capabilities (dimension 6). The questionnaire is used to perform user-based evaluation as reported in Section 4.3 and is evaluated on internal consistency reliability.

4 Experimental set-up

The use case considered in this study is concerned with the evaluation of a multimodal Intelligent Tutoring System designed to train metacognitive skills in a multi-issue bargaining setting - the Virtual Negotiation Coach (VNC). The system's goal is to make a negotiator aware of and reason about his negotiation behaviour and negotiation strategies and those of his opponent.

4.1 Context and scenario

The specific setting considered in this study involved a multi-issue bargaining scenario about anti-smoking legislation passed in the City of Athens. The negotiated regulations were concerned with four main *issues*: (1) smoke-free public areas (smoking ban scope); (2) tobacco tax increase (taxation); (3) effective anti-smoking campaign programs (campaign); and (4) enforcement policy and police involvement (enforcement), see Figure 1. Each of these issues involves four to five most important negotiation *values* with preferences representing negotiation positions, i.e. preference profiles. The strength of preferences was communicated to the negotiators through colours. Brighter orange colours indicated increasingly negative options; brighter blue colours increasingly positive options.

In our evaluation experiment, 28 participants aged 25-45, professional politicians or governmental workers, were interacting with the VNC (see Section 4.2) for an hour. Nine negotiation scenarios were used, based on different negotiators preference profiles. Users ('trainees') were assigned a City Councilor role and a random scenario. All sessions were recorded and numerous interaction parameters logged.

The trainees' task was to negotiate with a Small Business Representative (system) an agreement which assigns exactly one value to each issue, exchanging and eliciting offers concerning $\langle ISSUE; VALUE \rangle$ options. The task is considered as completed when for all four issues an agreement is reached. Negotiators were allowed to withdraw and/or re-negotiate previously made agreements within one session. The negotiation task quality is measured by the quality of the agreements reached.

[4] The system demo video and usability questionnaire is available online at https://www.lsv.uni-saarland.de/index.php?id=72

In integrative bargaining this can be determined by the number of *Pareto optimal outcomes*[5].

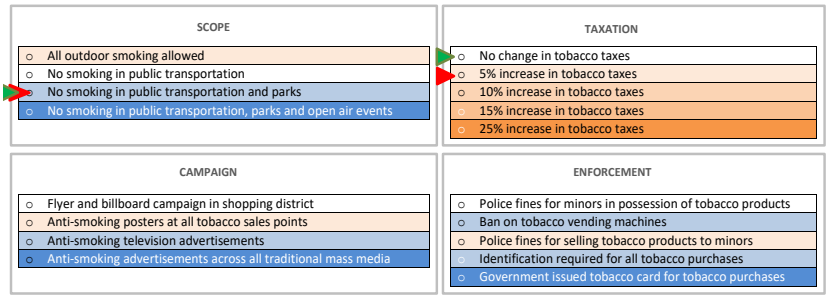

Fig. 1 Preference card: example of values in four negotiated issues presented in colours. Partners' offers visualized with red arrow (system) and green one (user).

The negotiation success is influenced by the negotiators' strategies. For integrative negotiations, where the negotiators strive for a balance between cooperation and competition, two main negotiation strategies are observed: cooperative and non-cooperative. Cooperative negotiators share information about their preferences and attempt to find mutually beneficial agreements. They are not engaged in positional bargaining[6] tactics, instead, they try to find issues where a trade-off is possible. Non-cooperative negotiators prefer to withhold their (true) preferences and focus on positional bargaining, rarely asking for or ignoring an opponent's preferences. They threaten to end the negotiation or make very small concessions. In our experiments, we calculated the *cooperativeness level* as the number of cooperative actions given the total number of task-related negotiation actions performed. A third task success metric is related to the the *number of negative deals*, i.e. dispreferred agreements on bright 'orange' options as shown in Figure 1.

4.2 Virtual Negotiation Coach

We designed the Virtual Negotiation Coach (VNC), a multimodal interactive system with the functionality described in the scenario section. The VNC gets a speech signal, recognizes and interprets it, and generates multimodal actions as response, i.e. speech and gestures of a virtual negotiator and positive and negative visual feedback of a virtual tutor. Figure 2 shows the VNC architecture and processing workflows.

Speech signals are recorded by multiple devices: wearable microphones or headsets, and an all-around microphone placed between participants. The speech signals serve as input for Automatic Speech Recognition (ASR). The Kaldi-based ASR component incorporates acoustic and language models developed using 759 hours

[5] Pareto optimality reflects a state of affairs when there is no alternative state that would make any partner better off without making anyone worse off.

