Speech, Image, and Language Processing for Human Computer Interaction:

Multi-Modal Advancements

Uma Shanker Tiwary
Indian Institute of Information Technology Allahabad, India

Tanveer J. Siddiqui
University of Allahabad, India

Information Science
REFERENCE

Managing Director: Lindsay Johnston
Senior Editorial Director: Heather A. Probst
Book Production Manager: Sean Woznicki
Development Manager: Joel Gamon
Development Editor: Myla Harty
Acquisitions Editor: Erika Gallagher
Typesetter: Jennifer Romanchak
Cover Design: Nick Newcomer, Lisandro Gonzalez

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Speech, image, and language processing for human computer interaction : multi-modal advancements / Uma Shanker Tiwary and Tanveer J. Siddiqui, editors.
p. cm.
Includes bibliographical references and index.
Summary: “This book identifies the emerging research areas in Human Computer Interaction and discusses the current state of the art in these areas”--Provided by publisher.
ISBN 978-1-4666-0954-9 (hardcover) -- ISBN 978-1-4666-0955-6 (ebook) -- ISBN 978-1-4666-0956-3 (print & perpetual access) 1. Human-computer interaction. I. Tiwary, Uma Shanker, 1960- II. Siddiqui, Tanveer.
QA76.9.H85S654 2012
004.01’9--dc23
2011047364

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 13
On the Development of Adaptive and User-Centred Interactive Multimodal Interfaces

David Griol
Carlos III University of Madrid, Spain

Ramón López-Cózar
University of Granada, CITIC-UGR, Spain

Zoraida Callejas
University of Granada, CITIC-UGR, Spain

Gonzalo Espejo
University of Granada, CITIC-UGR, Spain

Nieves Ábalos
University of Granada, CITIC-UGR, Spain

ABSTRACT

Multimodal systems have attained increased attention in recent years, which has made possible important improvements in the technologies for recognition, processing, and generation of multimodal information. However, there are still many issues related to multimodality which are not clear, for example, the principles that make it possible to resemble human-human multimodal communication. This chapter focuses on some of the most important challenges that researchers have recently envisioned for future multimodal interfaces. It also describes current efforts to develop intelligent, adaptive, proactive, portable and affective multimodal interfaces.

1. INTRODUCTION TO MULTIMODAL INTERFACES

With the advances of speech, image and video technology, human-computer interaction (HCI) has reached a new phase, in which multimodal information is a key point to enhance the communication between humans and machines. Unlike traditional keyboard- and mouse-based interfaces, multimodal interfaces enable greater flexibility in the input and output, as they permit users to employ different input modalities as well as to obtain responses through different means, for example, speech, gestures and facial expressions.

DOI: 10.4018/978-1-4666-0954-9.ch013

This is especially important for users with special needs, for whom the traditional interfaces might not be suitable (McTear, 2004; López-Cózar & Araki, 2005; Wahlster, 2006).

In addition, the widespread use of mobile technology implementing wireless communications such as personal digital assistants (PDAs) and smart phones enables a new type of advanced applications to access information. As the number of ubiquitous, connected devices continues to grow, the heterogeneity of client capabilities and the number of methods for accessing information services also increases. As a result, users can effectively access huge amounts of information and services from almost everywhere and through different communication modalities.

Multimodality has been traditionally addressed from two perspectives. On the one hand, human-human multimodal communication. Within this area we can find in the literature studies concerned with speech-gesture systems (Catizone et al., 2003), semiotics of gestures (Radford, 2003; Flecha-García, 2010), structure and functions of face-to-face communication (Bailly et al., 2010), emotional relations (Cowie & Cornelius, 2003; Schuller et al., 2011), and intercultural variations (Endrass et al., 2011; Edlung et al., 2008). On the other hand, human-machine communication and interfaces. Topics of interest in this area include, among others, talking faces, embodied conversational agents (Cassell et al., 2000), integration of multimodal input, fission of multimodal output (Wahlster, 2003), and understanding of signals from speech, text, and visual images (Benesti et al., 2008).

This chapter focuses on some of the most important challenges that researchers have recently envisioned for future multimodal interfaces. It describes current efforts to develop intelligent, adaptive, proactive, portable and affective multimodal interfaces. All these concepts are not mutually exclusive, for example, the system's intelligence can be concerned with the system's adaptation enabling better portability to different environments.

