Evaluation of Voice User Interface Design for Task-completion Problem

Advait Ambeskar
Computer and Information Science and Engineering
University of Florida
Gainesville, Florida
Email: ambeskaradvait@ufl.edu

Akash Barve
Computer and Information Science and Engineering
University of Florida
Gainesville, Florida
Email: akash.barve@ufl.edu

Sandesh Joshi
Computer and Information Science and Engineering
University of Florida
Gainesville, Florida
Email: sandeshjoshi@ufl.edu

Susmitha Are
Computer and Information Science and Engineering
University of Florida
Gainesville, Florida
Email: susmitha.are@ufl.edu

***Abstract*—Voice user interfaces (VUIs) are becoming a ubiquitous part of the modern-day technology ecosystem championed by their availability, simplified design, and growing acceptance. With improvement in the connectivity of devices, VUIs provide a low-latency solution to engage user-space without burdening the user with WIMP-based interfaces. While the use of VUIs is widespread, it is increasingly important to gain an insight and develop a better understanding of the design specifications that is inherently unique to the 'window-less' interface. Understandably, the design specification for each problem-space in the VUI domain cannot dictate a blanket solution. A type of problem that has been increasingly solved with the use of VUIs is the task-completion problem. It dictates a logical flow of successive instructions that engage and guide the user into completion of a given task.**

In study of this problem, the evaluation of the design specifications takes place using two major parameters. The study evaluates the effect of compression of instructions and shortness of task on accuracy and engagement time. It also evaluates the effect of conversation structure and logic flow on the user. The study also attempts to engage the user to determine the effect of short-term goals on the user performance. The aim is to understand and evaluate the factors to engage consistent design in the task-based problems designed for the VUIs.

***Index Terms*—Voice user interface, Speed-accuracy tradeoff, step-wise task, Design recommendations, Dialogue Structure**

I. Introduction

Recent technological advancements have brought in 'Post-WIMP' interfaces to the forefront; voice user interfaces (VUIs) leading the charge [10]. In such voice user interfaces, an important element of context and study is the perception of the interface, resting solely on the characteristics presented by the said interface. Cohen in [9] examined the effect of human-like personality traits displayed by the VUIs on human participants, exemplifying the need for personification of the digital assistant. The idea behind the development of a VUI has evolved to encompass the need to not just provide an 'unconventional' interface for humans to interact with the digital devices, but also to increase productivity, assist in tasks and provide a well-rounded service to improve quality of life.

The boom in the use of the VUIs is a direct result of the availability of various customizable features. The design of the VUIs, essential to a good user experience, heavily relies on the idea that the interactions draw influence from human conversational flow. The design of the VUIs is heavily inspired by the logic that dictates human conversation. This logic is inherently referred in development of the VUIs as noted through various structural components such as implicature, which dictate the flow that allows coherent understanding of the proceedings [3]. When further abstracted from the nature of human-to-human conversations, the design of the VUIs is highly dependent on the needs of the users. Since VUIs inherently are designed to mimic human interactions, the recommendations that dictate the modifications to the existing design of the VUIs is limited by our understanding of human needs, and the absence of a graphical interface [9].

An important consideration in design of any VUI is the limitation to its ability to speak. The need for the inclusion and study of characteristics that build a better conversation is an important design consideration, due to the underlying human bias, proven by psychology. Mairesse in [11] describes the dynamic nature of the nuances introduced through speech with a conversational agent is advantageous due to the underlying human-factor. However, it is an important design consideration is the limitations of the VUIs. Thus, it remains important to understand that the engagement and user satisfaction is derived based on task completion [8].

