# Spoken Dialogue Interfaces: Integrating Usability

Dimitris Spiliotopoulos, Pepi Stavropoulou, and Georgios Kouroupetroglou

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Panepistimiopolis, Ilisia, GR-15784, Athens, Greece
{dspiliot,pepis,koupe}@di.uoa.gr

**Abstract.** Usability is a fundamental requirement for natural language interfaces. Usability evaluation reflects the impact of the interface and the acceptance from the users. This work examines the potential of usability evaluation in terms of issues and methodologies for spoken dialogue interfaces along with the appropriate designer-needs analysis. It unfolds the perspective to the usability integration in the spoken language interface design lifecycle and provides a framework description for creating and testing usable content and applications for conversational interfaces. Main concerns include the problem identification of design issues for usability design and evaluation, the use of customer experience for the design of voice interfaces and dialogue, and the problems that arise from real-life deployment. Moreover it presents a real-life paradigm of a hands-on approach for applying usability methodologies in a spoken dialogue application environment to compare against a DTMF approach. Finally, the scope and interpretation of results from both the designer and the user standpoint of usability evaluation are discussed.

**Keywords:** Speech, Spoken Dialog Interface, Usability, Usability Evaluation, Auditory User Interface, Human Computer Interaction, Accessibility, Computer Mediated Communication.

## 1 Introduction

The late years' research in communication, the world-wide-web and Human Computer Interaction (HCI) has led to significant advances in information access and revolutionized the way people exchange knowledge, learn and communicate. However, despite the newly designed approaches and technological advances, accessibility issues still constitute a barrier for a significant percentage of possible users, both mainstream as well as people with disability. People accessing the web or other educational material are usually presented with interfaces that, although make the information accessible; require either specific knowledge or expertise by the user. Moreover these are only designed as supplementary interfaces for use by special user groups in a specified modality or format as an alternative accessible means. Interfaces that are designed with only the accessibility as a solitary guide usually fail to present the real user with a usable means of communication.

A. Holzinger and K. Miesenberger (Eds.): USAB 2009, LNCS 5889, pp. 484–499, 2009.

Web technology is rapidly reaching maturity making it possible to practically use for most applications by the majority of potential users in the recent years. With high speed internet availability providing access to demanding multimodal services to all homes, most people can reap the benefits of real-time services ranging from voice banking to online socialising and beyond. Most high-level services are provided solely through web pages in the traditional point-and-click manner. In an effort to include the people with disability and boost *customer experience* most providers deploy spoken dialogue interfaces as a means for universal access as well as naturalness of information access.

The Web Accessibility Initiative (WAI) of the W3C and the emerging Web 2.0 [20] provide recommendations for creating, maintaining, extending and communicating accessible content. The creation of such content represents a harder task since it requires innovative design and multimodal implementation [28] according to the design-for-all directives while maintaining the scope of high-level user experience. The issues concerning universal accessibility in intelligent environments have been identified and deemed of utmost importance and benefit by the government bodies such as the European Union and the US Department of Education showing increased interest through appropriate accessibility initiatives, such as eInclusion of the eEurope [10].

Due to the complexity of natural language interaction, it is becoming very important to build spoken language interfaces as easily as possible using the enabling technologies. However, not all technologies involved in the process are of the same maturity, let alone standardisation. Furthermore, there is only a handful of platforms available for building such systems. Given the range, variability and complexity of the actual business cases it is obvious that the enabling technologies may produce working systems of variable usefulness due to design and/or implementation limitations.

As with all human-computer interfaces, speech-based interfaces are built with the target user in mind, based on the requirements analysis. However, they differ from traditional graphical user interfaces and web interfaces. The use of speech as the main input and output mode necessitates the use of *dialogue* for the human-machine communication and information flow. Information is received by the speech interface and presented to the user in chunks, much alike a dialogue between two humans. The input is recognised, interpreted, managed, and the response is constructed and uttered using speech. The naturalness is indeed far more enhanced than using forms and buttons on a traditional web interface. Apart from that, the use of Dual-Tone Multi-Frequency (DTMF) navigation over telephony is also a very common approach. However, such approach is seldom tailored for all users, is static and very unnatural. As a result, the performance of the resulting application does not always meet satisfactory levels in usability. It is, therefore, imperative that usability is ensured by design and verified by evaluation in a spoken dialogue interface when that is either deployed from the start or replaces a DTMF menu-driven approach.

The rest of this paper discusses the background of speech-based Human-Computer Interaction and elaborates on the spoken dialog interfaces. It explores what usability is and how it is ensured for natural language interaction interface design and implementation, both from the designer and the application deployment (business use) points of view. Hands-on experience on business-oriented spoken dialogue interfaces has shown that the designer can benefit from summative evaluation in the pre-deployment phase. The benefit from the transition to a spoken dialogue interface from a DTMF interface that provides a relatively large amount of services to a wide range of users is of utmost importance. This work presents the results of usability testing performed over a pre-deployed natural language dialogue system and a DTMF approach for the same real-life application domain, showing how usability engineering is applied for the design and implementation of a state-of-the-art voice user interface and evaluated by the target users.