There are different levels in which the system can adapt to the user (Jokinen, 2003). The simplest one is through personal profiles in which the users have static choices to customize the interaction (e.g., whether they prefer a male or female system's voice), which can be further improved by classifying users into preference groups. Systems can also adapt to the users' environment, for example, Ambient Intelligence (AmI) applications such as ubiquitous proactive systems. The main research topics are the adaptation of systems to different expertise levels (Haseel & Hagen, 2005), knowledge (Forbes-Riley & Litman, 2004), and special needs of users. The latter topic is receiving a lot of attention nowadays in terms of how to make systems usable by handicapped and elderly people (Heim et al., 2007; Batliner et al., 2004; Langner & Black, 2005), and how to adapt them to user features such as age, proficiency in the interaction language (Raux et al., 2003) or expertise in using the system (Haseel & Hagen, 2005).

Despite their complexity, these characteristics for the design of user centred multimodal interfaces are to some extent rather static, i.e., they are usually gathered a priori and not during the dialog, and thus they are not used to dynamically adapt the multimodal interface at some stage in the interaction. There is another degree of adaptation in which the system not only adapts to the messages conveyed during the interaction, but also to the user's intentions and emotional states (Martinovski & Traum, 2003; Prendinger et al., 2003). It has been demonstrated that many breakdowns in man-machine communication could be avoided if the machine was able to recognize the emotional state of the user and responded to it more sensitively, for instance, by providing more explicit feedback if the user is frustrated. Emotional intelligence not only includes the ability to recognize the user's emotional state, but also the ability to act on it appropriately (Salovey & Mayer, 1990).

To deal with all these important topics required for the design of adaptive and user-centred interactive multimodal interfaces, this chapter is organized as follows. Section 2 provides an overview on the main architectures and toolkits available for the development of such systems. Section 3 describes the main principles involved in the development of multimodal interfaces which are adaptive to the user's location and activities without requiring explicit user inputs. The section also provides examples of multimodal systems implemented to incorporate such contextual information, and discusses various aspects concerned with emotion recognition and affective responsivity of multimodal systems. Section 4 describes our work related to the development of interactive multimodal interfaces. Finally, Section 5 presents the conclusions and outlines possibilities for future research directions.

2. DEVELOPMENT OF MULTIMODAL INTERFACES: ARCHITECTURES AND TOOLKITS

Multimodal interfaces involve several user senses simultaneously during the communication with the computer. We are particularly interested in systems which employ voice as a relevant communication modality for the input and output (Griol et al., 2008; Callejas & López-Cózar, 2008a). In this section we provide an overview on the main architectures and development toolkits available for the development of such systems.

Multimodal dialogue systems can be defined as computer programs designed to interact with users *similarly* as human beings would do, using more or less interaction modalities depending on their complexity (McTear, 2004; López-Cózar & Araki, 2005). These programs are employed for a number of applications, including tutoring (Forbes-Riley & Litman, 2011), entertainment (Ibrahim & Johansson, 2002), command and control (Stent et al., 1999), healthcare (Beveridge & Fox, 2006), call routing (Paek & Horvitz, 2004) and retrieval of information about a variety of services, for example, weather forecasts (Maragoudakis, 2007), apartment rental (Cassell et al., 1999) and travels (Huang et al., 1999). A detailed classification of these systems using different criteria (languages, domains, functionalities, interaction degrees, input and output modalities, etc.) can be found in López-Cózar and Araki (2005) and McTear (2004).

2.1. Approaches to Incremental Development

In order to develop usable multimodal interfaces, it is necessary to take the user perspective into account from the early stages in the development cycle. The system-in-the-loop technique is based on the fact that software systems improve cyclically by means of user interactions. For example, the performance of a speech-based multimodal interface can be improved by means of analyses of sentences previously uttered by users. If modifications are needed in the design of the system, the technique is employed again to obtain new experimental results. These steps (collection of data and test of system) are repeated until the system designers are satisfied with the performance. Among others, Van de Burgt et al. (1996) used this technique to implement the SCHISMA system. In particular, the technique was used to collect user utterances and analyse them in order to improve the performance of the system.

It is possible to collect user utterances from an early design of the system using the Wizard of Oz (WOz) technique, in which a human *Wizard* plays the role of the computer in a human-computer interaction (Fraser & Gibert, 1991). The users are made to believe that they interact with a computer but actually they interact with the Wizard, who decides the system's responses considering the current design planned for the interface. Salber and Coutaz (1993) discussed some requirements of WOz for multimodal systems. They indicated that a multimodal system is more complex to simulate than

a system based on speech only, which increases the task complexity and the bandwidth necessary for the simulation. For multimodal interaction, the authors suggested to employ a multi-wizard configuration, which requires properly organising the work of several wizards. A platform for multimodal WOz experiments must have a high performance and flexibility, and should include a tool to retrieve and manipulate data collected during the experiments.

