Vocal interaction model in web user interface involving natural language processing

Guillaume KOUM
ENSP Yaounde
g_koum@yahoo.fr

Augustin YEKEL
ENSP Yaounde
a_yekel@yahoo.fr

Serge TAMPOLLA
SOFT-TECH INC.
s.tampolla@soft-techint.com

Tam SANGBONG
SOFT-TECH INC.
tam@soft-techint.com

Abstract

The progress realized in the speech recognition technology throughout the last decade has opened up a new perspective in web user-interface interactions. Now, it is possible to design web portals which enable users to dialog with a page in a more natural and performant way. The actual objective is to gather all these technological resources available in order to design interfaces which users can accomplish tasks in a conversational environment with the machine. However, to implement a dialog, a vocal interface necessitates an enormous amount of work and adaptive strategies if we don't want the interactions to be too restrictive and frustrating for the user. The aim of this article is to go through procedures and techniques of the natural language understanding process as well as present a dialog model for vocal interactions, and a range of technical solutions to implement such a model.

1. Introduction

Interaction between humans and machines represent today a major technological, social and industrial stake. The problem is no more only to increase computers' scalability and to design more and more complex and sophisticated applications, but also to think about offering to the users interfaces that can easily be adapted to their needs in all their computerized communication activities.

This new state of things prompts companies to vary not only access channels (web browser, WAP browser, voice browser) to their applications but also to combine other modalities such as speech, touch (tactile screen) or stylet to traditional input/output modes (keyboard, mouse, graphical interface). Applications that offer this type of interaction are referred to as multimodal applications [1] [8]. The major challenges of this new concept are:

- On one hand, to find the best way to combine all those interaction modes
- On the other hand, to thoroughly study each of them to make them more efficient in this combination.

Speech has and will continuing having a very important role in this new type of interaction. In reality, amongst the communication media, speech is one of the most interesting from a user perspective. The recourse to the speech imposes itself in numerous new applications where the use of the keyboard is difficult, or even impossible: mobile or embarked systems, vocal servers, interactive boundaries, domestic systems, telephony.

Our goal is to explore the means to establish a model of dialogue, by exploiting the knowledge based on the linguistic model as well as those resulting from the survey of the behavior of users and from the progress of the dialogue process to design a robust, flexible and efficient dialogue at the level of a vocal interface.

Thus, our goal is to improve the issues faced with traditional systems such as *DTMF (Dual Tone Multiple Frequency)* and *IVR (Interactive Voice Response)* where the interactions are very restraining and coercive from a user point of view.

This document is organized in four sections. After this introduction, the next section examines the area of human-machine (computer) dialogue and defines some key concepts regarding. The next then section presents our proposed model of dialogue with some quality strategies to implement for an efficient human-machine interaction. Section three will present the realization methodology of a prototype. Finally before the conclusion, the fourth section presents the implementation process of our prototype.

2. Process of natural language understanding

The understanding of speech is a complex process requiring lexical, syntactic, semantic and pragmatic knowledge. Before one can create an interface capable to understand an oral and spontaneous statement, it is indispensable to understand the general working of the interactive system that uses it and to study the various difficulties that one can meet at the time of the understanding [5]. For every linguistic knowledge level, the system must solve some difficulties in order to succeed to an exact understanding of the intention of the use in the context and the environment of the speech. Several formalisms permitting to describe all these knowledge to the system exist.

- *Lexical:* This process consists of understanding the sense of a statement while searching for the sense of one or several words that constitutes it by putting together a collection of other prerecorded words [6].
- *Formal grammar:* permits the description of the syntax of the Language mainly. The non contextual grammars (regular grammars) have a particularly important role as data the processing bus of the effective syntactic. Sensors have been developed for this category of grammars: the automatons to finished states and the battery operated automatons.
- *Stochastic grammar:* improve the understanding of the statement while associating to the linguistic knowledge, the knowledge of the context of the interactions (Hidden Markov Model) [5].

3. Domain overview

3.1. Some key concepts of vocal applications

Before we go any further in this description, we will define some concepts as used in this document.