The proposed idea revolves around the study of various parameters using the VUIs built on top of the "Amazon Web Service" to evaluate standards for the design of the task-completion problem. A task completion problem is one that involves a succession of instructions that allow an end product that is a result of a continual output from each step. The idea is to understand and study the effect of various parameters of VUI design on the progression and completion of the task focusing on the user space evaluation; drawing correlations through measurement of time, accuracy, information retrieval and a qualitative evaluation from the user. The built system employs a 'Wizard of Oz' format, which allows for greater flexibility in evaluation and maintains variability of the test-case. A 'Wizard of Oz' system provides a developer-mediated test space for ease of evaluation where the output from the user is processed by a middleware system, usually the evaluator, to simplify maintenance of the test flow, and the accuracy of the evaluations.

II. Related Work

Human like personality traits influence people's perceptions of Voice User Interfaces. We see in [1] that more personification tends to a more user satisfaction. In the case of interfaces for task completion, our approach is to make the steps for these tasks more human like which would be divide the steps to the easiest possible level. Kamm in [2] identifies components of Voice applications that make them successful, one of them being dialogue flow which needs to be controlled by the user rather than the application, which can be implemented in instructional applications by allowing users to change the flow to their required state. We present this feature in all tasks used in our study. Kamm in [2] also presents an argument about identifying situations where there is some level of ambiguity, we try to tackle this by trying to present a description of what the origami artefact should look like wherever possible and ambiguity might occur. The same study also suggests for effectiveness, it is important to communicate the system expectations and system services to the user beforehand.

While presenting audio navigation patterns in [3] and [4], the authors focus on limits of the short-term memory of users. The authors present context, affecting forces, solution and structure for the intent of dividing the amount of information being delivered to the user at once into manageable pieces. They suggest grouping by number or according to some set criteria and also instruct to add commands for back and forth navigation. In a similar way, the paper also presents the intent of helping users stay oriented. This provides the designers questions to ask themselves, the answers to which would help user in navigating the system efficiently. These issues help in design of the first experiment of our study. Schnelle and Lyardet in [4] also presents a few dialog strategies which elaborates on guiding the user to a certain piece of information, this section also provides the first occurrence of using a Wizard of Oz system for studying this problem.

Our objective is to help practitioners design better voice user interface for step-wise task completion problem. The VUI framework in [6] helps to understand the known limitation of voice user interfaces and provides a solution to overcome them. Also, good dialogue design is an essential component for VUI. It improves the understanding, usability of service and reduces the risk of customer dissatisfaction. Stentiford and Popay in [7] discussed the dialogue style guide illustrating dialogue wording, structure, complexity and cognitive load on user. We have employed this design style in our experiment. Dialogue design is not always obvious and human evaluations are an important weapon in the battle to increase quality. That is the main reason for conducting the experiments to evaluate the performance based on the variation of dialogue design.

The limitations of the speech user interface must be apparent to users [8]. Task performance and user satisfaction vary considerably if limitations are not clear. The authors conducted a four-minute tutorial session to familiarize the system to novice users. User satisfaction ratings based on task completion, number of 'help requests' and mean recognition score are consistently lower to the non-tutorial group when compared to the tutorial group.

Grice in [3] provides a background towards design of VUIs as they inherently are dependent of logical proceedings of human conversations. Since human function is based on patterns learned through experiences and observations, a logical flow can be ascertained for the same. This knowledge can be further extended to understand conversations pursued by humans with other humans, machines, or interfaces. The context and structure of the sentences is key to understand the perceived meaning. Structures can be leveraged to improve quality of conversations, understand the implied meanings and the perceived human aspect (personality, verbal cues, tone) of the conversation. Cohen in [9] provides a framework to improve VUI by using factors that are closely related to the human conversational techniques. Since VUI essentially is a service that mimics human interactions, it provides a fair judgement and design recommendations that can be used to modify VUIs for more efficient performance and increased utility.

Our study focuses on size and the structure of instructions. Richard in [12] provided research literature on best way to construct instructions to enhance performance and learning of procedural tasks. A model [13] is suggested by Anderson to improve the design of procedural instructions.

