DORIS, a multiagent/IP platform for multimodal dialogue applications

J. L’Hour, O. Boeffard, J. Siroux, L. Miclet, † F. Charpentier, ‡ T. Moudenc

IRISA/Cordial – Université de Rennes 1
{johann.lhour,olivier.boeffard,laurent.miclet,jacques.siroux}@irisa.fr
†Télisma – Lannion, France, fcharpentier@telisma.com
‡France Telecom R&D – Lannion, France, thierry.moudenc@rd.francetelecom.com

Abstract
This article presents an effort to define a multimodal Agent-based dialogue platform cooperating with Internet technologies. We propose an open architecture integrating voice based and graphical user interfaces. This work is highlighted by a demonstrator, GEORAL, for the tourist information retrieval domain.

1. Introduction
Over the last years, speech technologies such as text-to-speech synthesis, speaker and speech recognition have made significant progress in providing efficient voice based user interfaces. Meantime, multipurpose data processing devices such as personal handheld computers became more and more powerful. The main advantage of such devices is their mobility but they offer a small visual panel to the user and the distinction between phones and handheld computers is less and less perceptible. From a transport network point of view, telecommunication operators offer solutions to the voice/data convergence issue. One can find more and more applications where speech cooperates with other communication modalities – for example a visual tactile interface, a gesture capture unit, etc. –. Mobile devices become smaller and smaller, the limit of this evolution depends on the main functionality and physical acceptability of such devices by the end-user. Concretely, it should be unrealistic to impose a 10” visual panel only for the reason that a service needs a specific visual functionality (like web surfing). In this context, one can’t circumvent the use of speech modalities. Users can get information from an interaction system using text-to-speech synthesis and react with their own voice. To relief the technological constraints of mobile access devices, speech is a natural way to exchange knowledge for humans. Moreover, one can take advantage from mixing the access modalities.

Considering all these previous points, this article presents an effort to define a multimodal vocal platform called DORIS whose goal is to achieve a full integration of speech, gesture, and dialogue technologies. Several publications have reported on efforts in building such a platform and several issues need to be addressed. Among those, we focus in this paper on the distributivity of the solution based on an Agent architecture, and on the use of Voice over IP solutions, and we illustrate such issues through a demonstration application built upon such an architecture. Additionally, this platform helps us to integrate different third-party solutions – speech bundles, VoIP protocols, applications, etc. – and test them in an acceptable technological environment.

The rest of the paper is organized as follows. In section 1, we describe the platform architecture, its organization and its specificity from a packet-data network point of view. Section 2 provides an overview of the main components of the platform (we do not focus here on the middleware components). Section 3 illustrates the use of the DORIS platform providing an application in the scope of a tourist information service called GEORAL. Section 4 finally summarizes the contributions of this paper.

2. System Architecture
DORIS, which abbreviates DialOg Research Integration System, is a platform designed to facilitate the process of multiple cooperative functions in order to build multimodal dialogue applications. Most of the platforms, e.g. DARPA-Communicator/GALAXY-II [2] or BLSTIP [1] share roughly the same kind of architecture. Our proposal focus on the interaction of a multimodal dialogue platform within an Internet framework. On one hand, we need a distributed software system. We choose an Agent-based middleware for its high-level of abstraction. Our choice was based on the FIPA – Foundation for Intelligent Physical Agent – standards [3]. FIPA releases specifications that describe protocols for cooperating Agents. The Agents methodology applies concepts from artificial intelligence and speech act theory. More precisely, an Agent is a piece of software which is autonomous, proactive and social within a network of peers. We choose the JADE1 implementation of the FIPA requirements. JADE is an open source project carried out by TiLAB developed with Java and tested over a wide range of environments [4]. On the other hand, we need a standard thin client access to the platform. We choose a client/server based solution implemented within the applet/servlet framework and operational for standard web browser. The client applet uses the JAVA media framework to access audio recording and

1JADE is a trademark of Tilab
Figure 1: Implementation of DORIS Agents over the JADE middleware which is a FIPA 2000 compliant middleware. We define a logical area where Agents (white disks with arrows) cooperate for a common task, e.g. GEORAL application. Proxy-Agents (dashed disks) communicate to the outside of the Agent part of the platform. Specific proxy-Agents, HTTP and streaming audio, are isolated in a DMZ zone for Internet access to the platform.

2.1. DORIS implementation

The DORIS platform is built over the JADE middleware. We have defined a DORIS class Agent which extends the JADE class Agent. This class provides specific methods to access Text-to-speech, TTS, and Automatic Speech Recognition, ASR, resources through the MRCP protocol [5].