## 2 Natural Language Interaction

People acquire communicative skills over time through the experience of using and operating the user interfaces. As the level of user adeptness rises, the speed and accuracy of the operation increases. The user adapts to the system and interacts more efficiently. The level of absolute efficiency corresponds to the actual system design, and can be assessed either as a full system or as a breakdown of its fundamental design modules or processes. In order to evaluate usability of such interfaces it is important to understand their design requirements and their architecture.

The use of speech as input/output for interaction requires a spoken language oriented framework that adequately describes the system processes. W3C has defined the Speech Interface Framework to represent the typical components of a speech-enabled web application [19].

Speech interaction is context-dependent. The context of the user input is analysed by the system in an attempt to understand the *meaning* and *semantics* within the application domain. The interaction itself is called a *dialogue*. The spoken dialogue interfaces handle human-machine dialogue using natural language as the main input and output. A general depiction of a Spoken Dialogue Interface is shown in Figure 1.

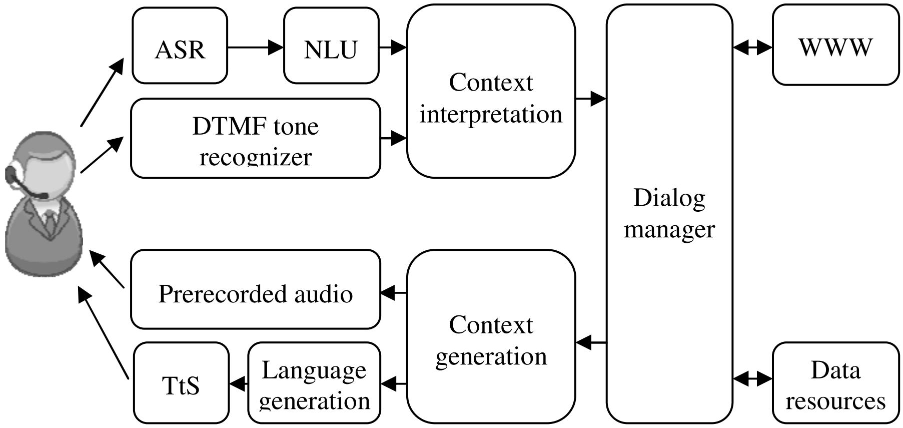

**Fig. 1.** Spoken dialog interface framework

Broadly speaking, a generic dialogue system comprises of three modules:

- Input – commonly includes automatic speech recognition (ASR) and natural language understanding (NLU). The ASR converts the acoustic user input into text while the NLU parses the text in order to semantically interpret it. Additionally, a DTMF tone recognizer may be included in order to allow for such input.
- Dialogue Management – is the core of the dialogue system. It handles a unique and complete conversation with the user, evaluating the input and creating the output. In order to do that, it activates and coordinates a series of processes that evaluate the user prompt. The dialog manager (DM) identifies the communicative act, interprets and disambiguates the NLU output and creates a specific dialogue strategy in order to respond. It maintains the state of the dialog (or belief state), formulates a dialog plan and employs the necessary dialog actions in order to fulfil the plan. The DM is also connected to all external resources, back-end database and world knowledge.
- Output – usually in includes a natural language generator (NLG) coupled with a text-to-speech synthesizer (TtS). The NLG renders the dialog manager output from communicative acts to proper written language while the TtS engine converts the text to speech and/or audio. A lot of applications, for the sake of customer satisfaction, use prerecorded audio queues instead of synthetic speech for output. In that case, the dialogue manager forms the output by registering all text prompts and correlating them with prerecorded audio files.

When building a speech-based human-computer interaction system, certain basic modules must be present [23]. The Dialogue Manager is responsible for the system behavior, control and strategy. In general, a dialogue with a machine is a sequential process and contains multiple turns that can be initiated by the machine (system initiative), the user (user initiative), or both (mixed initiative). The ASR and NLU recognize the spoken input and identify semantic values. The language generator and TtS or the prerecorded audio generator provides the system response. The dialogue is usually restricted within the thematic domain of the particular application. The performance of the particular modules is an indication of usability issues. The ASR accuracy and the lack of language understanding due to out-of-grammar utterances or ambiguity hinder the spoken dialogue. Moreover, the lack of pragmatic competence of the dialogue manager (compared to the human brain) and the response generation modules sometimes overcomplicate the dialogue and frustrate the user.

## 3 Usability

The term usability has been used for many years to denote that an application or interface is *user friendly*, *easy-to-use*. These general terms apply to most interfaces, including web interfaces and more importantly speech-based web interfaces. Usability is measured according to the attributes that describe it, as explained below [25]:

- Usefulness – measures the level of *task enablement* of the application. As a side results it determines the *will* of the user to actually use it for the purpose it was designed for.

- Efficiency – assesses the *speed*, *accuracy* and *completeness* of the tasks or a user's goal. This is particularly useful for evaluating an interface sub-system since the tasks may be broken down in order to evaluate each module separately.
- Effectiveness – quantifies the system *behaviour*. It is a user-centric measure that calculates whether the system behaves the way the users expect it to. It also rates the system according to the level of *effort* required by the user to achieve certain goals and respective *difficulty*.
- Learnability – it extends the effectiveness of the system or application by evaluating the user's effort required to do specific tasks over several repetitions or time for training and expertise. It is a key measure of user experience since most users expect to be able to use an interface effortlessly after a period of use.
- Satisfaction – it is a subjective set of parameters that the user's are asked to estimate and rank. It summises the user overall *opinion* about an application based on whether the product meets their *needs* and performs *adequately*.
- Accessibility – is a very important and broad discipline with many design and implementation parameters. For spoken dialogue interfaces, it can be through of as an extension of the aforementioned usability attributes to the universal user. Speech and audio interfaces are suitable for improved accessibility [11, 5, 12].

