

First Insight into Quality-Adaptive Dialogue

Stefan Ultes, Hüseyin Dikme, and Wolfgang Minker

Institute for Communications Engineering, University of Ulm

Albert-Einstein-Allee 43, Ulm, Germany

{stefan.ultes, huseyin.dikme, wolfgang.minker}@uni-ulm.de

Abstract

While Spoken Dialogue Systems have gained in importance in recent years, most systems applied in the real world are still static and error-prone. To overcome this, the user is put into the focus of dialogue management. Hence, an approach for adapting the course of the dialogue to Interaction Quality, an objective variant of user satisfaction, is presented in this work. In general, rendering the dialogue adaptive to user satisfaction enables the dialogue system to improve the course of the dialogue and to handle problematic situations better. In this contribution, we present a pilot study of quality-adaptive dialogue. By selecting the confirmation strategy based on the current IQ value, the course of the dialogue is adapted in order to improve the overall user experience. In a user experiment comparing three different confirmation strategies in a train booking domain, the adaptive strategy performs successful and is among the two best rated strategies based on the overall user experience.

Keywords: adaptive spoken dialogue system, user satisfaction, grounding, user study

1. Introduction

Most Spoken Dialogue Systems (SDS) are not capable of automatically adapting to changing situations, e.g., a changing environment or changing user needs. However, users are often not satisfied with the course of the dialogue, especially in a task-oriented setting. Hence, the users' needs change. For a conventional system, this usually results in bad performance and in an unsuccessful dialogue. However, a situation like that can be handled by rendering an SDS adaptive to the user by automatically recognizing the user satisfaction. This information may then be used to adapt the dialogue by influencing the action selection process of the dialogue management component. Hence, the user is put into the center of dialogue adaption.

However, a quality-metric for adapting the course of the dialogue during the interaction must fulfill certain requirements (Ultes et al., 2012), e.g., it must be automatically derivable for each system-user-exchange. The Interaction Quality paradigm by Schmitt et al. (2011) is the first metric which satisfies all requirements and thus can be used for this purpose. Hence, this contribution evaluates a quality-adaptive dialogue by conducting experiments with real users.

The remainder of the paper is organized as follows: the 2. Section presents significant related work. The Interaction Quality (IQ) paradigm, a more abstract form of user satisfaction which will be used in this work, is presented in the 3. Section. Furthermore, the test system is briefly presented in the 4. Section. The user experiment design and its results are finally presented in the 5. Section before the 6. Section concludes by summarizing the outcome of this work.

2. Significant Related Work

Much research has been performed on user-adaptive dialogue, very prominently work by Litman and Pan (2002). The authors identify problematic situations in dialogues by analyzing the performance of the speech recognizer (ASR) and use this information to adapt the dialogue strategy.

Each dialogue starts off with an user initiated strategy without confirmations. Depending on the ASR performance, the dialogue may eventually employ a system-directed strategy with explicit confirmations. Applied to TOOT, a system for getting information about train schedules, they achieved significant improvement in task success compared to a non-adaptive system. While Litman and Pan only adapt to the ASR performance being a system property as indicator for problematic dialogues, the user is put into the focus of adaption in this work by using an abstract form of user satisfaction.

San-Segundo et al. (2005) presented work following a similar approach. For a railway information system, a dialogue was created which adapts its confirmation strategy, i.e., the number of items to be confirmed (one or several) and the way of asking for confirmation (implicit vs. explicit), to the confidence of the speech recognizer. Furthermore, ASR confidence also influenced the sentence design. The authors successfully applied their system implementing adaption on four different levels in an evaluation with 105 users performing 335 calls. However, the performance has not been compared to a non-adaptive version of the system.

Further work on user-adaptive dialogue has been presented by Gnjatović and Rösner (2008). For solving the Tower-of-Hanoi puzzle with an SDS, they identify the emotional state of the user in order to recognize if the user is frustrated or discouraged. The dialogue is adapted by answering the questions "When to provide support to the user?", "What kind of support to provide?", and "How to provide support?" depending on the emotional state of the user. By that, the system is capable of providing well adapted support for the user which helps to solve the task.