- *Vocal server*: It is a machine on which software for the treatment of the speech is installed. When talking of a vocal server, we refer either to the software or to the machine on which the software is installed. Generally every software for speech processing will have three main components. Those components are used by vocal applications to analyze users' statements or to synthesize answers to send back to them. They are the following elements:
  - A speech recognition component
  - A component for semantic analysis
  - A component for speech synthesis
- *Vocal browser*: This is the equipment that permits a phone to access a vocal application installed on a *HTTP (HyperText Transfer Protocol)* Web server. Vocal browsers make it possible to build vocal applications for telephones from a simple HTTP server.
- *Vocal portal*: A vocal portal is a concept equivalent to the portal site of the web technology. The content and services found there are similar to those of a web portal. The difference here resides on the fact that information is structured for a vocal medium: simple and short information, different navigation.

3.2. Concept of human-machine dialogue

Literally the notion of "dialogue" underlies a conversational type of operation, an alternated intervention between the human and the machine [2].

It is important to make a difference between dialogue and communication. The dialogue is a particular way to communicate whereas communication can take several means based on languages or not [4]. During a dialogue, participants interact real-time simultaneously. Communication is made if there is mutual understanding of the interacting parties. A machine is not a social being and has neither intention nor culture. Thus, the term human-machine *interaction* is more appropriate than human-machine *communication*. The human-machine dialogue puts in presence, on one hand, a being of natural origin, endowed with a spontaneous and natural capacity to understand what surrounds him, and, on the other, a being of artificial origin of which the understanding capacity is not in anything comparable to that of the former [2]. This type of dialogue as we intend to present brings together an artificial agent and a human agent. In other words, the notion of dialogue is about the establishment of an oral conversation between the human and the machine.

There are different types of human-machine dialogues: objects-oriented dialogues (Here the system does not know any long-term goal. It just has to perform a series of detailed tasks), task-oriented dialogue (in this case tasks should be organized in order to perform the global known goal) and goal-oriented dialogue (The system should have some kind of knowledge to guess what the user really wants.). However, no matter the type of dialogue, a system of human-machine interaction should establish a method for dialogue management. Hence, the system possesses some knowledge, both on the task to perform and on the dialogue; some are static, and others dynamic [3]. Its dialogue manager handles the interaction from the reception of its interlocutor's intervention until the production of its own intervention.

3.3. Styles of dialogues

This describes the different situations of dialogue that the system is capable of tolerating. In practice there exist three types of interactions or styles of dialogue [10]:

- Controlled dialogue: In this style of dialogue, the system controls the interactions. It orientates the user in order to accomplish a very particular task. It asks some questions to the user at every level of the dialogue. The user must answer by words or with very specific sentences. The linguistic domain is very restricted and the vocabulary used is dependent of the domain of the application. If the users' answer is outside of this vocabulary, the system specifies to the user keywords to pronounce.
- Mixed initiative dialogue: The system completely controls the dialogue as in the previous section. However during the dialogue, the user can take initiative and act as the master while anticipating some answers or while orienting the dialogue in another sense. The user can answer a question other than what the system is asking.
- Natural language. The last style is known as natural dialogue because here the interactions are sometimes very structured and the user has the right to ask what he wants and the system should provide him an answer while taking into account the context and the history of the interaction.

4. Proposed model of dialogue

In this section, we propose a solution for the conception of a flexible and robust dialogue framework at the level of a vocal interface. The solution is presented as a model of dialogue because it could serve as a basis to effectively design and implement dialogue systems. The method used to present our model is as follows: a certain number of quality criteria that the dialogues should verify are set. We then, propose strategies adapted with respect to those criteria. The strategies proposed here integrate in them the following:

- The process to implement dialogues
- The effective progress of dialogues (questions, answers, navigation)
- The methods of control and validation of entered data.

4.1. Quality criteria of a dialogue

Any dialogic interactive system should have a certain number of qualities or at least should fulfill some ergonomic requirements if one wants the system to effectively be used.