A repository hosted at [14] was used to generate the base instructions for various Origami tasks. Since these tasks were explicitly written for the visually impaired children and adults, it provides a basis for easy conversion to voice instructions due to lack of dependence a visual medium.

III. Methodology

For the study, we have designed a 'Wizard of Oz' system. The system contains a text-to-speech (TTS) engine which converts English text into speech. We are using Amazon Polly TTS engine which is a service provided by Amazon Web Services. We are using easy origami instructions specifically written for the blind so that absence of visual cues won't hinder the participants. These instructions are fed to the TTS engine and converted to speech. Fig. 1 showcases the overall system designed for the study. We are conducting two controlled experiments where the participants will be presented with step-wise origami instructions. The participants are asked to create origami pieces following these instructions and the accuracy of their work is judged by observing them

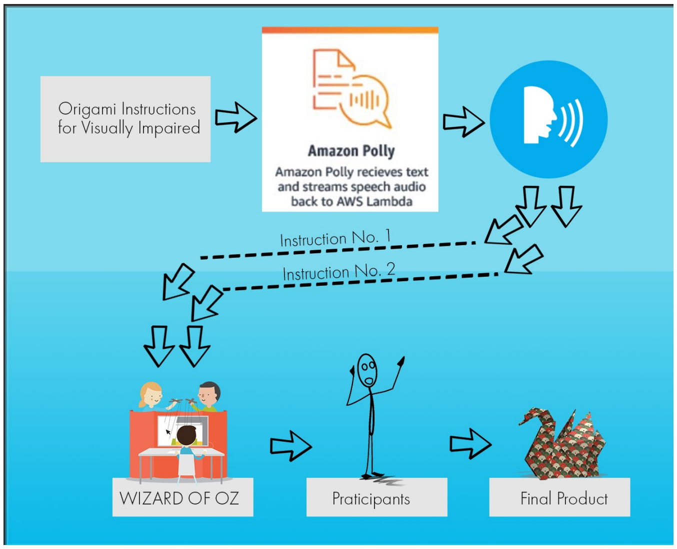
Origami Instructions for Visually Impaired
Amazon Polly
Amazon Polly recieves text and streams speech audio back to AWS Lambda
Instruction No. 1
Instruction No. 2
WIZARD OF OZ
Praticipants
Final Product

Fig. 1. Overview of the system used in the study.

for each instruction. The instructions are divided, organized and structured according to the experiments.

Baseline experiments were conducted to assign weights to each instruction which helped us assess the participants' accuracy of tasks. The baseline experiments consisted of four participants following the same instructions of each task. Average time was calculated for each instruction and weight was assigned to that instruction based on this time. Also overall time was recorded for each task between these four participants and whenever two tasks were compared side-by-side for a user, the time was normalized by the normalization factor calculated between those two tasks. For example, the average time to create a origami cup was less than a origami house. Whenever these two task were compared, a normalization factor of 1.37 (which was calculated by the timings we got for them) was multiplied to the cup tasks final timing to get fair comparison. Same was calculated for Fir tree and Party hat origami task and a normalization factor of 1.30 was observed. For both the experiments, participants performance was compared with two versions of the tasks performed by themselves. So, factors such as cognitive ability, origami expertise and native language would not distort the results.

Our study consisted of a total of 11 participants. All of them were students at the University of Florida within the age range of 18 to 28 years. Of the eleven participants, four were native English speakers and the rest were non-native English speakers with proficiency in English. None of the participants had previous experience with origami artwork. Out of the eleven participants, one participant opted out of the study due to personal reasons. Also, the total time to participate in both experiments was in the range of 60 minutes to 75 minutes per participant. The first experiment had ten study participants, with four native English speakers while the second experiment had five participants with two native English speakers.