Nothdurft et al. (2012) created a dialogue which is adaptive to the user knowledge. For the task of connecting a Blue-ray player with an amplifier using an HDMI cable, the multimodal system provides explanations on how to solve the task presenting text, spoken text, or pictures. The system makes assumption over the user knowledge by observing critical events within the dialogue (e.g., failed tries). Based

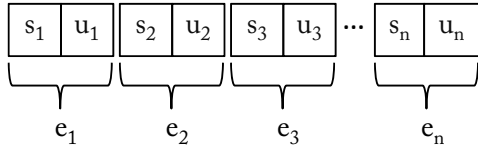


Figure 1: A dialogue may be separated into a sequence of system-user-exchanges where each exchange e_i consists of a system turn s_i followed by a user turn u_i .

on events extracted from the dialogue, the system generates explanations and selects the appropriate type of explanation so that the user can be expected to be capable of solving the task. The knowledge is stored in the knowledge model on a five-step scale where the knowledge fades over time.

3. Interaction Quality

For rendering an SDS adaptive to the user’s satisfaction level, a module is needed to automatically derive the satisfaction from the ongoing interaction. For creating this module, usually, dialogues have to be annotated with ratings describing the user’s satisfaction level. As a more objective means of describing the user’s satisfaction with the interaction, Schmitt et al. (2011) proposed a measure called “Interaction Quality” (IQ). For the authors, the main aspect of user satisfaction is that it is assigned by real users. However, they argue that this is impractical in many real world scenarios. Therefore, they propose the usage of expert raters instead. Further studies have also shown that ratings applied by experts and users have a high correlation (Ultes et al., 2013).

The IQ paradigm is based on automatically deriving interaction parameters from the SDS and feed these parameters into a statistical classification module. This predicts the IQ level of the ongoing interaction for the current system-user-exchange (Figure 1). The interaction parameters are rendered on three levels: the exchange level, the window level, and the dialogue level (see Figure 2). The exchange level comprises parameters derived from SDS modules Automatic Speech Recognizer, Spoken Language Understanding, and Dialogue Management directly. Parameters on the window and the dialogue level are sums, means, frequencies or counts of exchange level parameters. While dialogue level parameters are computed out of all exchanges of the dialogue up to the current exchange, window level parameters are only computed out of parameters from the last three exchanges.

These interaction parameters are used as input variables to a statistical classification module. The statistical model is trained based on annotated dialogues of the Lets Go Bus Information System in Pittsburgh, USA (Raux et al., 2006). Each of the 4,885 exchanges (200 calls) has been annotated by three different raters resulting in a rating agreement of $\kappa = 0.54$. Furthermore, the raters had to follow labeling guidelines to enable a consistent labeling process (Schmitt et al., 2012). An example of an annotated dialogue is shown in Figure 6.

Schmitt et al. (2011) applied a Support Vector Machine (Vapnik, 1995) (SVM) for estimating the Interaction

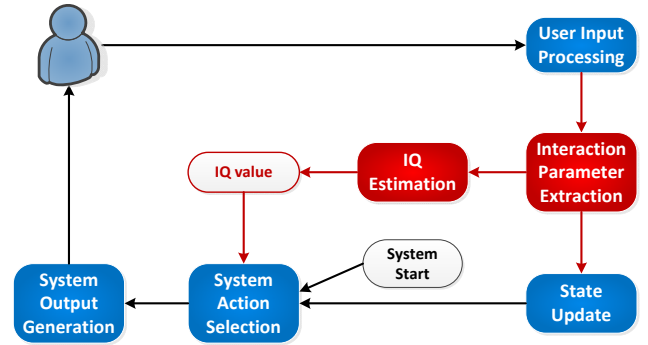


Figure 3: The IQ-adaptive dialogue processing cycle. For additional IQ-adaption, the red modules Interaction Parameter Extraction and IQ Estimation are integrated producing the estimation of the IQ value.

Quality achieving an unweighted average recall of 0.59 using a linear kernel.

4. The Test System

For evaluating the adaptive dialogue, a system based on the OwlSpeak dialogue manager (Heinroth et al., 2010) is used with added IQ-adaptivity. The basic principle shown in Figure 3 is rendered as an extension to the classic dialogue processing cycle (Ultes et al., 2014).

