VOICE USER INTERFACE USING VOICEXML
Environment, Architecture and Dialogs Initiative

Alexandre M. A. Maciel and Edson C. B. Carvalho
Center of Informatics, Federal University of Pernambuco, Recife, Brazil

Keywords: Voice User Interface, Voice Technologies, Dialog Initiative and VoiceXML.

Abstract: In this work we present a set of applications for Internet with voice user interface using VoiceXML language. Architecture, main platforms and dialog initiative ways were studied. Applicability and limitations were determined.

1 INTRODUCTION

According to (Lévy, 1993) a man-machine interface assigns to a set of programs and material devices that allows the communication between an information system and its human users. So, while humans and machines were not able to speak the same language, the interfaces will be necessary to mediate the communication between them.

The technologies associated with the construction of interfaces affect and guide our perception and the way we interact with the computational systems. It changes the way as we create and communicate with them (Johnson, 2001). If the metaphor of computer interface was another one, probably we would think different.

At the moment, surrounded by technology and machines anywhere, and, consequently of interfaces, applications since Internet maintains the same metaphor of the traditional computational systems, which limits our interaction. Voice interfaces offer, besides an easy and extremely intuitive communication form, other advantages in the machine interaction as, for example, the speed in the input data, safety in the speaker identification and accessibility of handicapped people.

(Gabriel, 2005) affirms that in almost all the technological areas, the boundaries between medias, technologies and concepts, are suffering a dissolution and hybrid process, such as all analysis, classification of processes and interactions become much more complex than before. In that way, when joining capacity and user-friendliness of voice interfaces to dialogue, with capacity and web diversity brings about a mean of communication with an enormous potential.

2 VOICE USER INTERFACE

Construction of voice interface applications is a challenge and the reason for this is that the language is deeply related to human behaviour (Schinelle, 2005). As a consequence, the expectations related to the interface become very high. This kind of interface tries to lead the user to the sensation that he could speak as if it he was talking with a human, however it was not perfectly achieved.

The main objective of a voice user interface project is to support the user navigation with options, commands and available information in a system to carry out a specifc task. Unfortunately, access information through navigation is more complex in the audio ambit.

A good voice interface projection can successfully attenuate the effect of these deficiencies using a user interaction structure that tries to carry out the required tasks successfully. For this, some factors must be considered in the voice interface design: the application requirements, potentialities and limitations of the technology and the population characteristics (Kamm, 1995).

Once understood those factors, the voice interface designer can anticipate some difficulties and incompatibilities that will affect the success of the application, minimizing its impacts.

The project and the execution of an interface are most successful as an interactive process with the interfaces tested empirically in groups of representative users where the problems are detected, corrected, and re-examined until the system achieves a steady and satisfactory performance.

M. A. Maciel A. and C. B. Carvalho E. (2007).
VOICE USER INTERFACE USING VOICEXML - Environment, Architecture and Dialogs Initiative.
In *Proceedings of the Second International Conference on Signal Processing and Multimedia Applications*, pages 370-374
DOI: 10.5220/0002138103700374

2.1 Interaction

Voice interfaces supply the information systems with an interesting alternative for input and output data such as a voice-only interface (phone) or a component of a multimodal and/or multimedia system.

A voice-only interface in an information system can become desirable for two reasons. First, the application can require free hands in the interaction. Second, the telephone system is a net technology truly robust and universal. Then, it makes sense to extend the information services from computer to phone (Dey, 1997).

Multimodal interfaces are a human-machine interaction for sequential or parallel applications of input/output data. Speech recognition, keyboard, mouse, mimic, gestures can be used as modality of input data and to get a synthesized reply voice, graphics or text message. These ways of interaction can be combined dynamically to provide bigger mobility to the user (Englert, 2006).

2.2 Dialogue Initiative

One of the fundamental aspects of the development of applications with voice interface is the way the dialogue initiative is taken. The strategy of management dialogue can be by system, user or mixed initiative (SPI Group, 2006).

In a system-initiative dialogue, the computer asks the user and when the necessary information is received, the solution is processed and the answer is given. Dialogues with user-initiative assume that the user knows what to do and how interact with the system. Generally, the system waits for the user input and answers it through operations. Applications with mixed-initiative assume that the initiative of the dialogue can be taken by the system or the user.

3 VOICEXML

VoiceXML is a markup language and its main objective is to bring the powerful Web development and to give the content for applications with voice interface. It allows the voice services integration with data services giving access to information and services in phone devices like the traditional Web (VoiceXML Forum, 2000).