1) Qualities from a user perspective:
   a) The user must be able to converse in a natural way; he should not have too many restrictions in his speaking habits. This is the criteria of system robustness (It includes the ambiguousness of the natural language and speaking and behavioral variability of the user. This criterion is linked to the model of the task).
   b) The user should be able, if he wants it, to interrupt the system, to reorientate his demand, to ask it to repeat or re-precise questions.
   c) The statement must be taken in context; the system must be able to solve language's ellipses, anaphors and other particularities of the natural language.
   d) Offer to the user a robust linguistic model with extended vocabulary in the whole application domain.
   e) The dialogue must be able to mix initiative dialogue to put the user at ease. Hence, the system should accept all information given by the user, even though they don't answer the asked question, and it must be able to guide the user if he is lost. This point contributes to create a cooperative dialogue.
   f) The user should be able to take the initiative to correct as soon as he detects a mistake due to a bad answer or to a bad interpretation by the system.

g) Possibility of the user to detect and to correct mistakes as quickly as possible.

2) *Qualities from the application perspective:*

a) The system should be able to negotiate, that means ask for the precisions or clarifications to the user if necessary.

b) The system should be able to ask for confirmations and re-formulations of answers in order to avoid possible mistakes.

c) System's reactions must be cooperative.

d) Answers generated from the system should be adapted to the user; the vocal synthesis should be intelligent, natural and adapted from a prosodic standpoint.

e) The system should function real-time. The user should not wait for the systems' reaction to his request.

4.2. Implementation process

Above we have stated the criteria of qualities that our model of dialogue should respect. This model takes into account the following features:

- User model (profile, intentions)
- Style of dialogues (situation of dialogue, progress)
- Linguistic model (lexicon, syntax, grammar).
- Model of the task and the domain i.e. what the application is supposed to achieve, its real expertise).

The following list presents steps generally followed when building a vocal interface:

- Determine the set of information that the application will need the user to enter
- Design dialogues using *UML (Unified Modeling Language)* diagrams (sequences, state transition …) as input material.
- Build grammars to recognize users' statements during dialogues
- Anticipate all possible answers of the system to users. Prepare pre-recorded files or texts to be synthesize.
- Perform tests of interactions between basic dialogues and application logic.

4.3. A word on technological infrastructure

Another strategy that can be applied in order to succeed in building a high-level quality dialogue is the good selection of technologies to use. To present a detailed description of the various technologies available in this area is out of the scope of this document. However, we want to note that there are two main tendencies in technology selection:

- The range of solutions proposed by the *W3C (World Wide Web Consortium)* for the vocal interfaces. The W3C proposes *VoiceXML (Voice eXtensible Markup Language)* as a standard for the development of telephonic applications over the web. VoiceXML permits to return the content and the services of the Web accessible from a simple telephone. It leans on the Internet infrastructure to enable the creation of a new generation of vocal services from a simple server HTTP [9].
- The Language *SALT (Speech Application Language Tag)* proposed by the SALT Forum [7].

5. Realization of a prototype

We described in the previous section a set of strategies and procedures to follow to build a vocal interface that fulfill some quality criteria. In order to put in evidence all these concepts and at the same time show how to implement them we propose the realization of a prototype. The analysis, the design, the implementation and the tests will highlight elements of the development cycle specific to vocal interfaces.

Since our model had to be applied to the collection of information, we opted for the gathering of information necessary to the server that process admissions management in an academic campus.

5.1. Subject matter area: Pervasive University Institute (PUI)

The academic process of the PUI is made up of the following elements: Orientation, pre-registration, registration, time schedule, exams. The application to build should be able to automate the above described process. Students, teachers, university's managers, etc. should be able to access the application from a vocal portal. The application should integrate every element of the academic process as described in the continuation.

a) Orientation: Orientation consists in providing to the candidates information on the different faculties available in the university center, the courses dispensed, the outlets as well as the conditions of entry in each of the faculties (departments) of the academic center. All information concerning the faculties and departments should be available on the portal

b) Pre-registration: Pre-registration has two phases: deposit of a candidacy file and lists publication. Any candidate willing to register for the first time to one of the university faculties shall go through the pre-registration phase. During that phase the candidate will have to provide in his file information comprising personal information, his background as well as his desired faculty. A pre-registration form exists on the vocal portal. A reference number is then assigned to the candidate. He can use it whenever needed.