For both the experiments, the following was done. Each participant was asked to do two simple origami tasks of similar difficulty. The participants were given a paper to make the origami by following the step-wise instructions delivered by the system. After each instruction is narrated by the VUI, the participants were supposed to perform the task specified. The participant had the option to ask the voice user interface to repeat the instructions. They could also jump to previous instruction by asking the voice user interface. Once they were done with a step, they could ask the VUI for next instruction. The accuracy of the final product created by the participants was judged by observing how each step is performed. Percent accuracy was assigned in the end of each task. The participants were also timed for the tasks and we also observed the repetition count of each step.

A. *Experiment One*

For the first experiment, the independent variable was the size of an instruction. This experiment was conducted to see whether the size of the instruction given to a user could affect the performance. We asked the participants to make a origami house and a origami cup. In one task, the instruction were smaller and atomic which couldn't be divided into further steps, while the other task had longer and composite instructions. From the related literature, we've observed that short and simple instructions are more efficient in a dialogue. Based on this we hypothesized the following:

Hypothesis

- The tasks with smaller/atomic instructions will have higher final accuracy when compared to tasks with composite instructions.
- Tasks with atomic instructions will take longer time to complete as they have more steps and pauses when compared to tasks with composite instructions.

B. *Experiment Two*

Information structure is an important parameter that drives human conversations. Thus, the study of the effect of structural components of conversations is an important parameter in design of the VUI. Information structure represents the flow of logic in the conversation which is increasingly important when the given flow dictates the understanding of the presented task. Thus, in the given set of problems, evaluation of this structure remains highly imperative.

The evaluation of this parameter has been designed with the need for consistent flow of logic. The proposed idea is to assess the effect of logical flow of the conversation and the possible disruption to the user. It engages the idea that non-linear flow of conversation can potentially benefit the user. The idea is based on the need for the VUIs to be more 'human-like' , and thus follow the non-linear aspect of general human conversations.

The experiment focuses on information structure. Participants were given two origami tasks. The first task has premonitions/warnings, repetition handling, while the second task is without them. We assessed the accuracy, time measure and

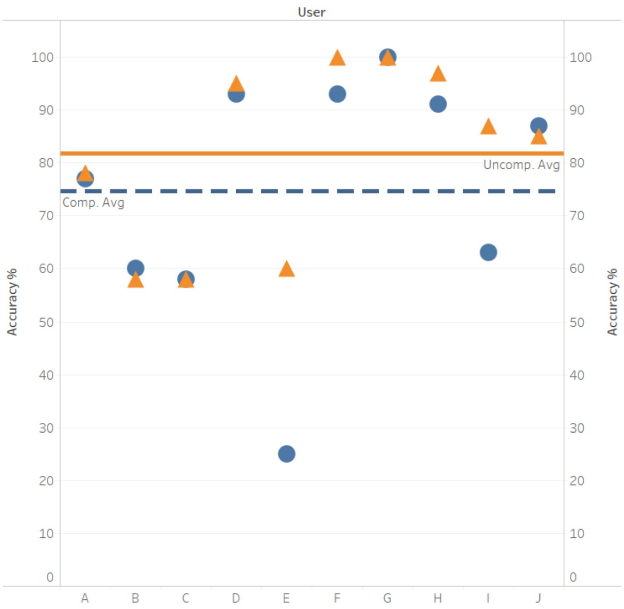

Fig. 2. Accuracy results of tasks with smaller instruction and composite instruction for all users, the orange triangle represents accuracy for smaller/compressed instruction whereas the blue circle represents accuracy for tasks with composite/uncompressed instructions. The lines represent the average accuracy for the respective tasks.

