EVALUATION OF A SPOKEN DIALOGUE SYSTEM FOR CONTROLLING A HIFI AUDIO SYSTEM

F. Fernandez Martinez[1], *J. Blazquez*[1], *J. Ferreiros*[1], *R. Barra*[1], *J. Macias-Guarasa*[2], *J.M. Lucas-Cuesta*[1]

[1] Speech Technology Group. Dept. of Electronic Engineering. Universidad Politécnica de Madrid.
[2] Department of Electronics. University of Alcala.

ABSTRACT

In this paper a Bayesian Networks, BNs, approach to dialogue modelling [1] is evaluated in terms of a battery of both subjective and objective metrics. A significant effort in improving the contextual information handling capabilities of the system has been done. Consequently, besides typical dialogue measurement rates for usability like task or dialogue completion rates, dialogue time, etc. we have included a new figure measuring the contextuality of the dialogue as the number of turns where contextual information is helpful for dialogue resolution. The evaluation is developed through a set of predefined scenarios according to different initiative styles and focusing on the impact of the user's level of experience.

Index Terms— spoken dialogue systems, Bayesian networks, usability, evaluation

1. INTRODUCTION

It is quite difficult to evaluate dialogue systems. In addition to the lack of evaluation standards within the dialogue community, at the same time, it is difficult to find performance figures from real-world applications that can be extrapolated to other systems or be worldwide accepted, as all of them are directly related to one specific dialogue system. Nonetheless, there is a general agreement on "usability" as the most important performance figure [2][3][4], even more than others widely used like "naturalness" or "flexibility".

Therefore, besides quality and efficiency metrics, automatically logged or computed, subjective tests have also been performed in order to assess the impact of the capabilities of the system on user satisfaction and to get a valuable insight on the shortcomings and advantages of the system.

2. PROTOTYPE DESCRIPTION

A conversational interface that allows users to drive a commercial Hifi audio system using natural language sentences is under evaluation. This system can be normally controlled by an IrDA remote control. Instead, users are going to control the system using a microphone. This interface makes the translation of speech into the corresponding IrDA commands needed to perform a specific set of control actions according to the user's intention. Its dialogue manager is based on BNs and it is provided with a set of domain independent dialogue strategies for handling contextual information (e.g. history of dialogue, information regarding the system's current setup or state, expert domain application knowledge, etc.). These strategies provide the ability to deal with dialogue phenomena such as: ellipsis, anaphora or deixis. It is possible to find a detailed description of the system, its architecture and the dialog strategies in [1].

3. EVALUATION SCENARIOS

A set of 15 dialogue goals were defined covering the typical functionality available in commercial Hifi systems (e.g. playing, recording, radio, volume, disc, track, or tape selection, etc.). From this goal set, different types of scenario were designed according to different initiative styles and task complexity levels. The whole set of defined scenarios added up to a total of 45 grouped into the following categories:

- **Basic** (strongly guided tasks aimed at demonstrating mandatory functionality): 23 in total. The user has to try to fulfil just one single dialogue goal (e.g. *"The user should try to stop the current disc playing"*). The dialogue context (dialogue memory and system's state) is timely prepared according to the targeted goal.
- **Advanced** (less guided but more complex scenarios): 19 in total. On one hand this type of scenario is aimed at demonstrating the flexibility, robustness, and adaptation capabilities of the system. On the other hand, users have to try to fulfil multiple dialogue goals (e.g. *"The user should try to play a particular track without referring to the specific disc the track belongs to"*). Similarly to the "basic" case, the dialogue context is timely prepared according to the targeted goals.
- **Free** (absolutely absence of guidance): 3 in total. This time the user is absolutely free to decide what to do with the system. Unlike the other two, the starting dialogue context is always set to a default (i.e. empty memory and system switched off).

Table 1. *Objective evaluation results: including partial results for each scenario type and user group.*

	Scenario Type			User Expertise		
	Bas	Adv	Free	Beg	Exp	**ALL**
Length (s)	67.6	92.7	167.0	92.3	92.9	92.6
Length (turn)	5.0	7.4	16.3	7.5	7.7	7.6
Help %	0.9	0.2	0.0	0.6	0.0	0.3
Cancel %	0.0	0.0	0.0	0.0	0.0	0.0
Timeout %	0.0	0.7	0.0	0.3	0.5	0.4
ASR reject %	13.9	13.8	7.7	18.1	7.6	12.8
NLU reject %	0.9	1.1	0.0	0.2	1.3	0.8
OOD turn %	1.8	1.2	0.8	0.8	1.6	1.2
ASR Rep %	3.6	3.5	2.0	3.4	2.9	3.2
NLU Rep %	3.6	5.6	3.7	4.6	4.9	4.8
Context turn %	47.6	57.7	53.9	56.1	53.7	54.8
System req %	30.7	30.3	15.9	29.9	25.1	27.3
Turn eff (turn)	0.66	0.67	0.63	0.69	0.63	0.66
Exec period (s)	8.89	8.34	6.49	8.56	7.63	8.03