The system starts with selecting a first system action. This can be seen as valid not only for system initiative dialogues if the set of system action also includes the action of only waiting for user input without producing any output. Based on the selected system action, system output is generated. Now, it is the user’s turn. The output of the user is processed as user input to the system. Usually, this involves automatic speech recognition and a semantic analysis. The resulting semantics are then used to update the internal state of the system. Furthermore, the input is also processed to extract interaction parameters necessary for IQ estimation. Based on the IQ value and the updated internal dialogue state, the system selects the next system action and the cycle starts anew.

As OwlSpeak uses VoiceXML as interface, a voice browser along with speech recognition and speech synthesis are needed. In this experiment, Voxeo Prophecy has been used as browser and Loquendo ASR and TTS as speech recognition and synthesis.

5. User Evaluation

The test system has been used to evaluate quality-adaptive dialogue management within a simple train booking dialogue with real users. Depending on the current IQ value, the confirmation strategy was adapted, i.e., each time the system requests a confirmation about a certain slot value from the user, the IQ value is used to decide whether the system uses an explicit or implicit confirmation prompt. In the following, the design and setup of the study will be presented before giving details about the results.

5.1. Design and Setup

For adapting the dialogue to the Interaction Quality, the confirmation strategy was selected out of one simple rea-

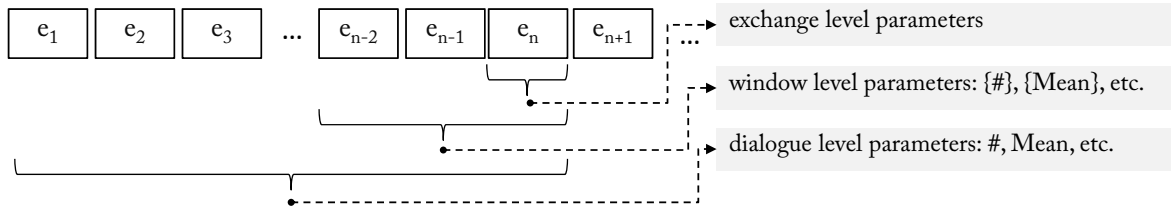


Figure 2: The three different modeling levels representing the interaction at exchange e_n : The most detailed exchange level, comprising parameters of the current exchange; the window level, capturing important parameters from the previous n dialog steps (here $n = 3$); the dialog level, measuring overall performance values from the entire previous interaction.

	fun with interaction	system reaction	hard to lose track	natural communication	user is in control	overall impression	user felt calm	user felt tense	dialogue length appropriate	user knew what to say	high level of concentration necessary	system understands user input	functionality	interaction was pleasant	interaction was not exhausting	interaction was not boring	interaction was diversified	interaction was not frustrating	system made no errors	error due to user behaviour	system is friendly	system is easy to use	system is flexible	system is understandable	reaction time	would use system again	human operator preferred	overall satisfaction
all-implicit	5.3 ^e	5.9 ^e	6.2	4.8 ^e	3.0	5.4 ^e	5.0	6.3 ^{ae}	5.4 ^e	6.0	2.7	5.9 ^e	5.5	5.6 ^e	5.6 ^e	4.7 ^e	3.9 ^e	5.9 ^e	5.9 ^e	3.3	5.8	6.3 ^e	4.1	5.8	5.0 ^e	5.2	4.7	5.5 ^e
all-explicit	3.7 ^{ai}	4.8 ⁱ	5.8	3.5 ⁱ	2.7	4.3 ⁱ	4.5	5.0 ⁱ	3.5 ^{ai}	6.1	3.3	4.2 ⁱ	4.6	4.3 ⁱ	4.3 ⁱ	3.3 ^{ai}	3.0 ^{ai}	4.4 ^{ai}	4.5 ⁱ	2.1	5.8	5.6 ⁱ	3.8	5.5	3.9 ^{ai}	4.2	5.1	4.4 ^{ai}
adapted	4.8 ^e	5.3	6.1	4.5	3.0	5.0	4.8	5.4 ⁱ	5.0 ^e	6.2	3.2	5.2	5.3	5.0	5.3	4.3 ^e	3.9 ^e	5.5 ^e	5.3	2.8	5.6	6.0	4.2	5.8	4.9 ^e	5.0	4.7	5.3 ^e