### 3.1 Interaction Design Lifecycle (Interfaces) and Usability

The basic interaction design process is epitomized by the main activities that are followed for almost every product. There are four stages in the lifecycle of an interface:

- Requirements specification and initial planning
- Design
- Implementation, testing and deployment
- Evaluation

In terms of usability there are three key characteristics pertaining to user involvement in the interaction design process [26]:

- User involvement should take place throughout all four stages.
- The usability requirements, goals and evaluation parameters should be set at the start of the development
- Iteration through the four stages is inevitable and, therefore, should be included in the initial planning.

Figure 2 shows how usability generally integrates with the development of an interface.

### 3.2 Speech Interfaces and Usability

Spoken dialogue interfaces may be of three types depending on their design:

a. DTMF replacement
b. Simple system or user-directed question-answering
c. Open-ended natural language mixed-initiative conversational system

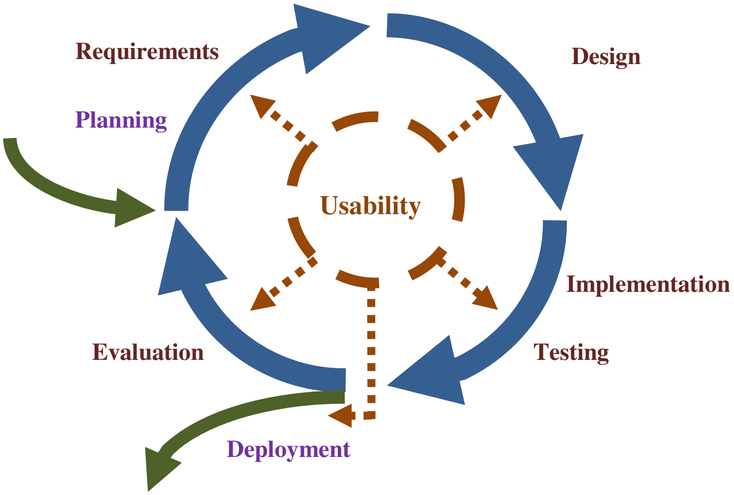

**Fig. 2.** Typical interface lifecycle and usability

Type (a) systems are the very basic menu-driven interfaces where a static tree-based layout is presented to the user. The user may respond with yes/no and navigate through the menu through options. Such systems are not user-friendly, typically used for very limited domain services, and require patience and time from the user in order to complete a task. The main advantage is that they are very robust, since the user is presented with only a few options at any time, and can only go forward or backwards in the tree-structured menu.

Type (b) systems use more advanced techniques in order to accommodate a more natural interaction with the user. The menus may be dynamic, have confirmation and disambiguation prompts as well as more elaborate vocabulary. Still, the system or the users have to use voice responses within the grammar. Such systems have reusable dialogue scripts for dialogue repair. The small grammars keep the system relatively robust. Such systems are used for most applications at the moment, providing a trade off between efficiency and robustness.

Type (c) systems are used for large scale applications. These systems are targeted for user satisfaction and naturalness. The users may respond to natural “how may I help you” system prompts with equally natural replies. The utterances may be long, complex and exhibit great variety. The dialogue is dynamic and the demand for successful ASR is high, as is the use of statistical or machine learning methods for interpretation. The dialogue management is task-based, the system creating tasks and plans of actions to fulfil. The users expect high-level natural interaction, a very important element to factorise in usability parameterisation.

It is obvious now that each type of design entails particular usability expectations. Each type is expected to excel in certain aspects.

**Table 1.** Usability impact on spoken dialogue interface development lifecycle

| Type | Requirements | Design | Implementation | Evaluation | Deployment |
|---|---|---|---|---|---|
| **DTMF** | low | medium | low | low | low |
| **Q&A** | medium | medium | low | medium | low |
| **Open** | high | high | medium | high | medium |

Table 1 shows how usability is taken into account in each stage of the product lifecycle, based on the experience from the development and testing of nationwide-size spoken dialogue business applications. The development of such applications is an iterative process, as mentioned before. Practitioners in industrial settings agree that usability parameters as well as testing is also part of the iterative process. Open systems possess the highest potential for usability integration. In that respect, the remainder of this chapter refers mostly to open systems and less to the other two types. These days, such systems are the centre of the attention by researchers, developers and customers alike, focusing on advanced voice interaction and high user satisfaction. The use of natural voice response (both acoustic and syntactic) and the natural dialogue flow constitute the state-of-the-art in spoken dialogue interfaces.