2.2. From Speech to Multimodality

The implementation of multimodal systems is a complex task in which a number of technologies are involved, including signal processing, phonetics, linguistics, natural language processing, affective computing, graphics and interface design, animation techniques, telecommunications, sociology and psychology. The complexity is usually addressed by dividing the implementation into simpler problems, each associated with a system's module that carries out specific functions. Usually, this division is based on the traditional architecture of spoken dialogue systems: automatic speech recognition (ASR), spoken language understanding (SLU), dialogue management (DM), natural language generation (NLG) and text-to-speech synthesis (TTS).

ASR is the process of obtaining a sentence (text string) from a voice signal (Rabiner & Huang, 1993). It is a very complex task given the diversity of factors that can affect the input, basically concerned with the speaker, the interaction context and the transmission channel. Different applications demand different complexity on the speech recognizer. Cole et al. (1997) identified eight parameters that allow an optimal tailoring of the recognizer: speech mode, speech style, dependency, vocabulary, language model, perplexity, signal-to-noise ratio (SNR) and transduction. Nowadays, general-purpose ASR systems are usually based on Hidden Markov Models (HMMs) (Rabiner & Juang, 1993).

SLU is the process of extracting the semantics from a text string (Minker, 1998). It generally involves employing morphological, lexical, syntactical, semantic, discourse and pragmatic knowledge. In a first stage, lexical and morphological knowledge allow dividing the words in their constituents distinguishing lexemes and morphemes. Syntactic analysis yields a hierarchical structure of the sentences, whereas the semantic analysis extracts the meaning of a complex syntactic structure from the meaning of its constituents. There are currently two major approaches to carry out SLU: rule-based (Mairesse et al., 2009) and statistical (Meza-Ruiz et al., 2008), including some hybrid methods (Liu et al., 2006).

DM is concerned with deciding the next action to be carried out by the dialogue system. The simplest dialogue model is implemented as a finite-state machine, in which machine states represent dialogue states and the transitions between states are determined by the user's actions. Frame-based approaches have been developed to overcome the lack of flexibility of the state-based dialogue models, and are used in most current commercial systems. For complex application domains, plan-based dialogue models can be used. They rely on the fact that humans communicate to achieve goals, and during the interaction, the humans' mental state might change (Chu et al., 2005). Currently, the application of machine-learning approaches to model dialogue strategies is a very active research area (Griol et al., 2008; Williams & Young, 2007; Cuayáhuitl et al., 2006; Lemon et al., 2006).

NLG is the process of obtaining texts in natural language from a non-linguistic representation of information. It is usually carried out in five steps: content organization, content distribution in sentences, lexicalization, generation of referential expressions and linguistic realization. The simplest approach uses predefined text messages (e.g., error messages and warnings). Although intuitive, this approach is very inflexible (Reiter, 1995). The next level of sophistication is

template-based generation, in which the same message structure can be produced with slight differences. This approach is used mainly for multi-sentence generation, particularly in applications where texts are fairly regular in structure, such as business reports (Reiter, 1995). Phrase-based systems employ what can be considered generalized templates at the sentence level (in which case the phrases resemble phrase structure grammar rules), or at the discourse level (in which case they are often called text plans) (Elhadad & Robin, 1996). Finally, in feature-based systems, each possible minimal alternative of expression is represented by a single feature to obtain the maximum level of generalization and flexibility (Oh & Rudnicky, 2000).

TTS synthesizers transform text strings into acoustic signals. A TTS system is composed of two parts: front-end and back-end. The front-end carries out two major tasks. Firstly, it converts text strings containing symbols such as numbers and abbreviations into their equivalent words. This process is often called text normalization, pre-processing or tokenization. Secondly, it assigns a phonetic transcription to each word, which requires dividing and marking the text into prosodic units, i.e., phrases, clauses, and sentences. The back-end (often referred to as the synthesizer) converts the words in text format into sound. Concatenative synthesis employs pre-recorded units of human voice that are put together to obtain words. It generally produces the most natural synthesized speech; however, differences between variations in speech and in the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches.