[6] Positional bargaining involves holding on to a fixed preferences set regardless of the interests of others.

of data from various available data sources[7]. The collected 'smoking ban' in-domain data [9] is used for language model adaptation. The ASR performance is measured at 34.4% Word Error Rate (WER) [6].

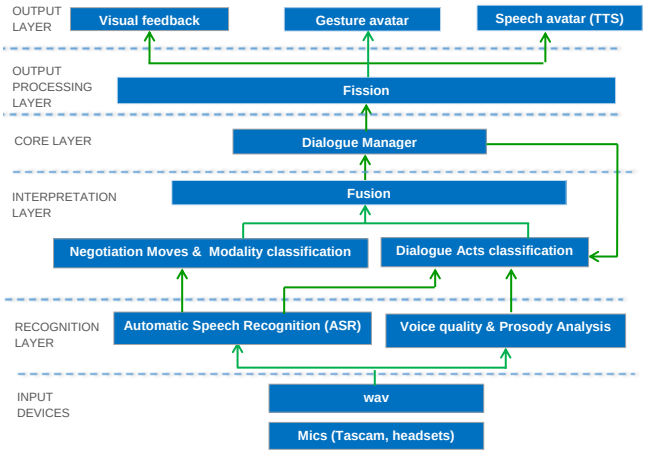

Fig. 2 Architecture of the Virtual Negotiation Coach system.

The ASR output is used for lexical, syntactic, and semantic analysis to perform negotiation moves, modality and dialogue act (DA) classification. Negotiation moves specify events and their arguments of $NegotiationMove(ISSUE;VALUE)$ type [21]. The sequence learning Conditional Random Field (CRF) models were trained to predict three types of classes: negotiation move, issue, preference value. A 10-fold cross-validation yielded the F-score of 0.7 on average. The modality classifiers (Support Vector Machine, SVM) show accuracies in the range between 73.3 and 82.6% [23]. For the DA recognition, SVM-based classifiers were applied with F-scores ranging between 0.83 and 0.86 [7]. Information related to the dialogue history has been used to ensure context-dependent interpretation of dialogue acts. Additionally, the trainee has a choice to select options using a graphical interface as depicted in Figure 2. As task progress support, partner offers and possible agreements are visualized with red arrow (system) and green one (user).

The Dialogue Manager (DM) is designed as a set of processes ('threads') that receive data, update the information state and generate output [22]. Integrated into the DM is the Negotiation Task Agent (NTA), which interprets and produces negotiation actions based on the estimation of the partner's preferences and goals and adjusts its strategy according to the perceived level of the partner's cooperativeness. It begins neutrally, requesting the partner's preferences; reacts with a cooperative negotiation move if it believes the partner is cooperative, and non-cooperatively if the partner's actions are interpreted as unconditionally non-cooperative, see [21]. The system reasons about the overall state of the negotiation task, and attempts to identify the best negotiation move for the next action. The DM computes (1) the system's counter-move, and (2) feedback sharing the system's beliefs about the user's

[7] Examples of resources are: the Wall Street Journal WSJ0 corpus, HUB4 News Broadcast data and the VoxForge corpus.

preferences and the user's negotiation strategy. Additionally, the DM takes care of actions concerning contact and social obligations management, as well as elaborate recovery and error handling actions.

The system includes Fusion and Fission components. The Fusion module currently fuses interpretations from two modules obtaining full semantic representations of user speech contributions. Given the dialogue acts provided by the Dialogue Manager, Fission generates system responses splitting content into different modalities, such as Avatar and Voice (TTS) for negotiation actions, and visual feedback for tutoring actions. The latter includes a representation of the negotiators' current cooperativeness, visualized by happy and sad face emoticons. At the end of each negotiation session, summative feedback is generated specifying the number of points gained or lost for each partner, the number of negative agreements, and the Pareto optimality of the reached agreements. This type of feedback accumulates across multiple consecutive negotiation rounds.

4.3 User-based evaluation: perception vs performance

The VNC system was evaluated measuring usability in terms of effectiveness, efficiency and satisfaction. Previous research suggests that there are differences in perceived and actual performance [10]. Performance and perception scores are correlated, but they are different usability metrics and both need to be considered when conducting quantitative usability studies. In our design, subjective perception of effectiveness, efficiency and satisfaction were correlated with various performance metrics and interaction parameters to assess their impact on the qualitative usability properties. We computed bi-variate correlations to determine possible factors impacting user perception of system usability and the derived performance metrics and interaction parameters from logged and annotated evaluation sessions.