## 4 Usability Evaluation

Usability evaluation is usually performed either during or at the end (or near the end) of the development cycle. The methodologies that can be used for that differ in their scope, their main difference being that, when a product is finished (or nearly finished), *usability testing* serves for fine-tuning certain parameters and adjusting others to fit the target user better. During the design phase, usability evaluation methods can be used to probe the basic design choices, the general scope and respective task analysis of a web interface. Some of the most common factors to think about when designing a usability study are:

- Simulate environment conditions closely similar to real world application use.
- Make sure the usability evaluation participants belong to the target user group
- Make sure the user testers test all parameters you want to measure
- Consider onsite or remote evaluation

### 4.1 Methodologies

Usability evaluation for speech-based web interfaces is carried upon certain usability evaluation methods and approaches on the specific modules and processes that comprise each application. Each approach measures different parameters and goals. They all have the same goal, to evaluate usability for a system, sub-system or module. However, each approach targets specific parameters for evaluation. Depending on whether they are deployed during the design or production phase, and whether they focus on the user interaction or a sub-system performance parameter, there are two distinct classes of methodology – evaluative usability testing and Wizard-of-Oz (WOZ) testing [14].

### 4.1.1 Wizard-of-Oz (Formative Evaluation)

It is a common approach that is used not only for speech-based dialogue systems but for most web applications. It enables usability testing during the early stages by using a human to simulate a fully working system. In the case of speech-based dialog systems, the human "wizard" performs the speech recognition, natural language understanding, dialog managements and context generation. Cohen et al. [4] list the main advantages of the WOZ approach:

- Early testing – it can be performed in the early stages in order to test and formulate the design parameters as early in the product lifecycle as possible.
- Use of prototype or early design – eliminates problems arising later in the development such as integration.
- Language resources - Grammar coverage for the speech recognition (ASR) and respective machine learning approaches for interpretation (NLU) are always low when testing a non-finalised product. Low scoring for ASR-NLU may hinder the usability evaluation, however, the use of the human usability expert eliminates such handicap.
- System updates – the system, being a mock-up, can be updated effortlessly to accommodate for changes imposed from the input from the test subjects, making it easier to re-test the updated system in the next usability evaluation session.

The WOZ approach is primarily used during the initial design phase to test the proposed dialogue flow design and the user response to information presentation parameterisation. Since errors from speech recognition and language interpretation are not taken into account, the resulting evaluation lacks the realistic aspect. Expert developers usually know what to expect from the speech recognition and interpretation accuracy because these are domain dependent.

There are two requirements for successful usability testing, the design of the tasks and the selection of participants. The participants are required to complete a number of tasks that are carefully selected to test the system. In a dialogue system the primary concern to evaluate is the dialogue flow. Other tasks consist of testing the natural language interface of such aspects as linguistic clarity, simplicity, predictability, accuracy, suitable tempo, consistency, precision, forgiveness, and responsiveness, which make the interface easy and transparent to use.

### 4.1.2 Usability Testing Using Working Systems (Summative Evaluation)

As mentioned before, usability testing can take place during the design phase, the implementation or after the deployment of a speech-based dialogue interface. At the end of the implementation, pre-final versions of the system should be tested by potential users in order to evaluate the usability. For spoken dialogue interfaces, a set of 15 objective (quantitative or qualitative) and subjective usability evaluation criteria have been proposed [7], including modality appropriateness, input recognition adequacy, naturalness, output voice quality, output phrasing adequacy, feedback adequacy, adequacy of dialogue initiative relative to the task(s), naturalness of the dialogue structure relative to the task(s), sufficiency of task and domain coverage, sufficiency of the system's reasoning capabilities, sufficiency of interaction guidance, error handling adequacy, sufficiency of adaptation to user differences, number of interaction problems [1] and overall user satisfaction.

Moreover, Bernsen & Dybkjær [2] have used evaluation templates for their DISC evaluation model as best practice guides while, later, they formed a set of guidelines for up-to-date spoken dialogue design, implementation and testing, covering seven major aspects [3]. These aspects can be used as the basis for usability evaluation strategies. Many frameworks and methodologies have been developed and used for evaluation of spoken dialogue systems in recent works [8, 9, 13, 15, 17, 18, 21, 24, 30].

## 5 Usability Evaluation of a Spoken Dialogue System: Case Study

There are two requirements for successful usability testing, the design of the tasks and the selection of participants. The participants are required to complete a number of tasks that are carefully selected to test the system. In a dialogue system the primary concern to evaluate is the dialogue flow.

A comparison between a DTMF system and a spoken language interface is expected to reveal no major differences accessibility wise. Both modalities are for example equally appropriate for a blind person trying to access a service by phone. Significant differences are however expected with respect to usability issues. In particular, DTMF systems are exclusively menu driven using a strictly hierarchical and static navigation process. This essentially means that the customer needs to navigate through various levels of menus listening to every option to finally be transferred to a human agent and be asked the general “how may I help you” question. This is especially true for complex domains taking into account that the options list should be kept short for cognitive reasons. Natural Language Interfaces, on the other hand, allow for a more dynamic and shorter dialog flow. Hence they allow for greater efficiency, flexibility and naturalness.

Furthermore, it is often the case that there is no exact mapping between the user’s request and the menu options presented to him and hence the user is left confused with no option that suits his needs (which in turn leads to hang-ups or mis-interpretations). On one hand, the large number of possible options is prohibitive resulting in a dysfunctional, over-complicated menu structure, on the other hand DTMF system design is most often developer rather that user centered. In contrast spoken dialogue interfaces utilize machine learning techniques in order to model user behaviour. The system is thus better adapted to the user’s mental model allowing for a more natural interaction.