Once the speech-based response of the system has been designed, it is possible to gradually incorporate other modalities. In order to do so, it is important to design a *fusion* module to combine information chunks provided by different input modalities of a multimodal interface. The result is a data structure that enables the multimodal system in handling different information types simultaneously. Using this data structure, the system's dialogue manager can decide what to do next. A number of methods have been proposed to represent the combined data. For example, Faure and Julia (1993) employed *Triplets*, which are a syntactic formalism to represent multimodal events in the form: (verb, object, location). The authors found this method very useful to represent speech information combined with deictic information generated by means of gestures. Allen (1995) proposed to use semantic structures called frames. The information from each modality was interpreted separately and transformed into frames, the slots which determined the parameters of the action to be made. Frames contain partial information if some slots are empty. During the fusion the frames are combined, fulfilling the empty slots. For example, Lemon et al. (2006a) used frames to combine multimodal information in a multimodal interface that provided information about hotels, restaurants and bars in a town. XML-based languages are other method to represent multimodal information. For example, Wahlster et al. (2001) used an XML-based language called M3L to represent all the information flows between the processing components of the SmartKom system.

2.3. Architectures for the Design of Multimodal Interfaces

It is important to properly select the architecture to be used for implementing a multimodal interface, since it should allow further enhancement or porting it from one application domain to another. We can find in the literature a number of architectures to implement multimodal interfaces.

Galaxy Communicator is a distributed, message-based, hub-centred architecture (Seneff et al., 1998), in which the main components are interconnected by means of client-server connections. This architecture has been used to set up, among others, the MIT's Voyager and Jupiter systems (Glass et al., 1995; Zue et al., 2000).

The Open Agent Architecture (OAA) architecture was designed to ease the implementation of agent-based applications, enabling intelligent, cooperative, distributed, and multimodal agent-based user interfaces (Moran et al., 1997). The agents can be developed in several high-level languages (e.g., C or Java) and platforms (e.g., Windows and Solaris). The communication with other agents is possible using the Interagent Communication Language (ICL). The cooperation and communication between the agents is carried out by means of an agent called Facilitator. Several authors have used this architecture to implement multimodal interfaces for a variety of application domains, including map-based tourist information (Moran et al., 1997), interaction with robots (Bos et al., 2003), and control of user movements in a 2D game (Corradini & Samuelsson, 2008).

The blackboard architecture was released considering principles of Artificial Intelligence. Its name denotes the metaphor of a group of expert people who work together and collaboratively around a blackboard to solve a complex problem. All the resources available are shared by the agents. Each agent can collaborate, generate new resources and use resources from other agents. A Facilitator agent controls the resources and acts as intermediary among the agents which compete to write on the blackboard, taking into account the relevance of the contribution of each agent. This architecture has been used to implement a number of multimodal interfaces (Raux & Eskenazi, 2007; Huang et al., 2007).

R-Flow is an extensible XML-based architecture for multimodal interfaces (Li et al., 2007). It is based on a recursive application of the Model-View-Controller (MVC) design. The structure is based on three layers: modality independent dialogue control, synchronization of logical modalities and physical presentation. Each one is codified in different XML-based languages. For example, State-Chart XML (SCXML) is used for dialogue control, whereas SMIL (Synchronized Multimedia Integration Language) and EMMA (Extensible Multimodal Interface Language) (Li et al., 2006) are used for modality synchronization and interpretation.

In addition to the architectures discussed above, which are amongst the most employed, it is possible to find other architectures in the literature. For example, Leßmann and Wachsmuth (2003) used the classical architecture Perceive-Reason-Act for the design of a multimodal interface. The *Perceive* module handles the input information, which is collected by auditory, tactile and visual sensors. The *Act* module generates the output information. Actions can be carried out by means of either *deliberative* or *reactive* behaviour. The component for deliberative behaviour uses knowledge about the domain, which is updated by perceptions, and generates intentions employing a plan library, which represents what the agent wants to do next. The second way of generating an action is by means of the reactive behaviour, which is reserved for actions that do not need deliberation, for example, making the agent appear more lifelike.

Following a different approach, Wei and Rudnicky (2000) proposed an architecture based on a task decomposition and an expectation agenda. The agenda is a list of topics represented by handlers. A handler encapsulates the knowledge necessary for interacting with the user about a specific information slot. The agenda defines a "plan" for carrying out a specific task, which is represented as a specific order of handlers.

2.4. Tools for the Development of Multimodal Systems

This section describes a number of tools and standards for developing multimodal systems that we consider relevant for this chapter. The discussion includes tools such as HTK, CSMU Sphinx, CMU SLM, NLTK, and standards such as VoiceXML, among others.