Table 1: The average results of the user questionnaires. Each question could be answered by a 7-point scale being translated to scores from one to seven. Significant differences are marked with a, e, and i marking significance with the adaptive, explicit, and implicit strategy respectively. (Please note: the original questionnaire was in German.)

sons: It is an easily adaptable concept which occurs in almost every dialogue in which the user has to provide information. A dialogue in the train booking domain was created asking the user for information about the origin, the destination, the day of the week and the time of travel. The user could choose out of 22 cities which were used as origin and destination alike. Furthermore, the time of travel was restricted to every full hour (1 pm, 2 pm, 3 pm, etc.). Three different dialogues were created: one only applying explicit confirmation (all-explicit), one applying only implicit confirmation (all-implicit), and one adapting the confirmation type to the current IQ value (adapted). Besides these differences, the dialogues were the same. The complete dialogue was system initiated and the course of the dialogue was predetermined, i.e., the order of information the user was asked to provide was given. A sample for the adapted strategy is illustrated in Figure 4. As only two different options for adapting the dialogue exist, i.e., either selecting implicit or explicit confirmation, the IQ value has been limited to only two values: two representing a satisfied user and one representing an unsatisfied user. If the user was recognized as being satisfied with the dialogue (high IQ value), slot values were confirmed implicitly while explicit confirmation was applied for unsatisfied users (low IQ value). In the end of the dialogue, the user was provided with a dummy message stating that the reservation has been made.

Before the experiment, each participant was presented with a sheet of paper stating all options they could say during the dialogue. This also included a list of all cities. Furthermore, each user participated in three runs of the dialogue—one for each type of confirmation strategy. During the experiment, the order of these dialogues has been alternated to get an

equal distribution over all combinations so that learning effects are taken account of. However, the user was not aware about the different dialogue types. After each dialogue, the participants were asked to fill out a questionnaire based on the SASSI questionnaire (Hone and Graham, 2000) to evaluate their overall impression with the dialogue. Each item was rated on a seven-point scale.

In total, there were 24 participants (8 female, 16 male) creating 72 dialogues with an average number of turns of 33.58. They were between 19 and 38 years old with an average age of 26.42. The participants were students from multiple disciplines.

5.2. Results

To evaluate the user experiment, the questionnaires are analyzed. The results for each question is depicted in Table 1. Each row shows the average score for one of the three different strategies. It is a well known fact that, for simple tasks like this, an all-implicit strategy is usually preferred over an all-explicit strategy (cf. (Fraser, 1994)). Hence, as expected, the all-implicit strategy performed best outperforming the all-explicit strategy clearly: it achieved a better score for almost all questions. The difference is even significant for 16 out of 25 values ($\alpha < .05$ applying the Mann-Whitney U test (Mann and Whitney, 1947)). Comparing the all-explicit to the adapted strategy gives a similar impression: The scores for almost all questions are better for the adapted strategy. However, this is not as significant having only 7 significant different values. More revealing is the conclusion drawn from comparing the all-implicit with the adapted strategy. While the all-implicit strategy again governs the scores, almost all results are not significantly