The advantage of using VoiceXML language to construct voice services is that companies can create voice automatized applications using a similar technology used to create Web visual sites, reducing significantly the construction cost of corporative voice sites (Kondratova, 2004).

In VoiceXML, an application is composed of a set of linked documents, all of them making reference to a main document called *root* of the application. All the applications begin from this document when it is loaded onto the server. See Figure1.

The content of a VoiceXML document is normally divided in a series of dialogues and sub-dialogues. The *dialogs* contains information for a particular transaction processing, for example, supply information considered in the next dialogue after the complement of current transaction. The *sub-dialogs* is generally treated as functions that are used for specific tasks, such as processing, and is called a dialogue "father" and returns it after completing the requested task.

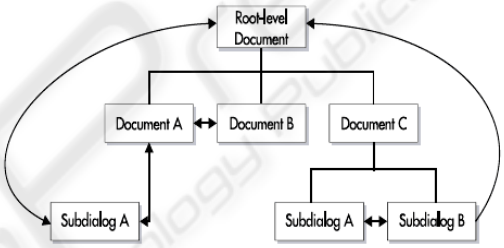

Figure 1: VoiceXML document flow.

Speech recognition systems enbale the computers "to listen" the user's speech and recognize what was said. Voice synthesis systems allow the "reading" of information. However to obtain satisfactory performance and time-out, the current systems limit what the user can speak inside of a context through grammars.

A grammar is a language definition. It can be used to describe natural languages, spoken and written by people, and formal languages such as programming languages, markup language documents, mathematical language and many others (Bringrt, 2005).

Grammars are based on a set of words and sentences that define the possible ways of interaction that can be used in an application. For example, there are many ways of asking for some product. "I´d like", "Give me", "I want". The grammatical rules can also specify kinds of user pronunciation, depending on regional accent (Enden, 1998).

The main standard grammars are Java Speech Grammar Format (JSGF), independent of platform and speaker based on Java technology, and Nuance Grammar Specification Language (NGSL) used in Nuance systems.

3.1 Architecture

In the VoiceXML applications, in the same way as in the Web applications, documents are stored on a Web server. In addition to this server, the VoiceXML architecture requires another server, the voice server, which deals with all interaction between the user and the Web server. The voice server works as a browser in the voice applications, interpreting all users input data and promoting audible messages as reply. In the case of the voice applications, the final user does not need to have a last generation computer with any sophisticating browser. It can access the voice application by a fixed/mobile phone or by VOIP (Voice Over Internet Protocol) software (Shukla, 2005). The cited architecture is shown in Figure 2.

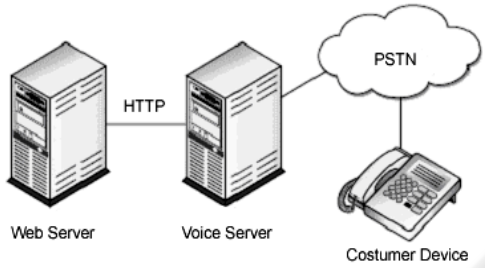

Figure 2: VoiceXML language architecture.

The voice server is an operational platform that executes the VoiceXML language services. Also called gateway, it works making a bridge between the world of telephony and Internet and assures that the development of the voice applications is successful in operation and maintenance.

A communication between those servers is made by means of HTTP (Hypertext Transfer Protocol), protocol for the Internet or Intranet. Communication between the voice server and the device customer is carried through the net PSTN (Public Switched Telephone Network).

To make this communication in an efficient way, the voice server has diverse technological resources (Beasley, 2001). The most important is the interpreter who supplies advanced characteristics similar to visual browsers (cache, favourites) and analyzes the source code that uses speech recognition and synthesis resources.

3.2 Platforms

In market exist many options of voice-based applications platforms. The main informatics companies possess production and publication tools each one with special characteristics.

Table 1: Main Voice Platforms.

	IBM	Nuance	Microsoft
Platform	WVS	NVP	MSS
Version	5.1	3.0	2004
Idiom	English, Portuguese	English, Portuguese	English
Source	Open	Open	Close
Hardware Required	Proc. 1.0Ghz Mem 2GB, Disc 3GB	Proc. 1.0Ghz Mem 1GB, Disc 3GB	Proc 2.5Ghz Mem. 4GB, Disc 20GB
Software Required	AIX, Linux, WIN2003	Solaris,Linux WIN2K	WIN 2003
Languages	VXML 1.0, 2.0	VXML 2.0	XML,SSML, SRGS,SALT

The Websphere Voice Server (WVS), one of the most known platforms, supports diverse standards of voice interface, which gives more freedom in relation to propriety technologies. The WVS has a refined speech recognition and voice synthesis resources due to the investments of IBM and offers a great amount of idioms (IBM, 2005).