On the Agent side, Agents take advantages of the FIPA specification in order to communicate in a standardized way through well known interaction protocols. At present, we use performative speech act message in a basic way. We plan to implement interaction protocols like the query or request FIPA protocol. One Agent is responsible for a well-defined function. That is, over different applications, we factorize common Agents which are not domain specific. More precisely, some Agents are devoted to access voice processing servers and hence are application independent. Besides they are built over the same interaction suite protocols, others Agents, e.g. the GEORAL’s thematic analysis Agent, are application specific.

Figure 3 illustrates the communication between Agents running the GEORAL application.

Figure 2: Chronogram of DORIS Agents processing a request for the GEORAL application. Each Agent, reco, anasyt, anathem, filter or fusion exchanges knowledge using communicative acts.

2.2. IP Network components

It is imperative to offer a wide and simple access to the platform. We choose the current client/server technology built on standard web browser for the client and HTTP/applet with JSP, dynamic HTML page generation, for the server side. Java applet for the client side benefits of the large diffusion of the JAVA runtime machine, J2SE for standard desktop system or J2ME for embedded systems like mobile phones.

For network security reasons, the front-end applets providing the user interface access to a DMZ (demilitarized zone) network zone. The applets communicate with a unique reference Agent acting as a dispatcher for other Agent resources. All the audio and text traffic goes through this Internet gateway, the rest of the platform is secured behind a firewall, the cooperative Agents and multimodal resources – TTS, ASR, Gesture –. The audio streams are built on the RTSP/RTP Internet protocol. On the client side we have implemented audio access with the Java Media Framework. Different coding payloads are available depending on which material is used: for example, a GSM coding when the applet runs on a cellular phone.

Another salient feature of our proposed solution is to mask the third-party API specificities behind the MRCP protocol [5], Media Resource Control Protocol. MRCP controls media service resources like speech synthesizers, recognizers, signal generators, signal detectors, fax servers etc. over a network. This protocol is designed to work with streaming protocols like RTSP (Real Time Streaming Protocol) which help establish control connections to external media streaming devices, and media delivery mechanisms like RTP (Real Time Protocol). RTSP protocol is a standard protocol for controlling the delivery of data with real-time properties. The main contribution of this protocol to our plat-
form concerns the negotiation of the RTP, setup parameters (client and server port numbers, session id) and the transport of MRCP messages between client and proxy-Agents dedicated to speech resources. We have defined half-duplex streaming. A client can initiate a session on the DORIS platform from one source, for example a PDA, and get a speech feedback from another source, for example with a cellular phone.

In order to facilitate technical access for some industrial partnerships, the platform includes fast secure network access. DORIS propose high throughput Internet connection with VPN access.

3. Basic Agent resources

In this section we introduce the basic multimodal resources integrated in the DORIS platform. Each resource is linked to a specific JADE Proxy-Agent. This Agent exchanges messages with its colleagues within the Agent platform and is responsible for accessing the required modal resources which is executed outside the Agent platform.

3.1. Dialogue resource

The core of a man-machine communication system has to be considered as an intermediary between the user and the application. It must have functions in order to understand and interpret user’s activities taking into the dialogue and application histories, to manage the application (which uses in general a database software) and to lead the interaction using the usual rules of dialogue, e.g. Gricean maxims [6]. Depending on the chosen application, these functions may be more or less complex, more or less integrated and can be processed in various orders. Agent-based technology appears as a good choice for implementing such a system. We propose a set of generic Agents in order to take into account these functions. Speech user inputs can be parsed by a parser based on a "list difference algorithm" written in Prolog. The semantic and pragmatic processes, written in Prolog, produce communication acts which are based on the speech acts theory and which have the same general format that the messages exchanged between the Agents. User’s activities on the touch screen are contextually processed at a semantic level (but taking into account specificities of the application), written in JAVA. The dialogue module Agent, written also in Prolog, is based on a simple automaton; it is able to manage interactions with the application database (actually a set of Prolog facts) and to produce co-operative responses to the users.