4. DATA COLLECTION

The system has been tested by students at the UPM. A total of 15 speakers, 3 female and 12 male, were recruited for the evaluation targeting an age range between 23 and 28 years. Participants were classified as "novice" (7) or "expert" (8) according to their experience level using speech interfaces.

Each participant was required to complete 10 dialogues or scenarios according to the following distribution: 3 basic, 6 advanced and 1 free scenarios. Thus, a total of 150 dialogues were collected. User-system interaction took place in a specially prepared living room equipped with the Hifi system where users promptly received a brief description of the tasks they were requested to accomplish for each scenario.

4.1. Collected metrics

Data labelling through manual transcription is a costly and time consuming work that has not been done. Instead of that, a combination of dialogue quality and efficiency measures have been automatically logged or computed [4]. Some of the considered metrics have been expressed as the percentage of turns where the specific event takes place.

4.1.1. Dialogue quality metrics

Help requests: the user interrupts the interaction to request some help.
Cancellation requests: the user explicitly quits the ongoing dialogue starting a new one.
Silence timeouts: the user did not respond in time.
Speech recognition rejections: either a low confidence ASR result or simply the absence of result for a particular phrase.
Understanding rejections: sentences for which no NLU result is obtained in spite of a valid recognition result.
Out of domain turns: user turns for which no goal is inferred as present or positively identified.
Number of repeated utterances: number of consecutive utterances for which either the same ASR result or the same NLU result is obtained. Both results are provided.

4.1.2. Dialogue efficiency metrics

Dialogue length: number of turns on average needed to complete a scenario. It can also be expressed as the elapsed time for the dialogue. Both results are provided.
Context-dependent turns: it can be estimated as the amount of turns in which some of the contextual information handling strategies is applied successfully.
System requests: number of turns where the system requests to the user some missing or deliberately omitted information.
Turn efficiency: average number of turns needed to accomplish a particular goal (execute a specific action).
Execution period: similar to the previous one, average time needed to accomplish a particular goal.

4.1.3. User satisfaction surveys

In order to get subjective ratings of the system we conducted user satisfaction surveys. First we requested users to rate task or scenario success after each scenario (Figure 2). Finally, after the evaluation, users also filled out forms (Table 2) rating on a 1 to 5 scale (i.e. 1-very poor, 2-poor, 3-fair, 4-satisfactory, or 5-highly satisfactory) typical: ASR and TTS performance, task ease, system response, etc.

5. OBJECTIVE EVALUATION

Instead of an individual analysis of each quality and efficiency metric (Table 1), we horizontally discuss all of them focusing on the most relevant issues. A detailed analysis aimed at measuring the impact of both the type of scenario and the user expertise on system performance is also presented.

5.1. On the initiative style

Each scenario type correlates closely with a particular initiative style. Results from [5] show that the initiative style significantly impacts system performance and user behavior.

"Basic" scenarios are more related to system's initiative whilst "advanced" and mostly "free" scenarios are much more related to mixed and user initiatives respectively. On the other hand, a more user-driven scenario means more open user utterances (i.e. "advanced" and "free" scenarios) that could be harder to recognize and understand and therefore lead to lower system performance.

On the contrary, most of the metrics improved instead of worsening. The dialogue length, as could be expected, increases as the user is allowed a greater initiative (e.g. the average "free" scenario length doubles the overall). However, a longer length does not mean a worse system's performance,

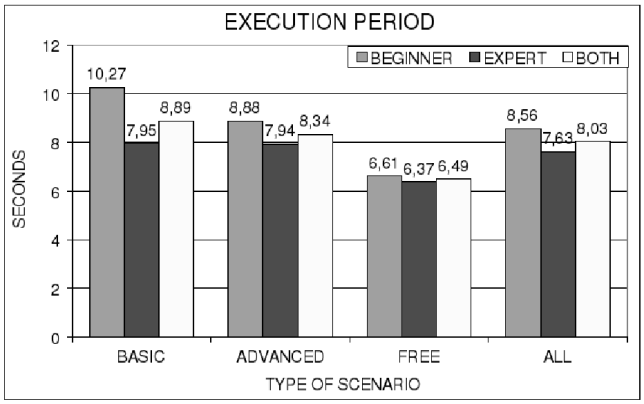

Fig. 1. Detail of the "Execution Period" results.

but just the user's intention of exploring the available functionality. "Turn efficiency" always dropped below 1 meaning that more than one action, almost two, can be accomplished in just one turn. Moreover, some measures clearly show that the system's behaviour is better for "free" scenarios than for any other type (i.e. "System request turns"). These results can be explained by a two-side effect.