5.2. Domain analysis

1) Actors:

a) The candidate. He is any person who has asked for a pre-registration in a department of the university

b) The jury. This is a group of people incharge of pre-registration files processing. They also publish the lists of accepted candidates.

c) The visitor. This is any person who consults admission information of the university.

2) Use cases:

TABLE I
SYSTEM USE CASES

Use case	Actor
Consult the orientation	Everybody
Fill the pre-registration form	Candidate
Study a candidacy file	Jury
Establish lists of candidates	Jury
Consult lists of candidates	Candidate

3) Database: ***The database schema is given in figure 1. below***

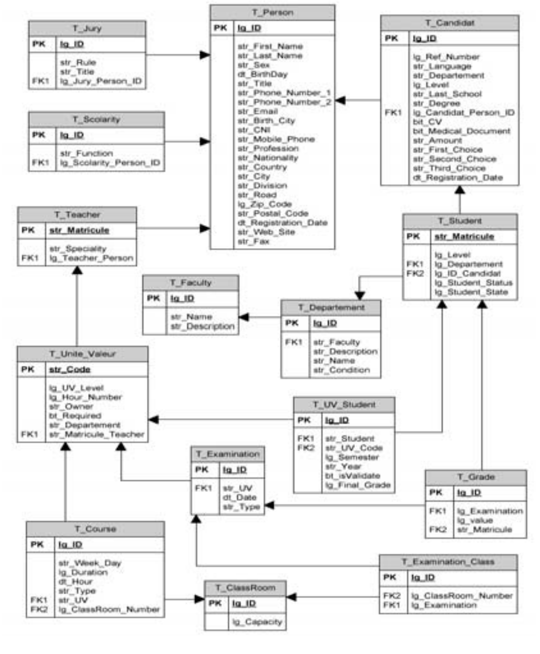

Figure. 1. Pervasive University Institute vocal portal database schema.

5.3. Dialogues and grammars design

At this point we want to specify that the only interaction mode is the voice. All interaction between the system and the user will be done orally. There are cases where the use of the keyboard and keypad of a phone would have been more appropriate (DTMF), but the objective here is to implement the correction strategy based on the cooperation between the intervenient parties. The design of dialogue scripts as well as their progress is useful for the development of the grammars. It is necessary before going further to the conception and implementation of the grammars that constitute the keys of success of the portal to identify all possible situations of dialogue.

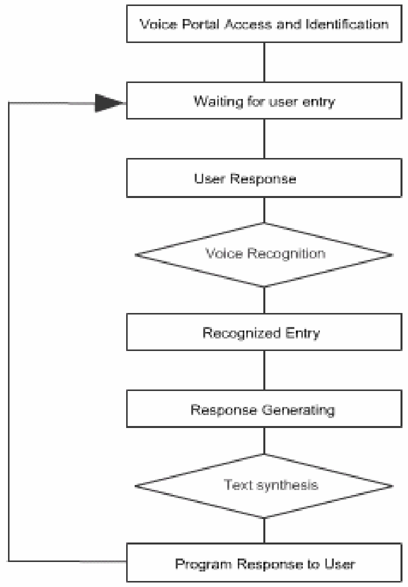

Figure. 2. Dialogues progress.

dialogues between the user and the system.

i) Scenario of progress dialogues

Figure 2 shows the progress of dialogues. It can be interpreted as follows:

1) The user connects to the portal
2) The system asks a question and waits the users' response.
3) The user answers the question
4) The system performs the input recognition.
5) The system builds its own answer and passes it to the synthesizer.
6) The synthesizer reads the asnswer.

ii) Scenario of progress dialogues

Real dialogues scenarios are very complex and take into account several aspects of a dialogue that could be difficult to

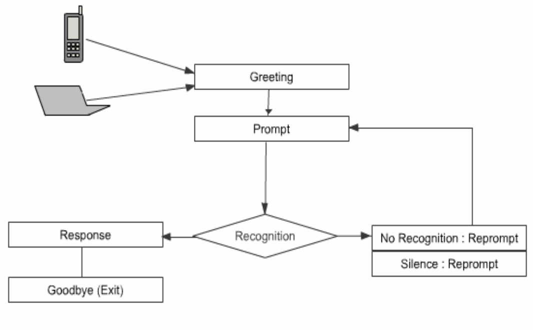

Figure. 3. Scenario of question/answer between the system and the user.