3.2. Text-to-Speech synthesis resource

A TTS system performs the automatic generation of speech from text. One can easily extend this definition to a variant starting from concepts or more precisely from an internal sentence model within the generative function of a dialogue system. The FTR&D speech synthesis engine generates speech by concatenating speech units stored in a reference database. The methodology consists in storing the waveform representation of each unit. Until recently, one should clearly make a compromise between the footprint of the reference database and the intelligibility and quality of the synthetic speech. Today, centralized TTS resources can take benefit of the availability of cheap mass storage. The corpus-based FTR&D speech synthesis engine follows the recent trend in the evolution of TTS systems that has been to store a very large acoustic corpus (about 250,000 acoustic units pronounced within different prosodic contexts). The system is organized in a way to separate clearly the linguistic and acoustic functions. First, the text or an utterance model is processed in order to define a prosodic template for the message to pronounce. Second, phonological information is used to find the best sequence of units, and concatenate speech segments in order to generate the stream of samples.

In corpus-based TTS systems, we move from a signal processing problem to a search strategy one. The main task is now to find the best sequence of units among an intractable combinatorial space of sequences. But, thanks to this large amount of speech units this new speech synthesis solution converts text to speech with a voice timber that is extremely close to natural speech. The TTS resource is located out of the Agent platform for third party integration constraints (SAPI 4 compliant). From the DORIS side, an Agent who needs a TTS resource requests a Proxy-Agent dedicated to speech technologies implementing the MRCP protocol (see section 2.2).

3.3. Speech recognition resource

Developed and marketed by Telisma, the core Philsoft® V3 multilingual Automatic Speech Recognition engine is based on multigaussian hidden Markov models at the phone level. The topology of each model depends on features at a phonetic level. The Adaptive Speech/Noise detection function offers a speech-based detection rather than energy threshold-based signal/noise detection. Grammars can be automatically generated in real time thanks to a Dynamic Grammar Builder. This functionality enables the design of voice applications with highly evolutive content, which is particularly suited to multi-topic dialogue applications. Additionally, the ASR process can be distributed over the network using a DSR functionality. This distributed mode is very powerful when small bandwidth or lossy channels are present in the end-to-end connection between the user and the platform. The ASR engine recognizes French, UK English, US English, German, Spanish, and several other languages. Thanks to the SNMP protocol associated to a MIB description, alarms and load activity reports can then be transmitted to monitoring applications. Within DORIS, the core technology is embedded in client/server solution. As the TTS resource, the ASR resource is located out of the Agent platform for third-party integration reasons. From the DORIS side, an Agent who needs an ASR resource requests a Proxy-Agent dedicated to speech technologies and implementing
the MRCP protocol.

3.4. Gesture resource

DORIS visual devices are provided with touch screens which allows a user to perform various types of designation: pointing, multiple pointing, drawing a line or an area or following a contour. These activities are collected by an Agent as events which contain the coordinates of each point touched on the screen. The Agent filters these events and possibly constructs a higher level representation like a surface or a curve. The final interpretation of the activity has to be processed by another Agent which will be depending on the application.

4. GEORAL

GEORAL is a multimodal application of which purpose is to provide information of a touristic nature to naïve users. Users can ask for information about the location of place of interests (city, beach, castle, church, ...) within a region, a small area, along river or road, or distance and itinerary between two localities. Users interact with the system using three different modalities: in a visual mode by looking to the map displayed on the screen, in a speech mode using natural language and in a gesture mode by pointing or drawing on the touch screen. The system itself uses both the speech channel (text-to-speech synthesis) and graphics such as displaying the interpreted touching activity of the user, the flashing of sites, routes and zooming in on subsections of the map, so as to best inform the user. Multimodality in input offers better conditions of use (user point of view). But, from a system point of view, it has two contradictory effects. Firstly, it is necessary to deal with the difficult problem of references: speech and gesture can be complementary (e.g. please give me the beaches in this area + drawing a zone). We merge the representation coming from the types of input by means of a plan-based approach [6] which operates on the representation of the communication acts. On the other hand, multimodality may bring the system some redundancy (e.g. “please give me the churches in Lannion + pointing the town Lannion”) which allows to reinforce word recognition confidence. Dialogue with users is semidirected: the system manages the dialogue by questioning users on necessary parameters, by producing cooperative and suggestive answers in order to avoid unproductive transaction; the user can interrupt the dialogue, and utter some phatic commands.

5. Conclusion

We have presented an Agent-based dialogue platform covering the field of human-machine interfaces. This work focuses on the third-party integration of speech technologies using the MRCP Internet protocol and streaming solutions over IP flows. In order to illustrate these technological building-blocks, we have developed a multimodal application, GEORAL, for providing information of a touristical nature.

![GEORAL Agents cooperating over the JADE middleware.](image)

6. References