With respect to learnability considerations, the intrinsically arbitrary nature of the mapping between concepts and DTMF tones makes options difficult to remember, especially in the case of complex menu trees and users that rarely call the system. Speech, of course, is for most users the most natural way to interact and the mode they are most experienced with.

Finally, DTMF systems are by definition restricted in choice of wording, which can sometimes be confusing when it comes to use of jargon, especially for elderly people who are not very technologically aware. For example, an elder who wants to make use of video calling capability on his mobile phone may not be aware that he needs to choose the 3G services option in the DTMF menu. In the case of a spoken language system, he could – in theory at least – express his request in his own words.

On the other hand, recognition technology in the case of spoken language systems is still error prone, and it has been shown that speech recognition accuracy greatly affects the user's experience. Furthermore, spoken language systems – especially mixed-initiative ones – are more difficult, time-consuming and costly to develop, ultimately posing the question whether there should be a trade-off between user satisfaction and engineering considerations. It should also be noted that the general population is not very familiar and experienced with such systems. Performance usually goes down first, before it goes up again, once the users are trained. DTMF systems are in contrast more familiar to users, easier to develop, simpler and thus almost error free. In addition, DTMF systems are constantly directing the user through the interaction, so most of the times the user knows what to do.

All issues mentioned above affect key usability criteria such as efficiency, effectiveness, user-satisfaction and ease of learning. Since these usability criteria apply to both types of systems, the two modalities can be compared based on metrics defined for each criterion, irrespectively whether the exact metric per se can apply to both types of systems. Such metrics can be hang-ups, no-input rates, no-match events (or pressing a non available DTMF key for DTMF systems respectively), interaction duration, number of dialog steps/turns (whereas the user input may be an utterance or a sequence of DTMF keys), successful task completion, use of barge-in capability, subjective evaluation questionnaires.

In the following section we will present a case study whereas a DTMF system of moderate complexity is compared to a spoken dialog interface. Both systems serve the same services for a Customer Care call centre of a Mobile Telephony company. Thus, they give users distant access to services by phone. We will first present the two systems in light of the usability criteria analyzed throughout this paper, and then we will present the usability evaluation of each system.

## 5.1 System Description

Figure 4 shows the basic architecture of both systems. The DTMF system has a three level menu with a number of options at each level ranging from 3 to 4. As is, the user needs to go through three steps to choose the deepest embedded in the tree option. The same can be accomplished within a single dialog turn in the case of the Spoken Dialogue system leading to more efficient interaction. Also worth noting that while both systems have barge-in functionality, DTMF users will still need to listen to all options if what they want is presented last. Furthermore, in the DTMF system the user is presented with 9 options – corresponding to user requests – while the Spoken Dialogue system at this point handles approximately 30 high-level request categories (that basically means that 21 categories are subsumed under the "other" option in the DTMF menu). Categorization of requests was based on the analysis of the domain and callers' utterances. As a result the speech based system is bound to achieve greater coverage and be more useful and effective.

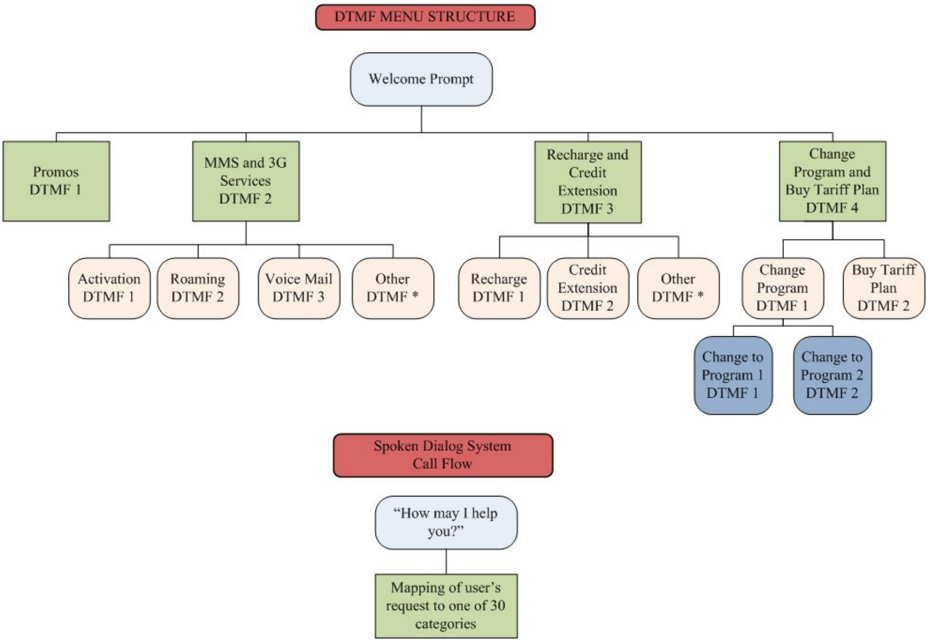
DTMF MENU STRUCTURE
Welcome Prompt
Promos DTMF 1
MMS and 3G Services DTMF 2
Recharge and Credit Extension DTMF 3
Change Program and Buy Tariff Plan DTMF 4
Activation DTMF 1
Roaming DTMF 2
Voice Mail DTMF 3
Other DTMF *
Recharge DTMF 1
Credit Extension DTMF 2
Other DTMF *
Change Program DTMF 1
Buy Tariff Plan DTMF 2
Change to Program 1 DTMF 1
Change to Program 2 DTMF 2
Spoken Dialog System Call Flow
"How may I help you?"
Mapping of user's request to one of 30 categories