First, "free" scenarios lack of a specific target. Consequently, the user may focus the dialogue in two different ways. Regarding "basic" and "advanced" scenarios the user may not clearly know how to achieve the requested goals. Thus, the user probably needs to find out how to do it during the dialogue itself resulting in a less efficient and fluent conversation as can be deduced from the slightly worse results presented in Table 1 for these types of scenario. On the contrary, when facing "free" scenarios the user may tend either to reuse some expressions or to address the system in an easier way. On the other hand, the absence of a specific purpose allows the user not to remain blinded by a requested target thus deriving in more efficient and fluent dialogues.

5.2. The value of experience

Already accomplished scenarios provide the user a significant accumulated experience that is helpful to complete the remaining scenarios. That experience has a strong impact on dialogue performance. The number of ASR rejections or system's requests are both good examples of this. The former rose up to almost 14% for both "basic" and "advanced" scenario sets (both together roughly correspond to the 80% of the user's turns). However, as can be observed in Table 1, the metric significantly decreased to 7.7% for "free" scenarios.

The user-system interaction improves as the user learns how to address the system and improves his dialogue skills. Furthermore, experts are more familiar with dialogue timing and turn-taking issues so they are supposed to obtain a better performance. As can be observed in the same table, there are less help requests and less ASR rejections and repetitions for expert users than for beginners. At the same time, expert users induce less system's requests and achieve a better turn efficiency (i.e. lower execution period), thus enabling a more fluent conversation. On the contrary, novice users obtain better results regarding NLU rejections and OOD utterances. A possible reason for this could be that newbies tend to use shorter and less complex sentences.

The efficiency improves gradually (i.e. execution period decreases) regardless of user's experience (see Figure 1). Therefore, users gradually need less time to perform an action, thanks to the increasing skill they are gaining. By comparing both "basic" and "free" results we can conclude that newbies have improved their dialog skills (i.e. execution period reduction thus improving turn efficiency) by 36% while experts have done it by 18%. Another interesting issue deduced from that figure is that, in the end, newbies have almost reached the ability of the experts, reducing gradually the gap regarding the execution period from 29.2% (2.32 seconds for "basic" scenarios) to only 3.6% (0.24 seconds for "free").

5.3. On the contextual information usefulness

Benefits obtained through the use of contextual information handling strategies can be regarded as highly significant. More than half of the turns, 55%, rely on the contextual resources available in the system.

In connection with the contextual capabilities, we have to analyze the amount of system's requests. This result should be, supposedly, fairly limited by the contextual capabilities. Thanks to the recovery of contextual information a significant number of requests are saved. Less than one of every three turns, 27.3%, involves a system's request. This number would be expected to increase, at least, up to 54.8% without contextual resources. In addition, both the expertise and the initiative style effects are evident as system's requests corresponding to "free" scenarios are almost 50% lower on average than that of other types. This is such a valuable result, specially considering that "free" scenarios lack of a starting dialog context coherent with the proposed tasks.

6. SUBJECTIVE EVALUATION

The subjective evaluation is based on the analysis of both a survey filled out by every participant just after the evaluation, and the user satisfaction rates obtained for each scenario.

6.1. On the initiative style

In order to have a running estimate of "task completion rate" without having to manually label each scenario, each user was asked about the level of success achieved for each scenario.

As can be observed in Figure 2 this rate was 0.6 points higher for "basic" and "advanced" scenarios than for "free".

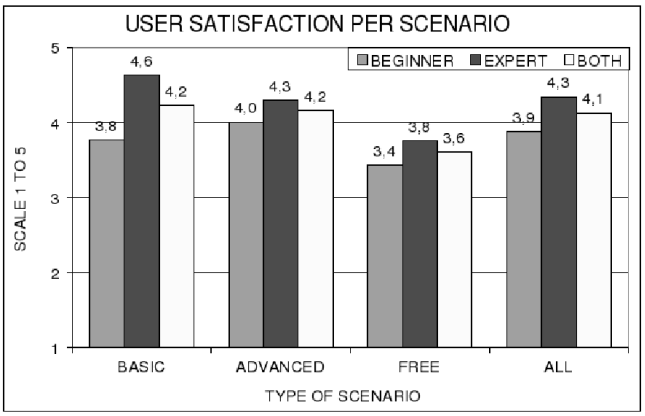

Fig. 2. Scenario success assessment.