The Nuance Voice Platform (NVP) is a VoiceXML open source platform optimized for the development, debugging and monitoring of voice solutions. An important characteristic of NVP is the distributed architecture, easily scaled and managed, specifically created to supply robustness and flexibility to massive applications (Nuance, 2007).

The Microsoft Speech Server (MSS) combines Web technology, voice processing services and telephony facilities in one integrated system, qualifying the companies to unify its infrastructure. It was not used in this work because it has not VoiceXML language support yet (Microsoft, 2007).

4 APPLICATIONS

Three applications have been developed with the objective to test the main characteristics of the VoiceXML language, its environment , architecture and dialogue initiatives. The infrastructure used in the applications is shown in Table 2.

Table 2: Development infrastructure

Interaction	Voice-Only
Initiative	System, User e Mixed
Grammar	JSGF, NGSL
Platform	NVP
Language	VoiceXML 2.0
Idiom	Portuguese
Input Device	Skype Software

4.1 News On-Line

Taking care with the system-initiative specifications, one application with voice interface accessed by phone was developed to offer the regional last news published in the Web sites by means of verbal communication, in flexible and comfortable way. System-initiative strategy reveals that those applications are efficient due to its simplicity in the interaction with the user (Lopéz, 2004).

The application starts the dialogue with a respectful greeting to the user and later it asks using voice commands, which canal the user wants to access. After that, the system verifies if it is an available option. In the positive case, the heading is read and if the user wants to listen to it he only has to answer yes or no; in the negative case, the system repeats the available options until the user chooses a valid one.

```
[System]:Bem vindo ao sistema de notícias.
         Selecione um dos nossos canais:
         Esportes, Economia ou Lazer.
[User]:     Economina
[System]:PIB brasileiro tem queda de 3.4% no
         primeiro trimestre. Você deseja
         ouvir o conteúdo da notícia?
[User]:Não
[System]:Dólar acumula queda de 16% no ano.
         Você ouvir o conteúdo da notícia?
[User]:Sim
[System]:O preço do dólar vem caindo devido
         as compras do Banco Central...
[System]:Obrigado por usar nosso sistema.
```

Figure 3: News application simulation

4.2 Dictionary

Systems with user-initiative tend to be sufficiently complex due to naturalness to recognize voice user´s input, mainly because it treats the recognition of concatenated words. In order to study the man-machine interaction, one portuguese dictionary application was developed achieving the requirements of this strategy. It is an adaptation from the book metaphor (visual) to the phone metaphor (vocal). This kind of applications is common in the industrial area executing control commands (Yankelovich, 1997).

The dictionay starts with a respectful greeting and waits for a user command. It is necessary that the user takes the initiative of the dialogue to get the interaction, or the call phone can be disconnected in a determined time. From the input data, the dictionary can give the translation or meaning of a word by consulting its grammar. If the input given by the user was understood, a desired answer will be listened, otherwise it emits an error message.

```
[System]:Bem vindo ao sistema de dicionário.
         O que deseja?
[User]:     Quero o endereço da UFPE
[System]:Expressão não compreendida. O que
         deseja?
[User]:Quero o significado de computador
[System]:Aparelho eletrônico concebido para
         realização de cálculos e operações
         lógicas...
 [System]:Obrigado por usar nosso sistema.
```

Figure 4: Dictionary application simulation

4.3 Airlines Information

Taking care with the mixed-initiative specifications, one application was developed to offer brazilian airlines information with schedules and fees. Those kinds of applications are common in dialogue systems. Other applications offer trips consult services, bookings and tickets sales (Lopéz, 2000).

The airlines information system starts with a respectful greeting followed by a question. Considering that the user has one brief knowledge of the purpose of the application, it must say something that demonstrates its interest in travel from one city to another. As if there's a mixed-initiative application, a small intervention of the system is necessary to guarantee the information security during the interaction. Therefore, the second stage of the dialogue is a confirmation of what was said by the user. In case of positive reply the available schedule is supplied by fees, otherwise the system will turn back to the initial condition.

```
[System]:Bem vindo ao sistema de informações
         aéreas. O que deseja?
[User]: Gostaria de viajar de Recife a São
        Paulo
[System]:Você deseja viajar partindo de
         Recife até São Paulo, sim ou não?
[User]:Sim
[System]:Foram encontrados 3 vôos saindo de
         Recife até são Paulo.
         Terça-Feira, 18:00h, R$389,00
         Quinta-Feira, 15:30h R$389,00
         Sábado, 0:15h, R$339,00
 [System]:Fim dos horários. Obrigado por
         usar nosso sistema.
```

Figure 5: Airline information application simulation

4.4 Evaluations

Each application was tested by four users of both sex, with different skills, 20 times in different environments (quiet or noisy). No differences of performance in relation to the environment and sex were determined. The experienced users had better performance in 15% than others. The Skype software had some transmission delays that harmed the test. An average of 30% of calls suffered unexpected interruptions.