**Fig. 3.** DTMF and Speech-based dialogue systems basic architecture. Optimal tree traversal is faster for the former. Error handling, universals, help sub-dialogs in the Spoken Dialog system and some sub-menus in the DTMF system are excluded for ease of presentation.

### 5.2 Usability Evaluation: Procedure, Results and Discussion

During usability testing seven participants were asked to perform five tasks each, using the DTMF and the spoken language functionality. None of the participants had used the particular systems before, all but one had used DTMF systems before, and only two had used natural language systems before. Choice of tasks was based on stats from real usage, so that the most common tasks were chosen. Also tasks were chosen so that they correspond to options found high and leftward in the tree structure, as well as more deeply embedded. Finally, one task corresponded to a request not represented in the DTMF menu. For each task, successful completion, call duration and dialogue steps, number of barge-ins and no-inputs were monitored. In addition, at the end of the testing session, each participant was asked to fill in a relevant questionnaire. Table 2 shows example questions (based on [6, 16, 22, 29], among others) used as subjective evaluation criteria and the usability aspect they correspond to. Participants were asked to assess the two systems on a Likert scale from 1 to 5. Finally, as a control parameter, participants were asked to compare the systems (e.g "Which system is more useful?") and state which system/modality they preferred.

Figures 5 and 6 present the results for the objective criteria. While task completion ratio is almost the same for the two systems, calls to the speech based system required fewer steps and were almost two times shorter compared to calls to the DTMF system. Barge-in functionality was most often used during the interaction with the DTMF system. This was expected, since users are more familiar with DTMF systems

**Table 2.** Example questions from the usability questionnaire. Each question is mapped to the usability parameter it assesses.

| Example Question / Statement | Usability Parameter |
|---|---|
| The system is useful | Usefulness |
| The interaction was boring | User satisfaction-Engagement |
| The interaction was frustrating | User satisfaction-Engagement |
| The interaction with the system is fast and efficient | Efficiency |
| There were many repetitions in the interaction | Efficiency |
| The interaction with the system is natural | Naturalness |
| I always knew what to do/say | User expertise-Expected behaviour |
| High concentration was required while using the system | Ease of use-Task ease |
| I could recover from errors quickly | Error handling / error tolerance |
| I felt I had control during my interaction with the system | Interaction control-dialog initiative |

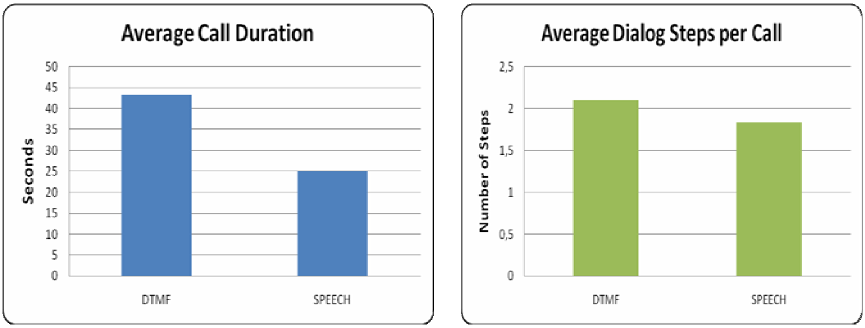

**Fig. 4.** Average call duration in seconds. Calls made to the Spoken Dialog System were significantly shorter. Number of Dialog Steps/Turns. A caller's utterance corresponds to a sequence of one or more DTMF tones. As within an utterance the user may provide more than one piece of information, similarly in a DTMF system an expert user may press more than one keys within a single turn.