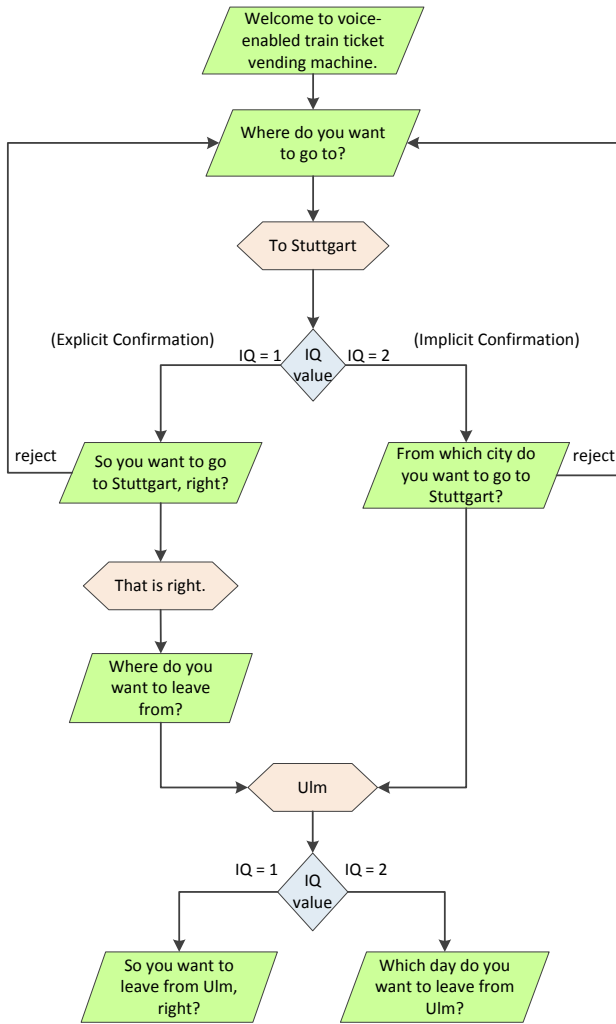


Figure 4: The dialogue flow for the adaptive strategy. Depending on the IQ value, the provided information by the user is either confirmed explicitly or implicitly within the next system question. (Please note: the original dialogue was in German.)

different. Hence and in contrast to the expectations, the adapted strategy did not perform significantly worse despite the dialogue being very simple.

This result is underpinned by looking at the users' overall satisfaction score with the dialogue as an emphasis was put on the question which strategy people liked best. A bar graph showing the average outcome of the user ratings grouped by the respective dialogue strategy is depicted in Figure 5. While the adapted strategy resulted in 45.6 % explicit and 54.4 % implicit confirmations, it is very interesting that it was not rated significantly different compared to the all-implicit strategy. That is even, although the ASR component made almost no errors (due to the limited number of options). Moreover, calculating Spearman's Rho (Spearman, 1904) shows significant correlation ($\alpha < 0.01$) with $\rho = 0.6$ between the users' overall satisfaction of the all-implicit and adapted strategy. Additionally, the dialogue length, which is one main indicator for user satisfaction in simple dialogues like this, is significantly higher for the adapted strategy compared to the

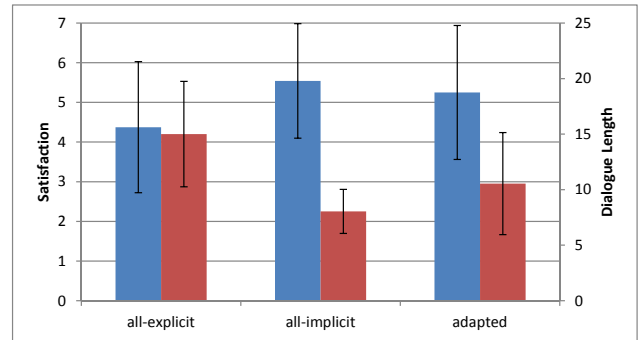


Figure 5: The overall satisfaction with the dialogue (left bar, left y-axis) and the average dialogue length in number of turns (right bar, right y-axis) according to questionnaire evaluation. Satisfaction for implicit and adapted do not differ significantly while all other differences are significant.

all-implicit strategy.

In other words, although the task was quite simple, there was no difference between the all-implicit and adapted strategies spurring the hope that for more complex dialogues, quality-adaption will perform best.

6. Conclusion

In this paper, a pilot experiment on adapting the dialogue to the current user satisfaction has been conducted with real users (in contrast to a simulated user). In the study, we investigated dialogues whose confirmation strategy was adapted to the Interaction Quality. We could show that, even for simple dialogues, the adaptive strategy was not significantly worse than a all-implicit strategy which is known to work best for simple dialogues like the one applied. Therefore, we believe that for more complex dialogues an adaptive strategy will perform best and plan on investigating this with more extensive user studies applying more complex dialogues.

7. References

- Fraser, N. M. (1994). The sundial speech understanding and dialogue project: results and implications for translation. In *Aslib proceedings*, volume 46, pages 141–148. MCB UP Ltd.
- Gnjatović, M. and Rösner, D. (2008). Adaptive dialogue management in the nimatek prototype system. In *PIT '08: Proceedings of the 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 14–25, Berlin, Heidelberg. Springer-Verlag.
- Heinroth, T., Denich, D., and Schmitt, A. (2010). Owl-speak - adaptive spoken dialogue within intelligent environments. In *IEEE PerCom Workshop Proceedings*, March. presented as part of SmartE Workshop.
- Hone, K. S. and Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (sassi). *Nat. Lang. Eng.*, 6(3-4):287–303.
- Litman, D. and Pan, S. (2002). Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, 12(2-3):111–137.