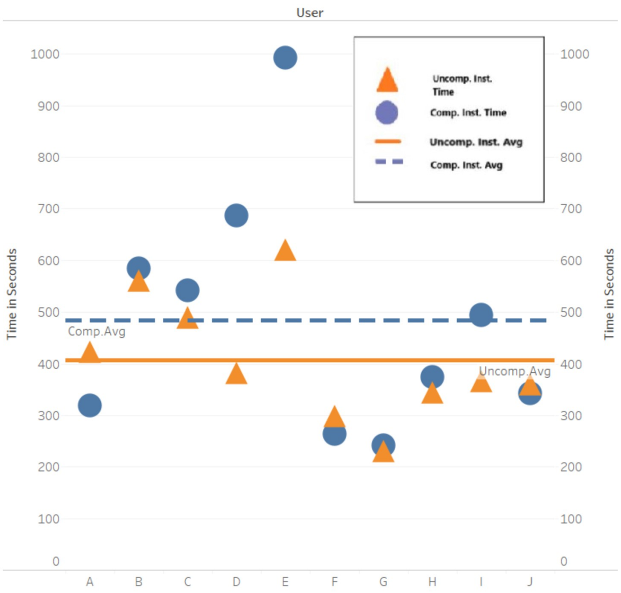

Fig. 3. Time results of tasks with smaller instruction and composite instruction for all users, the orange triangle represents time taken in seconds for smaller/compressed instruction whereas the blue circle represents time taken in seconds for tasks with composite/uncompressed. The lines represent the average time taken for the respective tasks.

likability of these two versions. Premonition provides the users with a nudge and allows them to anticipate the status of the system in the future.

The experiment evaluates the effect of structural flow on information disbursal and its effect on the user experience. The idea revolves around provisions in conversation between VUIs and user that allow the user to anticipate the 'next' step in the progression and providing 'hints' for the same. This evaluation would enable derivation of correlation between conversation flow and user accuracy for task completion. Such conversation flow can be explained through cursory warning instructions provided as part of the experiment to caution the user about expected changes in the task that is being performed and their intended effects. The example of such a variation would be instructions such as 'Please expect many folds along the horizontal center line' while performing the task before the 'folds' are actually performed.

Hypothesis

- Changing the flow of the interaction structure by properly adding premonitions wherever required, positively affects the user likability.
- Changing the flow of the interaction structure by properly adding premonitions wherever required will increase the accuracy of the interaction.

IV. Results

The study established the effect of independent variables such as number of instructions and the structure of instructions on the engagement time and accuracy of the result.

A. Experiment One

The measure of accuracy was noted at each instruction. The overall accuracy of task completion was measured through a weighted aggregate of accuracy at each step. Fig. 2 showcases the accuracy of each user for both their tasks and the average accuracy as provided for each task. It can be deduced from the graph that whenever there is a difference in accuracy between the two tasks for a user, the difference is significant.

The time for each task was also normalized according to the accuracy of the participant. All the completion times in the time graph is adjusted according to the accuracy. If the accuracy for task was fifty percent, the time was increased by a factor of two. Fig. 3 showcases that whenever there is a significant difference in the time taken for a task for a user, tasks with composite instruction took more time. This works against the second hypothesis we made. It was observed that participants kept repeating the complex steps for the tasks with composite instructions which increased the overall time taken to complete the task. Additionally, the study participant were also asked to take a survey to give their feedback for a preferred approach. Out of ten participants, seven participants said they liked atomic/ smaller instruction approach better while two participants liked the composite instruction approach. Only one participant was neutral.

B. Experiment Two

The participants were asked to make origami party hat and fir tree. The accuracy measure was calculated through a weighted measure of observed accuracy and factor derived from the baseline study. For this experiment, we were more