The perceptive assessments come from the user satisfaction judgments on different aspects after interacting with the system. The questionnaire designed for this purpose is, first, evaluated on internally consistency and reliability measuring Cronbach's alpha. The internal consistency of the factors (dimensions) were: (1) overall reaction, alpha=0.71; (2) perceived effectiveness, alpha=0.74; (3) system capabilities, alpha=0.73; (4) learnability, alpha=0.72; (5) visuals and animacy, alpha=0.75; and (6) real-time feedback, alpha=0.82. All alpha values were > 0.7, so we can conclude that all factors have sufficient internal consistency reliability.

As part of the user-based evaluation, users were asked to provide an overall rating of the system that they interacted with using six bipolar negative-positive adjective pairs such as frustrating-satisfying, difficult-easy, inefficient-efficient, unnatural-natural, rigid-flexible and useless-useful rated on 5-points Likert scales. Correlations between the mean overall satisfaction (3.64) and each of the other factors was measured as follows: effectiveness, $r = .79$; system capabilities, $r = .59$; learnability, $r = .87$; visuals and animacy, $r = .76$; and feedback, $r = .48$. Thus, users appreciate when the system effectively meets their goals and expectations and supports them in completing their tasks, is easy to learn how to interact with and offers flexible input and output processing and generation in multiple modalities.

As performance metrics, system and user performance related to task completion rate[8] and its quality[9] were computed. We also compared system negotiation performance with human performance on the number of agreements reached, the ability to find Pareto optimal outcomes, the degree of cooperativeness, and the number of negative outcomes[10]. It was found that participants reached a lower number of agreements when negotiating with the system than when negotiating with each other, 66% vs 78%. Participants made a similar number of Pareto optimal agreements (about 60%). Human participants show a higher level of cooperativity when interacting with the system, i.e. 51% of the actions are perceived as cooperative. This may mean that humans were more competitive when interacting with each other. A lower number of negative deals was observed for human-agent pairs, 21% vs 16%. Users perceived their interaction with the system as effective when they managed to complete their tasks successfully reaching Pareto optimal agreements by performing cooperative actions but avoiding excessive concessions. No significant differences in this respect were observed between human-human and human-system interactions.

As for efficiency, we assessed temporal and duration dialogue parameters, e.g. time elapsed and number of system and/or user turns to complete the task (or subtask) and the interaction as a whole. We also measured the system response time, the silence duration after the user completed his utterance and before the system responded. Weak negative correlation effects have been found between user perceived efficiency and system response delay, meaning users generally found the system reaction and the interaction pace too slow. Dialogue quality is often assessed measuring word and sentence error rates [1, 2] and turn correction ratio [11]. Many designers, however, noticed that it is not so much how many errors the system makes that contributes to its quality, but rather the system's ability to recognize errors and recover from them. This contributes to the perceived system robustness and is appreciated by the users. Users value if they can easily identify and recover from their own mistakes. All system's processing results were visualized to the user in a separate window, which contributes to the system observability. System's and user's applied repair and recovery strategies are evaluated by two expert annotators and agreement was measured in terms of kappa. Repairs were estimated as the number of corrected segments, recoveries as the number of regained utterances which were partially failed at recognition and understanding, see also [11]. While most annotators agreed that repair strategies were applied adequately, longer dialogue sessions due to frequent clarifications seem to be undesirable.

The VNC is evaluated to be relatively easy to interact with (4.2 Likert points). However, users found an instruction round with a human tutor prior to the interaction

[8] We consider the overall negotiation task as completed if parties agreed on all four issues or parties came to the conclusion that it is impossible to reach any agreement.

[9] Overall task quality was computed in terms of number of *reward points* the trainee gets at the end of each negotiation round and summing up over multiple repeated rounds; and *Pareto optimality* (see footnote 5).