The Hidden Markov Model Toolkit (HTK) is free software for building and using Hidden

Markov Models (HMMs), which was developed by Cambridge University (Young et al., 2000). In the community of multimodal interfaces this software is primarily used for ASR, but it also has been used for a number of applications including speech synthesis, character recognition and DNA sequencing. It consists of a set of libraries and tools that provide facilities for speech analysis, HMM training, testing, and results analysis. The software supports HMMs using both continuous density mixture Gaussians and discrete distributions and can be used to build complex HMM systems.

CMU Sphinx (Lee et al., 1990) is an ASR system developed at the Carnegie Mellon University. There are several versions of it (Sphinx 2 - 4), each including an acoustic model trainer (SphinxTrain). The recogniser can deal with continuous, speaker-independent speech using HMMs and n-gram language models. Sphinx 2 focuses on real-time recognition suitable for speech-based applications and uses a semi-continuous representation for acoustic modelling. Sphinx 3 adopted the prevalent continuous HMM representation and has been used primarily for high-accuracy, non-real-time recognition. Sphinx 4 is written entirely in Java with the goal of providing a more flexible framework for research. PocketSphinx has been designed to run in real time on handhelds and be integrated with live applications. There is also a number of proprietary software for ASR, including AT&T WATSON, Windows speech recognition system, IBM ViaVoice, Microsoft Speech API, Nuance Dragon NaturallySpeaking, MacSpeech, Loquendo ASR and Verbio ASR.

The Carnegie Mellon Statistical Language Modeling Toolkit (CMU SLM) is a set of Unix tools designed to facilitate language modelling (Rosenfeld, 1995). The toolkit allows processing corpora of data (text strings) in order to obtain word frequency lists and vocabularies, word bigram and trigram counts, bigram and trigram-related statistics and a number of back-off bigram and trigram language models. Using these tools it is also possible to compute statistics such as perplexity, out-of-vocabulary words (OOV) and distribution of back-off cases.

The Natural Language Toolkit (NLTK) (Bird et al., 2008) is a suite of libraries and programs for symbolic and statistical natural language processing for the Python programming language. Other tools include Phoenix, designed by the Carnegie Mellon University in combination with the Helios confidence annotation module (Ward & Issar, 1994), and Tina, developed by the MIT and based on context free grammars, augmented transition networks, and lexical functional grammars (Seneff, 1989).

The Center for Spoken Language Understanding (CSLU) at the Oregon Health and Science University developed a graphical tool called CSLU Toolkit for the design of dialogue managers based on finite-state dialogue models (McTear, 1998). The toolkit includes tools for working with audio, display, speech recognition, speech generation, and animated faces.

The AT&T FSM library is a set of Unix tools for building, combining and optimizing weighted finite-state systems (Mohri, 1997). Some systems based on finite states were created under the SUNDIAL (Müller & Runge, 1993) and SUNSTAR projects (Nielsen & Baekgaard, 1992).

VoiceXML is the W3C's standard XML format for specifying interactive voice dialogues between humans and computers (McGlashan et al., 2004). The language is the result of the joint efforts of several companies and institutions (AT&T, IBM, Lucent, Motorola, etc) which make up the so-called VoiceXML Forum. The language has been designed to ease the creation of multimodal dialogue systems employing audio, ASR, speech synthesis and recording, and mixed-initiative dialogues.

The W3C's Speech Interface Framework defines other standards related to VoiceXML, including:

- SRGS (Speech Recognition Grammar Specification),
- SISR (Semantic Interpretation for Speech Recognition),
- PLS (Pronunciation Lexicon Specification), and
- CCXML (Call Control eXtensible Markup Language).

SALT (Speech Application Language Tags) is also an XML based markup language that is used in HTML and XHTML pages to add ASR to web based applications. Multimodality using this language is possible in different ways: keyboard, speech, keypad, mouse and/or stylus. XHTML+Voice (X+V) is a new technology that combines XHTML and voice-based interfaces on small devices, such as PDAs and tablets. This technology uses web standards such as ECMAScript and JavaScript.

TRINDIKIT is a toolkit for building and experimenting with information states (TRINDIConsortium, 2001). The term information state means, roughly, the information stored internally by a system, in this case a dialogue system. A dialogue move engine (DME) updates the information state on the basis of observed dialogue moves and selects appropriate moves to be performed. Apart from proposing a general system architecture, TRINDIKIT also specifies formats for defining information states, update rules, dialogue moves, and associated algorithms. It further provides a set of tools for experimenting with different formalizations of information-state implementations, rules, and algorithms.

The Galatea Toolkit (Shin-ichi et al., 2003) is an open-source software for the development of anthropomorphic animated multimodal agents. The toolkit comprises four fundamental modules for ASR, speech synthesis, face animation, and dialogue control, which can be used to set up multimodal dialogue systems.