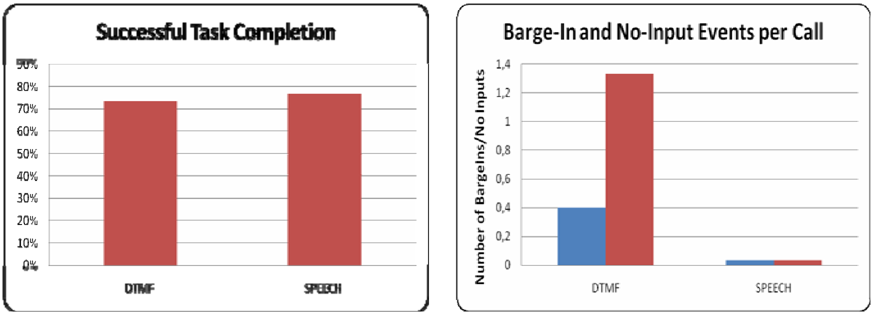

**Fig. 5.** Percentage of tasks completed successfully and average number of barge-ins and no-input events per call

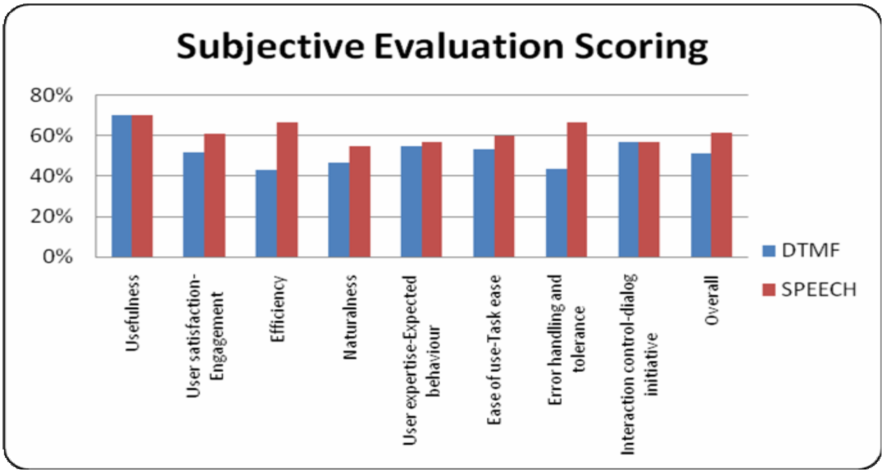

**Fig. 6.** Subjective Evaluation Scoring. Scores are presented as percentages of the highest score possible.

and DTMF prompts are inherently longer, as all options are listed there. Furthermore, natural language interaction has a different turn taking protocol. Number of no input events was also higher for the DTMF system as users sometimes felt that they had no menu item matching their request.

Figure 7 presents the results of the subjective evaluation. In general, the speech based system ranked higher and was preferred by all users but two. These users had experience with DTMF systems only, and felt more comfortable using such a system than talking to a machine. The speech based system scored significantly better with respect to efficiency, error handling and overall. It also scored better with respect to user satisfaction, naturalness and task ease, and only marginally better with respect to expected behavior and user expertise. The latter can be accounted for, considering that most participants had greater experience with DTMF. As far as error handling and dialog repair are concerned, natural dialog systems are more flexible and robust, whilst DTMF systems usually just force the user to start over again from the root of the menu. Finally, both systems were considered equally useful.

## 6 Conclusions

A usability evaluation paradigm was described that compared two systems of different modalities performing the same task. Despite recognition errors (25% natural language interpretation error rate) the speech based system proved to be faster, more efficient and user friendlier. And while the increase in speed of task completion is expected to be greater for the DTMF system, as users become more familiar with the application (adapting to menu structures, memorizing choices, using the barge-in functionality), it has been shown that expert, familiarized users are not necessarily satisfied users. Similarly, equally useful, effective (note that task completion rates were similar) and accessible interfaces, can differ vastly with regards to the end-user experience and satisfaction. Especially in the case of disabled users, who often have limited options for use of an alternative modality or a multi-modal interface, it is thus critical that the system is not only accessible but as usable as possible. Appropriate modalities are not necessarily usable modalities.

Finally, the following should be noted: Firstly, it has been shown in the literature [27] that there are differences between usability testing results and real use results. Our data from the first days of the speech-based system's launch corroborate this claim. In particular, there is a high ratio of no-input events and hang-ups (18,8% and 13,5% respectively). We do expect, however, that this percentage will decrease as users become more experienced with the interface. Lack of familiarity on behalf of the end user is a drawback for current open-ended speech interfaces.

Secondly, as mentioned above, speech interfaces can be costly and cumbersome to develop. Nevertheless, they can prove easier to manage and cheaper to upgrade. For example, due to the in-depth analysis of caller requests, future services may be broken down or reassigned to other sub- or top-level categories and new service queues effortlessly. New services can be added without any need to change the structure of the spoken dialogue application, as no menus or list options are actually presented to the user. Thus, since there is really no need to redesign, smooth transition is assured during major updates and the user experience is retained. The latter is considered to be a key parameter for user satisfaction. However, there are still issues to be considered with respect to system's need for re-training and change in content.

Finally, in this case study, a speech interface is compared to a DTMF system of moderate complexity. It might as well be the case that when compared to a simple DTMF system, there is no gain in user satisfaction. Our experience from such a simple call services application indicates that there is no difference in subjective evaluation. As an objective criterion, however, a 1.5% decrease in inner transfers (resulting from errors in routing) was monitored from the moment the speech-based system was launched. Similarly, we would expect a significant gain in user satisfaction for complex domains. In any case, early usability testing can facilitate the choice of the appropriate for each application modality, but more importantly help create an interface that it both functional as well as familiarly usable since the user feedback can be used right from the design phase.

## Acknowledgements

The work described in this paper has been funded by the KAPODISTRIAS Programme of the Special Account for Research Grants, University of Athens.