[10] We considered negative deals as flawed negotiation action, i.e. the sum of all reached agreements resulted in an overall negative value meaning that the trainee made too many concessions and selected mostly dispreferred bright 'orange' options (see Figure 1).

useful. Most users were confident enough to interact with the system of their own, some of them however found the system too complex and experienced difficulties in understanding certain concepts/actions. A performance metric which was found to negatively correlate with system learnability is user response delay, the silence duration after the system completed its utterance and the user proposed relevant dialogue continuation. Nevertheless, the vast majority of users learned how to interact with the system and complete their tasks successfully in the consecutive rounds. We observed a steady decline in user response delays from round to round.[11]

Users appreciated the system's flexibility. The system offered the option to select continuation task actions using a graphical interface on a tablet in case the system processing failed entirely. The use of concurrent multiple modalities was positively evaluated by the users. It was always possible for users to take initiative in starting, continuing and wrapping up the interaction, or leave these decisions to the system. At each point of interaction, both the user and the system were able to re-negotiate any previously made agreement.[12] As overall satisfaction, the inter-

Table 1 Summary of evaluation metrics and obtained results in terms of correlations between subjective perceived system properties and actions, and objective performance metrics (R stands for Pearson coefficient; * = statistically significant ($p < .05$)

Usability metric	Perception	Performance		R
	Assessment	Metric/parameter	Value	
effectiveness (task completeness)	mean rating score effectiveness 4.08	Task completion rate[8]; in %	66.0	.86*
		Reward points[9]; mean, max.10	5.2	.19
effectiveness (task quality)		User's Action Error Rate (UAER, in %)[10]	16.0	.27*
		Pareto optimality[9]; mean, between 0 and 1	0.86	.28*
		Cooperativeness rate; mean, in %	51.0	.39*
efficiency (overall)	mean rating score efficiency 4.28	System Response Delay (SRD); mean, in ms	243	-.16
		Interaction pace; utterance/min	9.98	-.18
		Dialogue duration; in min	9:37	-.21
efficiency (learnability)	3.3 (mean)	Dialogue duration average, in number of turns	56.2	-.35*
		User Response Delay (URD); mean, in ms	267	-.34*
efficiency (robustness)	3.2 (mean)	System Recovery Strategies (SRS) correctly activated (Cohen's κ)	0.89	.48*
		User Recovery Strategies (URS) correctly recognized (Cohen's κ)	0.87	.45*
efficiency (flexibility)	3.8 (mean)	Proportion spoken/on-screen actions mean, in % per dialogue	4.3	.67*
satisfaction (overall)	aggregated per user ranging between 40 and 78	ASR Word Error rate; WER, in %	22.5	-.29*
		Negotiation moves recognition accuracy, in %	65.3	.39*
		Dialogue Act Recognition; accuracy, in %	87.8	.44*
		Correct responses (CR)[14] relative frequency, in %	57.6	.43*
		Appropriate responses (AR)[13] relative frequency, in %	42.4	.29*

action was judged to be satisfying, rather reliable and useful, however, less natural (2.76 Likert points). The later is largely attributed to rather tedious multimodal generation and poor avatar performance. System actions were judged by expert annota-

[11] For now, this is only the general observation and the metric will be taken into consideration in future test-retest experiments.

[12] Performance metrics related to initiative and task substituitivity aspects and their impact on the perceived usability will be an issue for the future research.

tors as appropriate[13], correct[14] and easy to interpret. Other module-specific parameters reflecting widely used metrics computed by comparing system performance with reference annotations were various types of error rates, accuracy, and κ scores measuring agreement between the system performance and human annotations of the evaluation sessions. Recognition and interpretation mistakes turned out to have moderate negative effects on the user satisfaction. Table 1 summarizes the results.

Satisfaction questionnaires were constructed in such a way that, along with overall user satisfaction, we could also evaluate the system's tutoring performance. Participants indicated that system feedback was valuable and supportive. However, they expected more visual real-time feedback and more explicit summative feedback on their learning progress. Most respondents think that the system presents an interesting training skills application and would use it as a part of their training routine.

5 Conclusions and future research

We have presented an approach to multimodal dialogue system evaluation according to the available ISO standards on usability and qualitative metrics for effectiveness, efficiency and satisfaction. A prototype questionnaire was designed, based on established measures and best practices for the usability evaluation of interactive systems and interfaces. Potential questionnaire items were rated by respondents. Eight factors were selected as having a major impact on the perceived usability of a multimodal dialogue system and related to task success, task quality, robustness, learnability, flexibility, likeability, ease of use and usefulness (value). Performance metrics and interaction parameters were either automatically derived from logfiles or computed using reference annotations. Perception and performance were correlated to be able to quantify usability. It was observed that the overall system usability is determined most by the user satisfaction with the task quality, by the robustness and flexibility of the interaction, and by the quality of system responses.