HephaisTK is a toolkit for rapid prototyping of multimodal interfaces that uses SMUIML, a simple mark-up language for describing human-machine multimodal interaction and integration mechanisms (Dumas et al., 2010). This toolkit includes the exploration and assessment of different fusion mechanisms applied to data coming from different human-computer interaction means, such as speech, gesture or ink-based applications.

WebSphere Everyplace Multimodal Environment and WebSphere Multimodal Toolkit, from IBM, enable to integrate graphical and voice interaction in a single application.

The WAMI toolkit (Gruenstein et al., 2008) provides a framework for developing, deploying, and evaluating Web-Accessible Multimodal Interfaces in which users interact using speech, mouse, pen, and/or touch. The toolkit uses modern web-programming techniques, enabling the development of browser-based applications available on a wide array of Internet-connected devices. Several sophisticated multimodal applications have been developed using the toolkit, which can be used by means of desktop, laptop, tablet PCs and mobile devices.

The Multi-Modal Interface Designer (MMID) provides a combination of visual and speech-based interaction. The main way of interaction with the toolkit is vocal: speech-recognition and TTS. However, it includes an additional visual modality.

Festival (Clark et al., 2004) is a C++ general multi-lingual speech synthesis system developed at Centre for Speech Technology Research (CSTR) at the University of Edinburgh. It is distributed under a free software license and offers a number of APIs as well as an environment for development and research on speech synthesis. Supported languages include English, Spanish, Czech, Finnish, Italian, Polish and Russian. An alternative is FreeTTS (Walker et al., 2002), another open source speech synthesis system written entirely in Java. It allows employing markers to specify

when speech generation should not be interrupted, to concatenate speech, and to generate speech using different voices. There are also many commercial systems for TTS like Cepstral, Loquendo TTS and Kalliope.

Xface (Balci, 2005) is an open source toolkit for generating and animating 3D talking heads. The toolkit relies on MPEG-4 Facial Animation Parameters (FAPs) and a keyframe-based rendering which uses the SMIL-Agent scripting language. The toolkit is multi-platform as it can be compiled with any ANSI C++ standard compliant compiler.

The CSLR's Conversational Agent Toolkit (CAT) (Cole et al., 2003) provides a set of modules and tools for research and development of advanced embodied conversational agents. These modules include an audio server, the Sonic speech recognition system, and the Phoenix natural language parser. The CU Animate toolkit (designed for research, development, control and real time rendering of 3D animated characters) is used for the design of the facial animation system.

The Microsoft Agent toolkit (Walsh & Meade, 2003) includes animated characters, TTS engines, and speech recognition software. It is preinstalled in several versions of MS Windows and can be easily embedded in web pages and Office applications with VBScript. Microsoft also provides tools to create new agents, such as the Agent Character Editor.

Maxine (Seron et al., 2006) is an open source engine for embodied conversational agents developed by the University of Zaragoza (Spain). The agents created with this tool can interact with the user by means of text, voice, mouse and keyboard. The agents can gather information from the user and the environment (noise level in the room, position of the user to establish visual contact, image-based estimate of the user's emotional state, etc.), and are able to render emotional states that vary with the relationship that they establish with the user.

3. ADAPTIVE MULTIMODAL INTERFACES

Nowadays, we are surrounded by technology: mobile devices, wearable computing, smart environments and ambient intelligence applications provide new ubiquitous computing capabilities for which multimodal interfaces are in most cases essential (Nihei, 2004; Truong & Dustdar, 2009; Strauss & Minker, 2010). In this section we describe the main principles involved in the development of multimodal interfaces for this pervasive paradigm, highlighting the importance of the adaptivity of the interface.

Adaptivity refers to several aspects in dialogue systems. Novice users and experienced users may want the interface to behave completely differently, for example to have system-initiative instead of mixed-initiative. An example of the benefits of adaptivity in the interaction level can be found in Seneff et al. (2007). The processing of context is essential to achieve this adapted behaviour and also cope with the ambiguities derived from the use of natural language. For instance, context information can be used to resolve anaphoric references, to take into account the current user position as a data to be used by the system, or to decide the strategy to be used by the dialogue management module by taking into account specific user preferences.

3.1. The Role of Context in the Interaction

Although there is not a complete agreement on the definition of *context information*, the most widely accepted is the one proposed by Dey and Abowd (2000): "*Any information that can be used to characterize the situation of an entity (...) relevant to the interaction between a user and an application, including the user and the application themselves*". As can be observed from this definition, any information source can be considered context as long as it provides knowledge relevant

to handle the communication between the user and the system.