## References

1. Bernsen, N.O., Dybkjaer, H., Dybkjaer, L.: Designing Interactive Speech Systems: From First Ideas to User Testing. Springer, London (1998)
2. Bernsen, N.O., Dybkjær, L.: A Methodology for Evaluating Spoken Language Dialogue Systems and Their Components. In: International Conference on Language Resources and Evaluation, pp. 183–188. ERLA, Athens (2000)

3. Bernsen, N.O., Dybkjær, L.: Building Usable Spoken Dialogue Systems. Some Approaches. Int. J. Lang. Data Proc. 28(2), 111–131 (2004)
4. Cohen, M., Giancola, J.P., Balogh, J.: Voice User Interface Design. Addison-Wesley, Boston (2004)
5. Duarte, C., Carriço, L.: Audio Interfaces for Improved Accessibility. In: Pinder, S. (ed.) Advances in Human Computer Interaction, pp. 121–142. I-Tech Education and Publishing KG, Vienna (2008)
6. Dutton, R.T., Foster, J.C., Jack, M.A., Stentiford, F.W.M.: Identifying usability attributes of automated telephone services. In: European Conference on Speech Communication and Technology, pp. 1335–1338. ISCA, Berlin (1993)
7. Dybkjær, L., Bernsen, N.O.: Usability Issues in Spoken Language Dialogue Systems. Nat. Lang. Eng. 6(3-4), 243–272 (2000)
8. Dybkjær, L., Bernsen, N.O.: Usability Evaluation in Spoken Language Dialogue Systems. In: ACL Workshop on Evaluation Methodologies for Language and Dialogue Systems, pp. 9–18 (2001)
9. Dybkjær, L., Bernsen, N.O., Minker, W.: Evaluation and Usability of Multimodal Spoken Language Dialogue Systems. Speech Communication 43(1-2), 33–54 (2004)
10. eEurope 2005: An Information Society for All. Online Project Web Site (2005), http://europa.eu.int/information_society/eeurope/2005/index_en.htm
11. Fellbaum, K., Kouroupetroglou, G.: Principles of Electronic Speech Processing with Applications for People with Disabilities. Technology and Disability 20(2), 55–85 (2008)
12. Freitas, D., Kouroupetroglou, G.: Speech Technologies for Blind and Low Vision Persons. Technology and Disability 20(2), 135–156 (2008)
13. Hajdinjak, M., Mihelic, F.: The PARADISE evaluation framework: Issues and findings. Comp. Ling. 32(2), 263–272 (2006)
14. Harris, R.A.: Voice Interaction Design: Crafting the New Conversational Speech Systems. Elsevier, Amsterdam (2005)
15. Hartikainen, M., Salonen, E.-P., Turunen, M.: Subjective Evaluation of Spoken Dialogue Systems Using SERVQUAL Method. In: International Conference on Spoken Language Processing, pp. 2273–2276. ISCA, Jeju (2004)
16. Kamm, C.A., Litman, D., Walker, M.A.: From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems. In: International Conference on Spoken Language Processing, pp. 1211–1214. ISCA, Sydney (1998)
17. Kamm, C.A., Walker, M.A., Litman, D.: Evaluating spoken language systems. In: American Voice Input/Output Society Conference, AVIOS, San Jose, pp. 187–197 (1999)
18. Larsen, L.B.: Issues in the Evaluation of Spoken Dialogue Systems using Objective and Subjective Measures. In: 8th IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 209–214. IEEE Press, New York (2003)
19. Larson, J.A., Raman, T.V., Raggett, D.: W3C Multimodal Interaction Framework, http://www.w3.org/TR/mmi-framework/
20. Larson, J.A.: W3C Speech Interface Framework, http://www.w3.org/TR/voice-intro/
21. Litman, D.J., Pan, S.: Designing and evaluating an adaptive spoken dialogue system. User Modeling and User-Adapted Interaction 12(2-3), 111–137 (2002)
22. Love, S., Dutton, R.T., Foster, J.C., Jack, M.A., Stentiford, F.W.M.: Identifying salient usability attributes for automated telephone services. In: International Conference on Spoken Language Processing, pp. 1307–1310 (1994)
23. McTear, M.F.: Towards the Conversational User Interface. Springer, London (2004)

24. Moller, S., Englert, R., Engelbrecht, K., Hafner, V., Jameson, A., Oulasvirta, A., Raake, A., Reithinger, N.: MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations. In: 9th International Conference on Spoken Language Processing, pp. 1786–1789 (2006)
25. Rubin, J., Chisnell, D.: Handbook of Usability Testing, Second Edition: How to Plan, Design, and Conduct Effective Tests. Wiley Publishing, Inc., Indianapolis (2008)
26. Sharp, H., Rogers, Y., Preece, J.: Interaction Design: Beyond Human-Computer Interaction. John Wiley & Sons, Inc., New York (2002)
27. Turunen, M., Hakulinen, J., Kainulainen, A.: Evaluation of a Spoken Dialogue System with Usability Tests and Long-term Pilot Studies: Similarities and Differences. In: 9th International Conference on Spoken Language Processing, pp. 1057–1060 (2006)
28. van Kuppevelt, J., Dybkjær, L., Bernsen, N.O. (eds.): Advances in natural multimodal dialogue. Springer, The Netherlands (2005)
29. Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A.: Evaluating spoken dialogue agents with PARADISE: Two case studies. Comp. Speech Lang. 12(3), 317–347 (1998)
30. Walker, M.A., Kamm, C.A., Litman, D.J.: Towards developing general models of usability with PARADISE. Nat. Lang. Eng. 6(3-4), 363–377 (2000)