Further efforts will be undertaken to refine performance metrics and compute additional interaction parameters. We also plan to incorporate data coming from modern tracking and sensing devices to compute the affective state of the user during interaction with the system, as well as his level of motivation and engagement.

References

1. Walker, M.A., Litman, D. J., Kamm, C. A., Abella, A.: PARADISE: A framework for evaluating spoken dialogue agents. In: Proceedings of the 8th conference on European Chapter of the Association for Computational Linguistics, pp. 271–280 (1997)
2. López-Cózar, R., Callejas, Z. and McTear, M.: Testing the performance of spoken dialogue systems by means of an artificially simulated user. Artificial Intelligence Review, vol. 26 (4), pp. 291-323, Springer (2006)

[13] System action is appropriate given the context if it introduces or continues a repair strategy.

[14] System action is considered as correct if it addresses the user's actions as intended and expected. These actions exclude recovery actions and error handling.

3. Dzikovska, M., Moore, J., Steinhauser, N., Campbell, G.: Exploring user satisfaction in a tutorial dialogue system. In: Proceedings of the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial 2011), pp. 162-172 (2011)
4. Lewis, J. R.: Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ. ACM Sigchi Bulletin, 23(1), pp. 78-81 (1991)
5. Brooke, J. SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), pp. 4-7 (1996)
6. Singh, M., Oualil, Y., Klakow, D.: Approximated and domain-adapted LSTM language models for first-pass decoding in speech recognition. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH), Stockholm, Sweden (2017)
7. Amanova, D., Petukhova, V., Klakow, D.: Creating Annotated Dialogue Resources: Cross-Domain Dialogue Act Classification. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2016), ELRA, Paris (2016)
8. Chin, J.P., Diehl, V.A. Norman, K.L.: Development of an instrument measuring user satisfaction of the human-computer interface. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 213-218, ACM (1988)
9. Petukhova, V., Stevens, C.A., de Weerd, H., Taatgen, N., and Cnossen, F., Malchanau, A.: Modelling Multi-Issue Bargaining Dialogues: Data Collection, Annotation Design and Corpus. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2016), ELRA, Paris (2016)
10. Nielsen, J.: User Satisfaction vs. Performance Metrics. Nielsen Norman Group (2012)
11. Danieli, M., Gerbino, E.: 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, Vol. 16, pp. 34-39 (1995)
12. Walker, M., Kamm, C., Litman, D.: Towards developing general models of usability with PARADISE. Natural Language Engineering 6.3-4, pp. 363-377 (2000)
13. Hone, K.S., Graham, R.: Subjective assessment of speech-system interface usability. In: Proceedings of the 7th European Conference on Speech Communication and Technology. (2001)
14. Fraser, N.: Assessment of Interactive Systems. In: Handbook on Standards and Resources for Spoken Language Systems (D. Gibbon, R. Moore and R. Winski, eds.), pp. 564–615, Mouton de Gruyter, Berlin (1997)
15. Möller, S.: Quality of telephone-based spoken dialogue systems. Springer Science & Business Media (2004)
16. Bartneck, C., Kulić, D., Croft, E. and Zoghbi, S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics, 1(1), pp.71-81 (2009)
17. Linek, S. B., Marte, B., Albert, D.: The differential use and effective combination of questionnaires and logfiles. In Computer-based Knowledge and Skill Assessment and Feedback in Learning settings (CAF), Proceedings of the ICL (2008)
18. Kooijmans, T., Kanda, T., Bartneck, C., Ishiguro, H., Hagita, N.: Accelerating Robot Development Through Integral Analysis of HumanRobot Interaction. IEEE Transactions on Robotics, 23(5), pp.1001-1012 (2007)
19. Dix, A.: Human-computer interaction. In: Encyclopedia of database systems, pp. 1327-1331, Springer US (2009)
20. Root, R. W., Draper, S.: Questionnaires as a software evaluation tool. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 83-87, ACM (1983)
21. Petukhova, V., Bunt, H., Malchanau, A.: Computing negotiation update semantics in multi-issue bargaining dialogues. In: Proceedings of the SemDial 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue, Germany (2017)
22. Malchanau, A., Petukhova., Bunt., Klakow D.: Multidimensional dialogue management for tutoring systems. In: Proceedings of the 7th Language and Technology Conference (LTC 2015), Poznan, Poland (2015)
23. Lapina, V. and Petukhova, V.: Classification of Modal Meaning in Negotiation Dialogues. In: Proceedings of the 13th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-13), pp. 59-70, Montpellier, France (2017)