Kang et al. (2008) differentiate two types of context: *internal* and *external*. The former describes the user state (e.g., communication context and emotional state), whereas the latter refers to the environment state (e.g., location and temporal context). Most of studies in the literature focus only on external context. However, it is very important to combine both types of context information to provide a personalized and meaningful interaction which takes into account both the users' current location and their preferences (Strauss & Minker, 2010).

In the literature, there are several approaches developing mobile and context aware systems such as platforms, frameworks and multimodal applications for offering context-aware services. These applications include location-based services, e.g., suggesting points or events of interest taking place near the user's current location (Poslad et al., 2001). Other types of context information are device profiles, user preferences, user's activities and interactions, devices, and the network status. These types of context play an important role when context is used to support adaptation in service/task selection (Prezerakos et al., 2007; Truong et al., 2008).

Context information is usually gathered from a wide variety of sources, which produces heterogeneity in terms of quality and persistence. This is why some authors distinguish between static context, which deals with invariant features, and dynamic context, that is able to cope with information that changes (Henricksen et al., 2002). The frequency of such changes is very variable and can deeply influence the way dynamic context is obtained and shared. It is reasonable to obtain largely static context directly from users, and frequently changing context from indirect means such as sensors. To share context some authors have develop tools that make the transfer of contextual information transparent to the interface, which can be placed at a higher level of abstraction. This has been addressed for example by using web services (Keidl & Kemper, 2004).

An important issue to be considered is which language and model is best suited to describe context. A number of methods have been proposed to create these models, from the simple key-value method (in which a variable contains the actual context), to tagged encoding approaches (which uses context profiles to enable modelling and processing context recursively, and to employ efficient context retrieval algorithms), and object oriented models (which have the benefits of encapsulation and reusability). UML, XML, RDF, and OWL-based representations are also widely used because they are considered open and interoperable. In existing context-aware systems, XML is already used widely for modelling and implementing context information.

Regarding context storage techniques, relational databases are frequently employed to store context information in context aware systems out of the web services domain (Naguib et al., 2001; Henricksen et al., 2002). A number of formalisms have been defined to represent the information of the user interaction captured by the sensors in the environment. Many multimodal dialogue systems typically employ the semantic representation based on the concept of dialogue acts (DA) (Stolcke et al., 2000). A DA represents the meaning of the user and system utterances (e.g., question, answer, response, etc.). User's DAs are usually represented by frames (Minsky, 1975). A frame is a structure for representing a concept or situation. Each concept in a domain has usually associated a group of attributes (slots) and values. Recently, machine-learning approaches have been applied to create simple statistical user models trainable on existing human-computer dialogue data, which provide more dynamism to compute internal context (Eckert et al., 1998; Georgila et al., 2005; Schatzmann et al., 2007).

The performance of a dialogue system highly depends on context information. In fact, the result of the interaction can be completely different

depending on the environment conditions (e.g., people speaking near the system, noise generated by other devices) and user skills. In the literature we can find different methodologies to take into account contextual information for adapting the different modules of a dialogue system. In Pargellis et al. (2004) a profile manager is integrated in a spoken dialogue system to code the user preferences about services and modify the dialogue structure by taking them into account. In different dialogue systems, users' skills and preferences are also used to personalize the interaction, set the system initiative, and select specific prompts and modalities (Minker et al., 2004; Seneff et al., 2007).

3.2. The Role of Affect

One of the main research objectives of multimodal systems is to achieve human-like communication between people and machines. This eliminates the need for keyboard and mouse in favour of more intuitive ways of interaction, such as natural language, thus leading to a new paradigm in which technologies can be accessed by non-expert users or handicapped people.

However, multimodal human-computer interaction is still not comparable to human dialogue. One of the reasons for this is that human interaction involves exchanging not only explicit content, but also implicit information about the affective state of the interlocutor. Systems that make use of such information are described as incorporating affective computing as they emulate human emotional intelligence as they are able to recognize, interpret, manage and/or generate emotions. The concept of emotional intelligence was introduced in Salovey and Mayer (1990) to denote "the subset of social intelligence that involves the ability to monitor one's own and others' feelings and emotions, to discriminate among them and to use this information to guide one's thinking and actions". Salovey and Mayer proposed a model that identified four different factors of emotional intelligence: the perception of emotion, the ability reason using emotions, the ability to understand emotion and the ability to manage emotions. According to Salovey and Mayer, the four branches of their model are "arranged from more basic psychological processes to higher, more psychologically integrated processes." For example, the lowest level branch concerns abilities of perceiving and expressing emotion.