Due to the absence of a specific purpose (i.e. "free" scenarios), users tend to explore the available functionality. This freedom favours situations in which the user may attempt to do something that actually is not allowed in the system, thus leading to mistakes, and thus to a worse valuation. This score (3.6 over 5) correlates with the assessment made by users of the available functionality (i.e. Q8 in Table 2), 3.8 over 5.

6.2. The survey: summary of findings

Almost every question is better rated by expert users than by newbies. Along the same line, user satisfaction per scenario (Figure 2) also receives better scores from expert users.

Answers to question 1 can be considered as a sign of good ASR and NLU performance. Although system's prompts were shortened as much as possible removing non-critical information, question 4 showed that users, specially novices, found the listened prompts a little annoying and too verbose to be duly assimilated. This result is quite consistent with those corresponding to questions 2 and 3. Actually, a poor understanding of the system's prompts may cause a worse valuation of the system's understanding ability and behaviour. On the other hand, though not significantly penalizing question 5, some users felt a bit frustrated due to the lack of barge-in.

Question 6 denoted an excellent easiness of use (also reflected by question 9 results) while answers to question 10 showed that participants were fairly determined to use the system. Newbies were slightly more willing to use the system than experts, surely due to a better knowledge of the real limitations of the system (and also of its quality, thus rating the system better).

For the same reason, beginners tend to oversize expectations of what the system can actually do. Therefore, regarding question 7, beginners' expectations fulfilled in a lower level than experts' ones.

Finally, the global assessment of the system (i.e. Q9) is 4.0. Besides being a very positive evaluation, it is well correlated with the user satisfaction per scenario, 4.1 (Figure 2).

Table 2. *Survey results for each user group.*

	Survey	Beg	Exp	Both
1	Did the system understand what you said?	3.57	3.88	3.73
2	Was the system able to act coherently with dialogue context (e.g. system's status, previously executed actions, etc.)?	3.43	4.25	3.87
3	Did the system execute requested actions?	3.71	3.88	3.80
4	Was the system easy to understand?	3.57	3.88	3.73
5	Does the system respond quickly enough?	3.86	3.88	3.87
6	Is the system easy to use?	4.29	4.50	4.40
7	Did the system work the way you expected?	4.00	4.63	4.33
8	Was the available functionality acceptable?	3.81	3.81	3.81
9	How would you rate the system globally?	3.86	4.13	4.00
10	Do you think you'd use the system regularly instead of the IrDA remote control?	4.14	3.75	3.93

7. CONCLUSIONS

More natural, flexible and robust dialogue is possible thanks to the suggested BN based dialogue modelling approach [1]. This is supported by a good user satisfaction rate and the obtained results for the collected metrics.

In this regard, the strategies for handling contextual information have been proved to be essential, saving a significant amount of system's requests, and thus speeding up the dialogue. Experience has turned to be another important factor regarding dialogue performance. However, thanks to the negotiation capabilities of the BN based dialogue manager, the user is able to rapidly react improving his dialogue skills and resulting in more fluent and efficient dialogues. On the other hand, according to the ratings obtained from the performed survey, it is clear that there is much room for improvement at several levels (e.g. NLU and response generation modules).

8. REFERENCES

[1] F. Fernandez et al., "Speech interface for controlling an hi-fi audio system based on a bayesian belief networks approach for dialog modeling," in *Eurospeech*, Lisboa (Portugal), 2005, pp. 3421–3424.

[2] M. Turunen et al., "Evaluation of a spoken dialogue system with usability tests and long-term pilot studies: Similarities and differences," in *ICSLP*, Pittsburgh, USA, 2006, pp. 1057–1060.

[3] S. Schulz and H. Donker, "An user-centered development of an intuitive dialog control for speech-controlled music selection in cars," in *ICSLP*, Pittsburgh, USA, 2006.

[4] M.A. Walker et al., "Towards developing general models of usability with paradise," *NLE, Special Issue on Best Practice in Spoken Dialogue Systems*, 2000.

[5] A. Raux et al., "Doing research on a deployed spoken dialogue system: One year of let's go! experience," in *ICSLP*, Pittsburgh, USA, 2006.