show on a flow chart. However, to have an idea, the figure 3 is a simple example that shows what really takes place during exchanges between the user and the system.

iii) Scenario of new user detection on the site

When a user connects to the portal, the system asks him if he has ever visited the site before. He should either answer

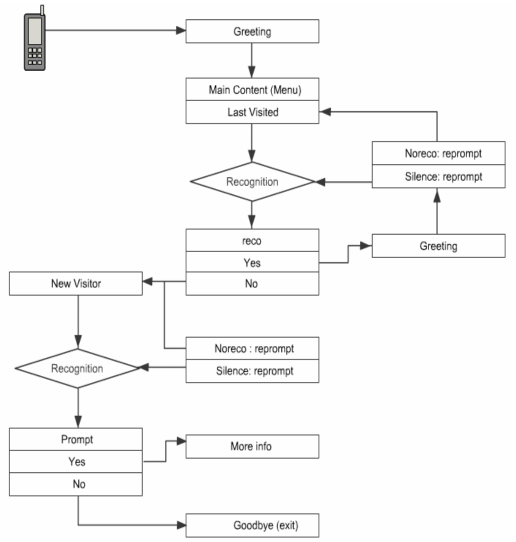

Figure. 4. Scenario of a new user detection on the portal.

affirmatively or negatively, but he is not obliged to answer by “yes” or “no” as shown on the figure 4. The user has the possibility to build whole sentence as their answer. Situations where the system asks the user to enter a phone number. This is a typical scenario for all types of data.

5.4. Construction of grammars

We have already mentioned that the power and the robustness of an interactive vocal system relies on the flexibility of the grammar used to perform the interpretation of users’ statements. Once the diverse dialogues’ scenarios are designed, it important to build the grammar to recognize and process whatever shall be said by the user during a dialogue session with the interface. For example the grammar to recognize affirmative and negative answers of the user should recognize expressions such as “yes”, “yes please”, “yes thanks”, “correct” for the affirmative; and “no”, “no thanks”, “no thank you”, for the negative. Within the environment *.Net Speech SDK (Software Development Kit)* from Microsoft Corporation, a subset of the previous grammar is represented.

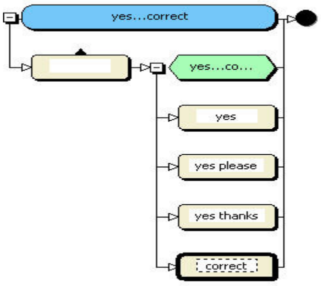

Figure. 7. Affirmative grammar within the .Net Speech SDK (Grammar [8]). as shown on the figure 7.

6. Development and test

Our development uses the .NET Speech SDK with the language SALT that we have already introduced.

For the test we used Internet Explorer and specified the parameter “voice only” in the page URL. The following shows the progression of a dialog with a visitor on the portal (NB: The application was developed in the context of a multimodal application. This is why all along the scenarios are references to pages load):

1) *System:* Welcome on the university portal. If you want an admission, then say “admission” or if you are just a visitor say “visit”.
2) *Visitor*: [ ]
3) *System:* I didn’t understand you. Say “admission” or “visit”.
4) *Visitor*: Admission!
5) *System [the admission form is loaded]:* Welcome on the admission page. If you have registered before, then give your reference number to check your information. Otherwise demand a new registration.
6) *Visitor*: New registration
7) *System:* [the new registration page is loaded].

When the registration page is loaded the dialogue continues as follows:

1) *System:* Welcome on the university portal. Please enter your personal information.
2) *System:* What is your first name?
3) *Candidate*: I am mister Omar Bakary
4) *System:* [the system saves the title, the sex, the last name and the first name of the candidate] Mister Omar Bakary, please what

is your birth date?

5) *Candidate*: the 7 December, 1986

6) *System:* [the system saves the day, the month and the year of birth of the candidate] the 7th ?