Figure 6: Example of a dialogue annotated with Interaction Quality for each system-user-exchange. As the system (S) fails to interact properly with the user (U) in the end, the IQ value drops down to one. Utterance represents either the system utterance or the recognized user utterance. The example is taken from the LEGO corpus (Schmitt et al., 2012) identified by CallID 2061123000.

Utterance	IQ
S: Welcome to the CMU Let's Go bus information system.	
U: -	5
S: To get help at any time, just say Help or press zero.	
U: -	5
S: What can I do for you?	
U: PENN AT BUTLER TIME IS IT IS THE CENTRE AT HIGHLAND FROM RUNS THE 61A EIGHT THIRTY 61B CAN'T OAKLAND	5
S: Leaving from OAKLAND . Is this correct?	
U: OAKLAND	4
S: Leaving from OAKLAND . Is this correct?	
U: YES	4
S: Right	
U: -	4
S: Where do you want to go?	
U: SQUIRREL HILL OF FIFTH AVENUE	4
S: Going to FIFTH AVENUE . Is this correct?	
U: YES	5
S: Alright	
U: -	5
S: When do you wanna travel?	
U: SUNDAY	4
S: At what time do you wanna travel?	
U: AT FIVE FOUR VERY ABOUT ELEVEN O'CLOCK	4
S: Leaving at 11 a.m. . Is this correct?	
U: YES	5
S: Right	
U: -	5
S: Hold on. Let me check that for you.	
U: -	5
S: OAKLAND and FIFTH AVENUE are both the same stop. Please provide a different start or end point.	
U: -	3
S: To ask about a different trip, you can say, start a new query. If you are finished, you can say goodbye.	
U: 61A	2
S: You can say start a new query, or goodbye.	
U: FIVE DUQUESNE	1
S: To ask about a different trip, you can say, start a new query. If you are finished, you can say goodbye.	
U: LAST THE NEXT	1
S: You can say start a new query, or goodbye.	
U: START WH	1
S: To ask about a different trip, you can say, start a new query. If you are finished, you can say goodbye.	
U: THE	1
S: You can say start a new query, or goodbye.	
U: THANK YOU QUERY ELEVEN BACK	1
S: To ask about a different trip, you can say, start a new query. If you are finished, you can say goodbye.	
U: GOODBYE	1
S: Thank you for using the cmu Let's Go Bus Information System. Goodbye.	
U: -	1

- Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 18(1):50–60.
- Nothdurft, F., Honold, F., and Kurzok, P. (2012). Using explanations for runtime dialogue adaptation. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 63–64. ACM.
- Raux, A., Bohus, D., Langner, B., Black, A. W., and Eskenazi, M. (2006). Doing research on a deployed spoken dialogue system: One year of let's go! experience. In *Proc. of the International Conference on Speech and Language Processing (ICSLP)*, September.
- San-Segundo, R., Montero, J. M., Macías-Guarasa, J., Ferreiros, J., and Pardo, J. M. (2005). Knowledge-combining methodology for dialogue design in spoken language systems. *International Journal of Speech Technology*, 8(1):45–66.
- Schmitt, A., Schatz, B., and Minker, W. (2011). Modeling and predicting quality in spoken human-computer interaction. In *Proceedings of the SIGDIAL 2011 Conference*, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Schmitt, A., Ultes, S., and Minker, W. (2012). A parameterized and annotated corpus of the cmu let's go bus information system. In *International Conference on Language Resources and Evaluation (LREC)*.
- Spearman, C. E. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103.
- Ultes, S., Schmitt, A., and Minker, W. (2012). Towards quality-adaptive spoken dialogue management. In *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SD-CTD 2012)*, pages 49–52, Montréal, Canada, June. Association for Computational Linguistics.
- Ultes, S., Schmitt, A., and Minker, W. (2013). On quality ratings for spoken dialogue systems – experts vs. users. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 569–578. Association for Computational Linguistics, June.
- Ultes, S., Dikme, H., and Minker, W. (2014). Dialogue management for user-centered adaptive dialogue. In

Proceedings of the 5th International Workshop On Spoken Dialogue Systems (IWSDS). Springer, January.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA.