COLLECTION OF USER JUDGMENTS ON SPOKEN DIALOG SYSTEM WITH CROWDSOURCING

Zhaojun Yang[1], *Baichuan Li*[2], *Yi Zhu*[2], *Irwin King*[2], *Gina Levow*[3‡], *Helen Meng*[1†]

[1]Department of System Engineering and Engineering Management
[2]Department of Computer Science and Engineering
[1,2]The Chinese University of Hong Kong, Shatin.N.T., Hong Kong
[3]Department of Linguistics, University of Washington, Seattle, WA 98195 USA

ABSTRACT

This paper presents an initial attempt at the use of crowdsourcing for collection of user judgments on spoken dialog systems (SDSs). This is implemented on Amazon Mechanical Turk (MTurk), where a Requester can design a human intelligence task (HIT) to be performed by a large number of Workers efficiently and cost-effectively. We describe a design methodology for two types of HITs – the first targets at fast rating of a large number of dialogs regarding some dimensions of the SDS's performance and the second aims to assess the reliability of Workers on MTurk through the variability in ratings across different Workers. A set of approval rules are also designed to control the quality of ratings from MTurk. At the end of the collection work, user judgments for about 8,000 dialogs rated by around 700 Workers are collected in 45 days. We observe reasonable consistency between the manual MTurk ratings and an automatic categorization of dialogs in terms of task completion, which partially verifies the reliability of the approved ratings from MTurk. From the second type of HITs, we also observe moderate inter-rater agreement for ratings in task completion which provides support for the utilization of MTurk as a judgments collection platform. Further research on the exploration of SDS evaluation models could be developed based on the collected corpus.

Index Terms— spoken dialog system, user judgment, crowdsourcing, Amazon Mechanical Turk, Let's Go

1. INTRODUCTION

Spoken dialog systems (SDSs) are becoming increasingly pervasive in supporting information access by the masses. There have been many SDSs developed recently, such as flight information, restaurant guide, or tutoring student learning. It is highly desirable to evaluate these systems and compare their performance. However, principled ways of evaluating an SDS remain an open research problem. Such evaluation can be categorized into *component-based* perspective and *holistic* perspective. The component-based perspective covers the performance of individual components such as the correctness in speech recognition, language understanding, the appropriateness in response generation, as well as the naturalness of the synthetic speech in conveying the responses. The holistic perspective involves the perceived level of system usability, system intelligence, and abilities in error recovery by considering the system entirely [1].

A thorough evaluation of an SDS needs to consider all relevant evaluation metrics covering the functionalities for all the system components [2]. As a result, different kinds of evaluation metrics have been proposed in previous work, such as query density and concept efficiency for measuring the system's ability in understanding [3]. According to the functionalities that the metrics are related to, metrics have also been classified into the five categories: *dialog-related*, *meta-communication-related*, *cooperativity-related*, *task-related*, and *speech-input-related* [2].

For holistic evaluation, it also needs to cover the wide variety of users' impressions relating to all dimensions of quality of an SDS [4]. A usual and desirable approach may be to invite subjects to fill out a questionnaire after interacting with the SDS, which could cover all aspects of perceptions for the system such as task completion or user satisfaction. However, the main disadvantage of this method is that it is slow, tedious, and costly. More importantly, due to constrained resources, this approach can only be limited to a small number of evaluators, whose feedbacks may not be statistically representative of the large user population that can access the SDS. Previous evaluation frameworks such as PARADISE [5] have been proposed for automatic inference of overall user satisfaction of unrated dialogs. Nevertheless they still need at least a small number of evaluated dialogs to train the predictive model. Furthermore, in some situations where the system has already been in deployment, it is often difficult to ask real users to patiently complete an evaluation survey.

In this paper, we present an alternative approach towards collection of user judgments on SDSs, which uses

†H. Meng is the corresponding author.

‡This work was done while Dr. Gina Levow was a Visiting Scholar at The Chinese University of Hong Kong.

crowdsourcing through Amazon Mechanical Turk (MTurk)[1]. Crowdsourcing means to outsource a task to a crowd of people. MTurk is a crowdsourcing marketplace that makes use of human intelligence online to perform tasks which cannot be completed entirely by computer programs. Tasks on MTurk can be completed by a large number of people over the Internet in a cost-effective, efficient, and flexible manner. Much previous work has been devoted to the creation of speech and language data through MTurk [2] [6]. In our work, we design a basic type of HITs for efficiently collecting user judgments on numerous dialogs, as well as an extended type of HITs for assessing the reliability of MTurk Workers. In addition, a set of approval rules are developed. They are necessary to exclude random submissions with nonsensical ratings, which will affect the overall quality of the ratings obtained. To the best of our knowledge, this is the first attempt at the use of crowdsourcing to collect user judgments for SDS evaluation. We believe that this work is a good complement to conventional approaches. The methodology in assigning HITs to crowdsourcing can also bring about further research.

The rest of this paper is organized as follows. Section 2 describes the experimental dialog corpus and a simple dialog classification algorithm in terms of task completion. Section 3 introduces the methodology of user judgments collection using crowdsourcing. Collected results are exhibited in Section 4, together with analysis of the results. Section 5 presents the conclusions and possible directions of future work.

2. EXPERIMENTAL CORPUS AND AUTOMATIC DIALOG CLASSIFICATION

The corpus used in this study contains about 50,000 dialogs drawn from the log files collected through the Let's Go dialog system developed by Carnegie Mellon University [7]. This system provides bus schedule information for Pittsburgh residents when the customer service hotline is not staffed. Input utterances from users are transcribed automatically by automatic speech recognition (ASR). Fig. 1 shows examples of the Let's Go dialogs.

We survey the dialog logs and identify several cue phrases that enable us to develop a set of heuristics to automatically classify the dialog logs into five categories in terms of task success. The dialog logs are tracked from the end towards the beginning with a set of ordered rules:

1. Dialogs with no more than 3 turns are classified as the "TooShort" category.

2. Dialogs with the cue phrases, "Start a new query", "New query", etc., in the user turns are classified as the "MultiTask" category.

Case 1: TaskComplete
S: *Welcome to the CMU Let's Go bus information system.*
S: *What can I do for you?*
U: *When is the next 64A from Forbes and Murray to East Liberty?*
S: *The 64A. Did I get that right?*
U: *Yes.*
S: *Leaving from Forbes and Murray. Correct?*
U: *Yes.*
S: *Going to East Liberty. Am I right?*
U: *Yes*
S: *I think you want the next bus.*
U: *Yes*
S: *Hold on. I'll look that up.*
S: *The next 64A leaves Murray Avenue at Forbes at 10:32pm and arrives at Highland Avenue at Penn Avenue at 10:55 pm.*

Case 2: TooShort
S: *Welcome to the CMU Let's Go bus information system.*
S: *What can I do for you?*
U: *OAKLAND .ANY. .FIND. .SOUTH SIDE .WANT.*

Case 3: MultiTask
. . . . . .
S: *There is a 28X leaving THE AIRPORT* · · ·
U: *NEW QUERY*
S: *Okay, let's start from the beginning.*
. . . . . .

Case 4: OutofScope
. . . . . .
S: *I'm sorry, I don't know any routes* · · ·

Case 5: TaskIncomplete
. . . . . .
S: *Leave at 8 a.m. Is this correct?*
U: *YES*
S: *Hold on, I'll look that up.*

Fig. 1. *Examples of the Let's Go dialogs.*

3. Dialogs with the cue phrase "...result[3]..." in the system turns are classified as the "TaskComplete" category.

4. Dialogs with the cue phrase "I'm sorry...." in the system turns are classified as the "OutofScope" category.

5. If none of the above rules apply, the dialog is classified as the "TaskIncomplete" category.

The distribution of the classification results is shown in Fig. 2. About 35% of the dialogs are in the TaskComplete category, while the remaining categories share comparable percentages. As will be described later, this rough classification is used to determine the proportion of the different types of dialogs that are uploaded to MTurk for crowdsourcing. The

[1] www.mturk.com

[2] See NAACL-HLT 2010 Workshop on the use of MTurk for speech and language collection/annotation/evaluation.

[3] This cue phrase is observed in the grammar tags of the Let's Go System's dialog logs.

rough classification also provides a means by which we can verify whether the human raters on the MTurk platform have devoted efforts in rating our dialogs, as opposed to providing random (and unreliable) inputs. We will elaborate on these points in the following sections.

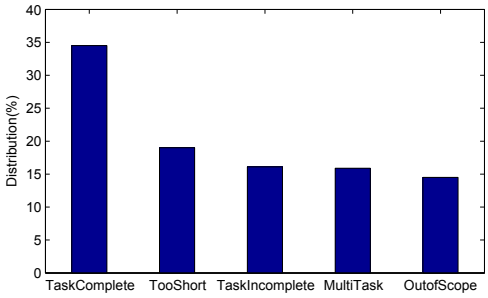

Fig. 2. *Distribution of the five dialog categories based on the automatic, heuristics-based classification.*

3. USER JUDGMENTS COLLECTION WITH CROWDSOURCING

The objective of collecting user judgments with crowdsourcing is to get a large number of gathered dialogs evaluated by numerous people (hence more statistically representative) in an efficient and cost-effective manner, which is difficult to be realized using traditional methods. The MTurk platform organizes work in the form of human intelligence tasks (HITs). An HIT is designed by the "Requester" (i.e., our research team) and is completed by many "Workers" (i.e., anyone who is interested in the task) over the Internet. Each HIT has a preset payment and number of Workers. Before accessing the HIT, each Worker needs to meet the qualification requirement, i.e., the Worker's approval rate should be higher than or equal to 98%. The Workers' inputs to the HITs will undergo an "approval process" by the Requester. Approved inputs will result in automatic payment from the Requester to the Worker via the MTurk platform.

3.1. HITs on Dialog Evaluation

This type of HITs are designed to outsource the assessment of the SDS to MTurk Workers. The assessment focuses on selected dimensions of performance regarding the SDS, based on a large number of selected dialogs from the logs. To achieve this goal, we have authored a set of questions that constitute the HIT in Table 1.

As shown in Table 1, we include the explanation of the aim for each question, but this is not shown to the MTurk Workers. These questions cover user's confidence (user expertise), the perceived task completion, expected behavior, overall performance, and categorization of task success. In particular, for Question 5, the initial set of three answer options led to much disagreement among Workers. Many also sent us comments about the lack of a clear definition of task completion versus task incompletion. Consequently, we revised to include seven answer options (see Table 2) based on the ITU Recommendation [8]. We have purposely designed the questions in such a way that they can cross-validate each other (Q2 and Q5 both aim to assess task completion), which will be used for approval of ratings from MTurk later.

Table 1. *Questions constituting the HIT on Dialog Evaluation. (Q-Question, Opt-Options)*

Q1	**Do you think you understand from the dialog what the user wanted?**
Opt	1) No clue 2) A little bit 3) Somewhat 4) Mostly 5) Entirely
Aim	*elicit the Worker's confidence in his/her ratings.*
Q2	**Do you think the system is successful in providing the information that the user wanted?**
Opt	1) Entirely unsuccessful 2) Mostly unsuccessful 3) Half successful/unsuccessful 4) Mostly successful 5) Entirely successful
Aim	*elicit the Worker's perception of whether the dialog has fulfilled the informational goal of the user.*
Q3	**Does the system work the way you expect it?**
Opt	1) Not at all 2) Barely 3) Somewhat 4) Almost 5) Completely
Aim	*elicit the Worker's impression of whether the dialog flow suits general expectations.*
Q4	**Overall, do you think that this is a good system?**
Opt	1) Very poor 2) Poor 3) Fair 4) Good 5) Very good
Aim	*elicit the Worker's overall impression of the SDS.*
Q5	**What category do you think the dialog belongs to?**
Opt	1) Task is incomplete 2) Out of scope 3) Task is complete
Aim	*elicit the Worker's impression of whether the dialog reflects task completion.*

Table 2. *Definitions of different levels of task success, based on the ITU Recommendation [8].*

TS:S	Succeeded (task for which solutions exist)
TS:Cs	Succeeded with constraint relaxation by system
TS:Cu	Succeeded with constraint relaxation by the user
TS: CsCu	Succeeded with constraint relaxation both from the system and from the user
TS:SN	Succeeded in spotting that no solution exists
TS:Fs	Failed because of the system behavior, due to system inadequacies
TS:Fu	Failed because of the user behavior, due to non-cooperative user behavior

Each HIT contains the text transcription of one dialog and the questionnaire in Table 1 for assessment by the Workers, who are paid USD $0.05 for each task completed. We have uploaded 11,000 dialogs in total, including samples from the three major dialog categories and in proportions that follow

the percentages obtained from the automatic classification, i.e., TaskComplete (55%), TaskIncomplete (27%), OutofScope (18%). TooShort and MultiTask dialogs are excluded from the HIT. The former is easily detectable as unsuccessful. The latter can be easily segmented into mono-task dialogs, which can then follow the three-way categorization (TaskComplete / TaskIncomplete / OutofScope) directly.

3.2. HITs on Inter-rater Agreement

This type of HITs are the extensions of those in Section 3.1 and are designed to assess the reliability of MTurk Workers through inter-rater agreement across different raters. Each HIT includes the text transcriptions of 30 selected dialogs (10 dialogs from the categories of TaskComplete, TaskIncomplete, and OutofScope respectively). Each dialog is associated with the questionnaire in Table 1. Workers are paid USD $1.5 for each task completed. Altogether, we have 3 groups of Workers (each with 16 individuals) rating two sets of dialogs (each with 30). Groups 1 and 2 evaluate the first set of dialogs, while Group 3 evaluate the second set. In this way, we can assess whether the inter-rater agreement varies across different raters and different dialogs.

3.3. Approval of Ratings

It is important to verify the quality of inputs from a large number of MTurk Workers. Since the quality of the ratings directly impacts the credibility of the SDS evaluation, some basic rules have to be set to ensure the Workers are devoting efforts and to guarantee the reliability of ratings, in addition to the qualification requirement preset for the Workers. We have developed the approval mechanism, as follows:

R1. We reject HITs for which the working time is less than 15 seconds, since we feel that careful (and thus high quality) ratings cannot be completed within such a short period.

R2. If an MTurk Worker completes a large number of HITs (e.g., over 20) but provides identical answers for all of them, his/her work will be rejected.

R3. Approval requires consistency between the answers to related questions (Q2 and Q5). Consistency is based on four main heuristics:

- Answers to Q2 being "Entirely successful" or "Mostly successful" can go with answers to Q5 being TS:S, TS:CS, TS:Cu, or TS:CsCu.
- Answers to Q2 being "Entirely unsuccessful" or "Mostly unsuccessful" can go with answers to Q5 being TS:Fs or TS:Fu.
- The answer to Q2 being "Half unsuccessful / successful" can go with any answer in Question 5.
- The answer to Q5 being TS:SN can go with any answer to Q2.

R4. Approval requires consistency between the answers to Q5 and the automatic classification of the dialogs (see Section 2). In particular, the heuristics are:

- TaskComplete can match with TS:S, TS:Cs, TS:Cu, and TS:CsCu.
- TaskIncomplete can match with TS:Fs and TS:Fu.
- OutofScope can match with TS:SN.

R5. If these above heuristics are not satisfied, the dialog will be checked carefully. Random (incorrect) ratings are rejected. However, we have approved some ambiguous cases, as they will be explained in Section 4.2.

4. RESULTS AND ANALYSIS

4.1. Approval Rates and Comments from Mturk Workers

11,000 HITs are rated by around 700 online Workers in 45 days. Three persons in our team complete the verification of the rated HITs and approve 8,394 of them. The total expenditure paid to the Workers is USD $350. Approval rates for each dialog category, i.e., TaskComplete, TaskIncomplete, and OutofScope, are 79.59%, 65.23%, and 90.65% respectively. OutofScope is the highest because some Workers consider a task to be successful if they think that the absence of the information is due to the database but not the ability of the system. Others consider such cases as failures since the system does not provide the requested information for the users. We approve either decision from the Workers.

Rejected dialogs led to some controversies. Some apologized for their errors and others complained about the rejections. We received feedbacks from the MTurk Workers concerned, many of which are useful to help enhance our understanding of SDS evaluations. Here we list some typical comments associated with their implications as follows.

- The system does not provide exact information that the user wanted although it provides some related results. (*Retrieval result from database is a vital aspect of SDS performance.*)
- The understanding ability of system is very important on the user's first try, so good understanding may lead workers to choose task success even if the system does not provide any information to the user. (*Good language understanding ability plays an important role in improving user satisfaction.*)
- The system succeeds in providing a message based on the user's initial inputs but fails to follow up with the user's updated information. (*Timely updating the dialog history impacts users' perception on SDS performance positively and greatly.*)

4.2. Consistency between Automatic Dialog Classification and Manual Ratings

To assess the quality of ratings from MTurk, we investigate the consistency between automatic dialog classification (see Section 2) and the manual ratings from MTurk Workers (with respect to Questions 2 and 5) based on the approved HITs of about 8,000 dialogs.

According to the approval rules, the HITs whose ratings of Q5 are not consistent with the automatic classification will be rejected (see R4 in Section 3.3). However, as mentioned in Section 3.3, we still approve some ambiguous dialogs for which it is difficult to determine their success in task completion. The manual ratings of Q5 for the ambiguous dialogs do not agree with automatic classification. Case 5 of Fig. 1 shows an example, where the dialog is terminated midway. Some Workers regard the dialog as successful (in Question 5), because they feel that the system would have completed the task had the user not hung up. Nevertheless, since the dialog does not have the necessary cue phrase (i.e. "..result..") in the log, it is automatically classified as TaskIncomplete. Percentages of ambiguous dialogs over the total approved ones in each category of TaskComplete, TaskIncomplete, and OutofScope are 13.5%, 17%, and 46.5% respectively. Note that a higher percentage of ambiguous dialogs are approved for the OutOfScope category, mainly due to MTurk Workers' diversified understandings of such kind of dialogs, as we have discussed in Section 4.1.

Moreover, Fig. 3 plots the mean answer scores of the two questions for approved dialogs in each of the three categories, where a higher score maps to a higher level of success, namely: scores to Q2 range from 0 for "Entirely unsuccessful" to 1 for "Entirely successful"; scores to Q5 range from 0 for "TS:Fu" to 1 for "TS:S". The scores to Q2 and Q5 have been normalized in the same range from 0 to 1.

Generally, despite some ambiguous dialogs are approved, we still observe reasonable agreement, i.e., the dialogs automatically classified as TaskComplete receive high scores from the Workers, those automatically classified as TaskIncomplete receive low scores, and those in OutofScope category receive neutral scores. Such consistency verifies the reliability of the approved ratings from MTurk to some extent.

4.3. Inter-rater Agreement

As mentioned earlier, the second type of HITs (see Section 3.2) are designed to assess the level of inter-rater agreement among the MTurk Workers. We adopt Cohen's weighted kappa measure which is often applied to ordinal categories,

$$Kappa = \frac{\sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}(n_{ij}/N - n_{i.}n_{.j}/N^2)}{1 - \sum_{i=1}^{c}\sum_{j=1}^{c} w_{ij}n_{i.}n_{.j}/N^2}, \quad (1)$$

where c is the number of categories (i.e., answer options for each question here, $c = 5$ for Q1-Q4 and $c = 7$ for Q5),

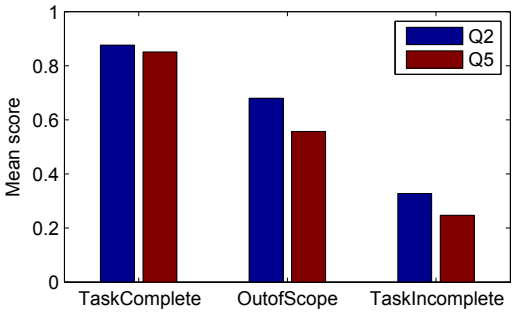

Fig. 3. *The normalized mean scores of Q2 and Q5 for approved ratings in each category. A higher score maps to a higher level of task success.*

$w_{ij} = 1 - \frac{(i-j)^2}{(1-c)^2}$, n_{ij} is the element in the observed matrix, $n_{i.} = \sum_j n_{ij}$ and $n_{.j} = \sum_i n_{ij}$. Details can be found in [9]. A higher kappa value indicates a higher inter-rater agreement.

Recall that we have three groups of Workers rating two sets of dialogs. These ratings are accepted directly and do not undergo the approval process. For any pair of Workers in each group, we compute the weighted kappa value for each question. We then compute the mean weighted kappa value for each question over the entire group. Results are shown in Fig. 4. Despite the fact that groups 1 and 2 evaluated the same dialog set, while group 3 evaluated a different dialog set, the three kappa plots remain close, which illustrates that the inter-rater agreement for each question remains stable across different raters and different dialogs. In particular, Q5 (categorization of task success) achieves mean weighted kappa values above 0.6 and Q2 (perceived task completion) achieves reasonable values above 0.4, which is indicative of a moderate level of agreement [10]. Q2 and Q5 are about task success which can gain "official" or somehow objective ratings from reliable raters, so the moderate and stable agreement partially shows the reliability of MTurk Workers and provides support for the utilization of MTurk as a judgment collection platform. On the other hand, Q3 (expected behavior) and Q4 (overall impression on system performance) have low values below 0.3, which is indicative of a lack of agreement. This suggests that evaluation based on overall user satisfaction may be quite subjective. The low agreement in user satisfaction may lead to the low prediction accuracy for the evaluation model, which has been analyzed in [11].

We conceive of several factors that cause low inter-rater agreement in overall user satisfaction. First, different users may emphasize different aspects of system performance, ranging from the system's intelligence, task completion, dialog efficiency, etc. Second, raters with different levels of domain knowledge may have different expectations of the system. It may not be meaningful to compute an overall average score of user satisfaction across a diversity of users. Instead, we may consider the following possibilities:

- Evaluate SDSs along different dimensions individually.

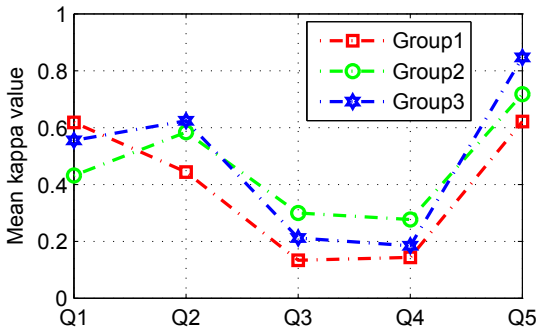

Fig. 4. *Mean values of weighted Kappa of five questions for 3 groups. The Kappa values are stable across different raters and different dialogs. Q2 and Q5 which can gain "official" ratings from reliable raters have high levels of inter-rater agreement.*

- Evaluate SDSs in terms of different types of user queries, targeting at different system functionalities.
- Evaluate SDSs based on different user groups with different levels of domain knowledge.

5. CONCLUSIONS AND FUTURE WORK

This paper presents our initial attempt at the use of crowdsourcing for collection of user judgments on spoken dialog systems through MTurk. We describe a design methodology for two types of HITs - the first targets at fast collecting ratings of a large number of dialogs efficiently and the second aims to assess the reliability of MTurk Workers through inter-rater agreement. A set of approval rules are also developed to take care of the quality of ratings from MTurk.

Compared with the traditional method of inviting subjects to fill out a questionnaire after interaction, the results we achieved show that the crowdsourcing method is more efficient, flexible, and inexpensive and could access more statistically representative population. At the same time, the quality of ratings can also be controlled. Reliable ratings for 8,394 dialogs rated by around 700 online Workers are approved. Approval rates for each dialog category, i.e., TaskComplete, TaskIncomplete, and OutofScope, are 79.59%, 65.23%, and 90.65% respectively. Reasonable consistency between the manual MTurk ratings and the automatically classified dialogs in terms of task success is an indicator of the reliability of the approved ratings from MTurk. The moderate level of inter-rater agreement for ratings in task completion partially verifies the reliability of MTurk Workers. It also suggests that the MTurk platform is suitable for collecting users' perceptions on system quality and can be further explored in terms of research.

A drawback of the approach is that MTurk Workers are not the real users of the system. The Workers' quality perceptions of dialogs may not reflect the users' real needs and experiences with the system. Further research may consider publishing the SDS online through MTurk so that Workers can communicate with the system and evaluate their own interactions. The current methodology is specific for SDS evaluation task through MTurk and needs to be compared with other traditional methods. It will also be expanded in the future to evaluate SDSs in more dimensions, such as system intelligence or error recovery abilities. Moreover, the collected corpus of ratings on the Let's Go system from MTurk will be used to develop a quantitative SDS evaluation framework.

6. ACKNOWLEDGEMENT

The project team is a participant in the Spoken Dialog Challenge 2010 (http://dialrc.org/sdc), which is organized by Professor Maxine Eskenazi and Professor Alan Black of the CMU Dialog Research Center. The work is partially supported by a grant from the HKSAR Government Research Grants Council (Project No. 415609) and that from MSRA FY09-RES-OPP-103 (Reference No. 6902682). This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies.

7. REFERENCES

[1] M.F. McTear, "Spoken dialogue technology: enabling the conversational user interface," *ACM Computing Surveys*, vol. 34, no. 1, pp. 169, 2002.

[2] S. Möller, "Parameters for quantifying the interaction with spoken dialogue telephone services," in *Proc. of the SIGDIAL*, 2005.

[3] J. Glass, J. Polifroni, S. Seneff, and V. Zue, "Data collection and performance evaluation of spoken dialogue systems: the MIT experience," in *Proc. of ICSLP*, 2000.

[4] H. Bonneau-Maynard, L. Devillers, and S. Rosset, "Predictive performance of dialog systems," in *Proc. of LREC*, 2000.

[5] M.A. Walker, D.J. Litman, C.A. Kamm, and A. Abella, "PARADISE: a framework for evaluating spoken dialogue agents," in *Proc. of European chapter of the Association for Computational Linguistics*, 1997.

[6] I. McGraw, C. ying Lee, L. Hetherington, and J. Glass, "Collecting voices from the crowd," in *Proc. of LREC*, 2010.

[7] A. Raux, B. Langner, D. Bohus, A. Black, and M. Eskenazi, "Let's go public! taking a spoken dialog system to the real world," in *Proc. of Interspeech*, 2005.

[8] ITU P series Rec, "Parameters Describing the Interaction with Spoken Dialogue Systems," *ITU, Geneva*, 2005.

[9] M.M. Shoukri, *Measures of interobserver agreement*, 2004.

[10] J.R. Landis and G.G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[11] K.P. Engelbrech, F. Gödde, F. Hartard, H. Ketabdar, and S. Möller, "Modeling user satisfaction with Hidden Markov Model," in *Proc. of the SIGDIAL*, 2009.