7) *Candidate*: yes

8) *System:* June?

9) *Candidate*: no December

10) *System:* December

11) *Candidate*: yes

12) *System:* 1986?

13) *Candidate*: [silence]

14) *System:* In which town were you born?

15) *Candidate*: Maroua in Cameroon

16) *System:* [saves the town and the country of the candidate] what is your profession?

17) *Candidate*: student

18) *System:* [saves the candidate profession] could you say your national identity card number?

19) *Candidate*: 1.1.2.0.3.4.9

20) *System:* [saves the ID number]. 1.1?

21) *Candidate*: yes

22) *System:* 2.0.3?

23) *Candidate*: [silence]

24) *System:* 4.9?

25) *Candidate*: [silence]

26) *System:* Here is the summary of your personal information. You are mister Omar Bakary,

27) *Candidate*: No, my last name is Boubakary.

28) *System:* Here is the summary of your personal information. You are mister Omar Boubakary, born on the 7 December 1986 in Maroua in Cameroon. You are a student and your national ID number is 1.1./2.0.3./4.9.What is your first name? Please could you say next to continue your registration?

29) *Candidate*: Next

30) *System:* [the system loads the next page]

7. Conclusion

This paper was aimed at the provision of a vocal interface for conversations. We have presented:

- The principle and the elements necessary for its functioning
- A dialogue model as well as a guide for the conception of the vocal interfaces.

At the end of this paper, we have enough reason to confirm that today we have sufficient resources to provide such a concrete solution to the difficult issue of establishing a dialogue between humans and machines. It is possible to provide vocal interfaces for professional as well as occasional use, integrating in it the most important aspect of human–machine dialogue which is spoken language with the aid of technologies like VoiceXML and SALT. These two languages offer utilities permitting the formulation of a flexible and at the same time a very powerful vocal interface. One of these languages is meant strictly for vocal applications while the other is use for multi-mode applications. Standard languages like SRGS (Speech Recognition Grammar Specification) are languages used for describing grammar attributes. This language permits the conception of some applications with knowledge of a given semantic level.

8.References

[1] S. Ndongna, “Application multimodale : combinaison et synchronisation des modes d’entrées sorties,” M.Eng. thesis, Dept. Genie Informatique., Ecole Nationale Supérieure Polytechnique, Yaoundé, Cameroon, 2003.

[2] Caelen Jean. CLIPS-IMAG. (2003, March). Interaction multimodales. Available: http://www-geod.imag.fr/caelen/cours

[3] Programme de Recherche en sciences Cognitives de Toulouse. Approche pluridisciplinaire du traitement de l’erreur dans le dialogue oral homme-machine. Available: http://www.irit.fr/ACTIVITES/PRESCOT/Prescot.f.html

[4] Bob Carpenter, Jennifer Chu-Carrol, Natural Spoken Diaog System, Lucent Technologies Bell Labs, U.S.A , 20 Juin 1999.

[5] BOUSQUET-VERNHETTES Caroline, Mémoire de DEA, Décodage Conceptuel Robuste De La Parole Spontanée Dans Les Serveurs Vocaux Interactifs, UNIVERSITÉ Paul SABATIER, Année 1998.

[6] ZUE V., GLASS J., GOODINE D., H. Leung, M. Phillips, J. POLIFRONNI & S. SENEFF, the proceedings of the second DARPA Speech and Natural LanguageWorkshop, held in Harwichport, MA, 15-18 october 1989.

[7] Speech Application Language Tags 1.0 document specification, http://www.saltforum.org/downloads/SALT1.0.pdf, dernière consultation : 24 Avril 2003.

[8] Speech Recognition Grammar Specification for the W3C Speech Interface Framework, W3C Voice Browser Activity, http://www.w3.org/TR/speech-grammar , derniere consultation mars 2003.

[9] Voice Extensible Markup Language (VoiceXML) Version 2.0, W3C Voice Browser Activity http://www.w3.org/TR/voicexml20 , derniere consultation Mai 2003.

[10] An Introduction to IBM Natural Language Understanding An IBM White Paper, http://www.us.ibm.com/, 2001.