A quantitative evaluation of the voice synthesis system was not possible because it can not be measured. The result was considered satisfactory by the users. The effectiveness of the recognition system was 100% for news online, 90% for airline information and 86% for dictionary. The results were obtained by a simple average of rigthness and errors in the system. However, two factors made difficult a correct analysis. The error can be in the imperfection of transmission or in the grammars especification.

5 CONCLUSIONS

The speech recognition based on the telephonic net offers an enormous potential because it is extremely spread out. It is also a difficult technique because it is impossible to control use conditions. Problems involve a great and unexpected population, differences in the microphones of the devices, noise and short band. The most succeed systems are those which limit the vocabulary size.

The VoiceXML is an excellent language for voice applications with well defined criteria. However, it is not the perfect tool for all kind of projects. Factors that influence the choice of the VoiceXML are architecture, hardware, operational system, idioms and the platform used in a particular project. Then a correct assessment must be done in order to decide for its use.

Regarding the dialogue initiative. Although the system-initiative has an excellent performance in the speech recognition, it allows a little interaction with the user, this is ideal only for informative systems as exchange, time forecast, etc. Nevertheless, the user-initiative requires natural language and needs a deep user´s knowledge about the application, and it is recomended to corporative applications, like agenda and email, for example. The mixed-initiative is the interaction way which has a largest applicability because its confirmation strategy allows a good quality in the speech recognition and it has got a good interaction with the user. This type of initiative can be used as a substitution for traditional call-centers.

ACKNOWLEDGEMENTS

We thank the Center of Informatics for infrastructure and technical support on server's installation.

REFERENCES

Beasley, Rick. et al. *Voice Application Development with VoiceXML*. Sams Publishing, 2001.

Bringrt, Björn. *Embedded Grammars* . Master Thesis. Göteborg University, Sweden, 2005.

Dey, Anind K., et al. *Developing Voice-Only Applications in Absence of Speech Recognition Technology*. GVU Technical Report, Submitted to DIS '97.

Enden, Jarkko. *Java Speech API*. Technical Report of University of Helsink, 1998.

Englert, Roman, et al. *Architecture of Multimodal Mobile Application*. 20th International Symposium on Human Factors in Telecommunication. France, 2006.

Gabriel, M. C. Cruz. *Entre a Máquina e o Homem*. Revista Eletrônica Cibercultura, N:1679-6756, 2005.

IBM Websphere Voice Server for Multiplataforms V5.1.1/5.1.2 Handbook.IBM Coorporation. 2005.

Johnson, Steven. Cultura da Interface: *Como o Computador Transforma Nossa Maneira de Criar e Comunicar*. Rio de Janeiro, 2001.

Kamm, C. *User Interfaces for Voice Applications*. Proceedings of the Natural Academy of Science of the USA. PND:1995;92;10031-10037.

Kondratova, Irina. *Performance and Usability of VoiceXML Application*. 8th World Multi-Conference on Systemics, Cybernetics and Informatics, 2004.

Lévy, Pierry. *As Tecnologias da Inteligência: O Futuro do Pensamento na Era da Informática*. São Paulo, 1993.

Microsoft, Speech Server Web Site. http://www.microsoft.com/sppech, 2007.

Nuance Voice Plataform DataSheets. http://www.nuance.com/voiceplatform/, 2006.

Schnelle, Dirk, et al. *Audio Navigation Patterns.* EuroPLop, 2005.

Shukla, Charul. et al. *VoiceXML 2.0 – Developers Guide*. Dreamtech Software Índia INC, 2000.

SPI - Speech-based & Pervasive Interaction Group. University Tampere. Http://www.cs.uta.fi/hci/spi/ddsi

VoiceXML Fórum. *Voice Extensible Markup Language*. Manual, versão 1.0, 2000.

Yankelovich, Nicole – Using Natural Dialogs as the Basis for Speech Interface Design, Sun Microsystems Laboratories, 1997.