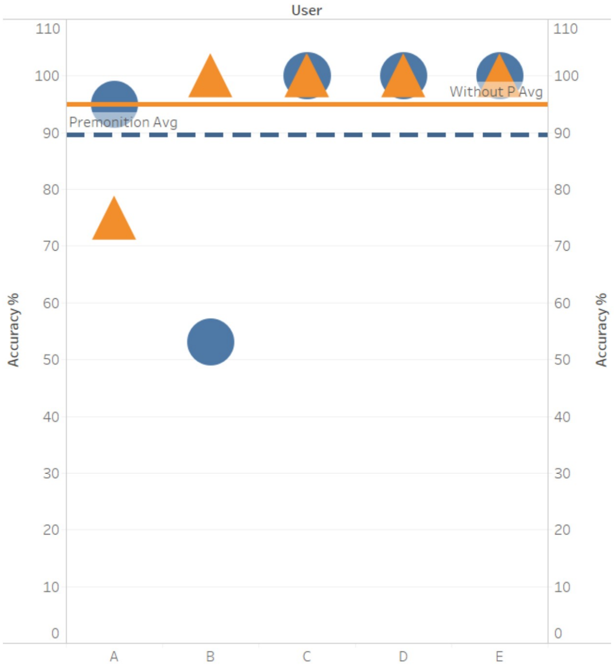

Fig. 4. Accuracy results of tasks with warnings/premonition and tasks without premonitions for all users, the orange triangle represents accuracy for tasks without premonition whereas the blue circle represents accuracy for tasks with premonition. The lines represent the average accuracy for the respective tasks.

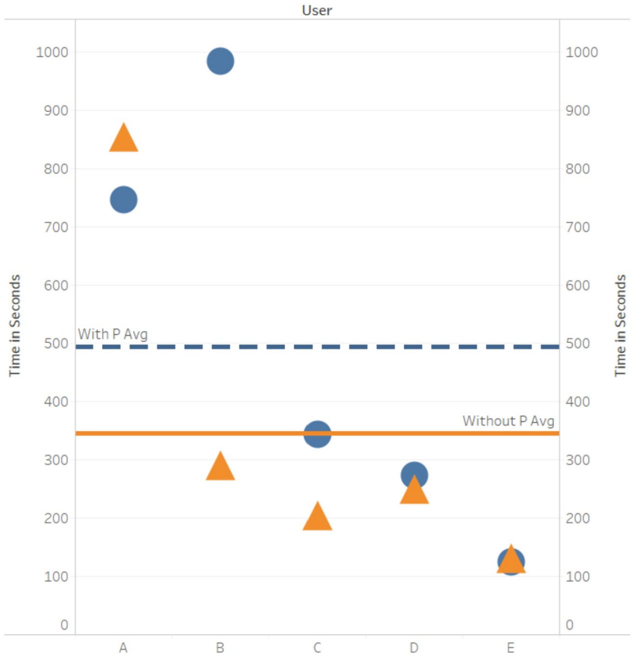

Fig. 5. Time results of tasks with warnings/premonition and tasks without premonitions for all users, the orange triangle represents time taken in seconds for tasks without premonition whereas the blue circle represents time taken in seconds for tasks with premonition. The lines represent the average time for the respective tasks.

interested in qualitative results than quantitative results. The qualitative results support the first hypothesis made for this experiment. 4 out of 5 participants preferred the tasks when instructions had some warnings in them. The quantitative results do not support the second hypothesis for this experiment. As observed through Fig. 4, there is not a significant difference between the task accuracy of the participants. But we can not be very confident as the sample space is too small to make any observation. Fig. 5 expresses the time measure for each task. As expected, tasks with warnings in instruction had more content and took more time to complete as compared to tasks without warnings.

V. Discussion

The results from the experiments were interesting, even if not surprising. We had two hypotheses for the first experiment. The results support the first hypothesis, which states that the tasks with smaller atomic instructions will have better accuracy. However, the experiment led to interesting observations regarding the second hypothesis. The results do not support the second hypothesis for the first experiment, which states that tasks with smaller instruction will take more time to complete. It was observed that participants kept repeating the complex steps for the tasks with composite instructions which increased the overall time taken to complete the task.

For the second experiment, the results support the first hypothesis but not the second. While there is no significant effect on accuracy measure observed, the user preference indicated a positive bias towards a structured premonition-based interaction.

Below are some insights that were drawn from the study.

Break down the difficult steps: It was observed that when it came to difficult steps, the participants kept repeating the composite instruction. We recommend that the complex steps should always be broken down into coherent atomic steps for higher accuracy.

Easy small steps frustrate the users: It was observed that when easy steps were represented in atomic way, participants seemed frustrated and some of them even mentioned this in the post experiment questionnaire. It is recommended to combine the easier continuous steps into one. This will give an organic flow to the VUIs instructions and will add naturalness to them.

More instruction does not mean more time: Even though the smaller instructions have more break time between them, they make up for the time by conveying the instruction with coherency with little repetition. The compressed steps were repeated several times by participants increasing the overall time taken to finish the task, which is reflected in the results.

Checkpoints are good: Participants seemed very confident and relieved when their own product matched the checkpoint. Also, when participants were not following the steps correctly, the checkpoints helped them realize the problem and they went back to previous steps, which in turn increased accuracy. The checkpoints help establish a common ground between the system/ VUI and the user.

Warn the users ahead: It is recommended that giving subtle suggestion pertaining what to expect next does help the users to follow steps properly. This will make the user aware of the context. It was observed that in task when users were not warned, they looked quite surprised at certain instruction which asked them to do something they were not expecting.

VI. Conclusion and Future Work

The design and structure of instructions is key for successful VUI's. We see that such voice interfaces remove the complexity of using screens and having to deal with controls when trying to do something along with interaction with the interface. Proper study and design of these VUIs will help them gain momentum for day to day use.

It was observed that native English speakers had higher accuracy when compared to non-native speakers. The participants in this study were mix of native and non-native speakers. Even though the non-native speakers were proficient in English, two separate experiments taking native and non-native speakers as participants could be performed in the future. This might bring out some more interesting results and insights. Additionally, we feel that a bigger sample size would help the study to be more conclusive. Given more resources and a proper setting, a similar experiment could be conducted with a VUI giving instructions for other related tasks like cooking. This would feel more natural to the users as these are the use cases that are being used by a large variety of voice-enabled smart devices currently sold in the market.

References

[1] Purington, A. Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017, May). Alexa is my new BFF: social roles, user satisfaction, and personification of the amazon echo. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 2853-2859). ACM.

[2] Kamm, C. (1995). User interfaces for voice applications. Proceedings of the National Academy of Sciences, 92(22), 10031-10037.

[3] Grice, H. P. (1975). Logic and conversation. 1975, 41-58.

[4] Dirk Schnelle, Fernando Lyardet, and Tao Wei. Audio Navigation Patterns. InEuroPLoP 2005 Conference Proceedings, 2005.

[5] Schnelle, Dirk, and Fernando Lyardet. ”Voice User Interface Design Patterns.” EuroPLoP. 2006.

[6] Bernhard Suhm. Towards Best Practices for Speech User Interface Design. E UROSPEECH 2003 - GENEVA

[7] F W M Stentiford, P A Popay. The design and evaluation of dialogues for interactive voice response services. .BT Technology Journal; Ipswich Vol. 17, Iss. 1, (Jan 1999): 142-148

[8] Kamm, Candace A. / Litman, Diane J. / Walker, Marilyn A. (1998): ”From novice to expert: the effect of tutorials on user expertise with spoken dialogue systems”, In ICSLP-1998, paper 0883.

[9] Cohen, Michael H., Michael Harris Cohen, James P. Giangola, and Jennifer Balogh. Voice user interface design. Addison-Wesley Professional, 2004

[10] Van Dam, A. (1997). Post-WIMP user interfaces. Communications of the ACM, 40(2), 63-68

[11] Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of artificial intelligence research, 30, 457-500.

[12] Elsa Eiriksdottir, Richard Catrambone. Procedural Instructions, Principles, and Examples: How to Structure Instructions for Procedural Tasks to Enhance Performance, Learning, and Transfer. Georgia Institute of Technology, Atlanta, Georgia

[13] Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 1322–1340.

[14] http://accessibleartsandcrafts.blogspot.com/