To endow multimodal interfaces with affective computing capabilities makes it possible to recognize the user's emotions and adapt the interface functionalities to better accomplish his requirements, thus partly simulating the four branches of human emotional intelligence mentioned. Stern (2003) also provides empirical evidence that if a user encounters a virtual character that seems to be truly emotional, there is also a potential to form emotional relationships with each other. Emotion recognition has been used in Human Computer Interaction (HCI) systems for several purposes. In some application domains it is necessary to recognize the affective state of the user to adapt the systems to it or even change it. For example, in emergency services (Bickmore & Giorgino, 2004) or intelligent tutors (Ai et al., 2006), it is necessary to know the users' emotional state to calm them down, or to encourage them in learning activities. However, there are also some applications in which emotion management is not a central aspect, but contributes to the better functioning of the system as a whole. In these systems emotion management can be used to resolve stages of the dialogue that cause negative emotional states, as well as to avoid them and foster positive ones in future interactions (Burkhardt et al., 2005). Furthermore, emotions are of interest not just for their own sake, but also because they affect the explicit message conveyed during the interaction: they change peoples' voices, facial expressions, gestures, speed of speech, etc. This is usually addressed as "emotional colouring" and can be of great importance for the interpretation of the user input. For example, Wahlster (2006) use emotional

colouring in the context of the SmartKom system to detect sarcasm and thus tackle false positive sentences.

Additionally, the similarity-attraction principle states that users have a better attitude toward agents which exhibit a personality similar to their own. Thus, personality plays a very important role on how users assess multimodal interfaces and their willingness to interact with them. In the same way as humans understand other humans' behaviour and react accordingly to it in terms of the observation of everyday behaviour (Lepri et al., 2009), the personality of the system can be considered as a relatively stable pattern that affects its emotion expression and behaviour and differentiates it from other multimodal interfaces (Xiao et al., 2005).

Emotion recognition for multimodal interfaces is usually treated as a classification problem in which the input is the user last response (voice, facial expressions, body gestures…) and the output is the most probable emotional state. Many different machine learning classifiers have been employed for emotion recognition and frequently the final emotion is decided considering the results of several of these classification algorithms (López-Cózar et al., 2008). Some of the classifiers most widely used are K-nearest neighbours (Lee & Narayanan, 2005), Hidden Markov Models (Pitterman & Pitterman, 2006; Ververidis & Kotropoulos, 2006), Support Vector Machines (Morrison, Wang, & Silva, 2007), Neural Networks (Morrison, Wang, & Silva, 2007; Callejas & López-Cózar, 2008) and Boosting Algorithms (Sebe et al., 2004; Zhu & He, 2008).

Emotion recognition can be carried out with invasive and non invasive methods. Invasive methods are based on physiological measures like breathing rate or conductivity of skin (Picard, 1997). One of the most widespread methods consists in measuring the galvanic skin response (GSR) as there is a relationship between the arousal of emotions and changes in GSR (Lee et al., 2005). Some other methods are EMG, which measures facial muscles (Mahlke, 2006), heart rate or more recently the usage of brain images (Critchley et al., 2005). Non invasive methods are usually based on audio and video. On the one hand, audio emotion recognition can be carried out from the acoustic information or from linguistic information. Speech is deeply affected by emotions: acoustic, contour, tone, voice quality and articulation change with different emotions, a comprehensive study of those changes is presented in Cowie et al. (2001). Language information deals with linguistic changes depending on the emotional state of the user. For this purpose the technique of word emotional salience has gained remarkable attention. This measure represents the frequency of apparition of a word in a given emotional state or category and it is calculated from a corpus of user-system interactions (Lee et al., 2005). On the other hand, video recognition usually pays attention to facial expression, body posture and movements of the hands; a summary of all these features can be found in Picard and Daily (2005). Other authors emphasize that emotions are influenced by cultural and social settings and defend an "interactional approach" (Boehner et al., 2007) to be considered along with physiological, audio or video measures.

Due to its benefits and huge variety of applications, affective computing has become an outstanding research topic in the field of HCI, and numerous important international and interdisciplinary related projects have appeared. Some of the latest are, to mention just a few:

- **MEGA** (Camurri et al., 2004): Its purpose was the modelling and real-time analysis, synthesis, and networked communication of expressive and emotional content in non-verbal interaction (e.g., music or dance) by multi-sensory interfaces, from a multimodal perspective.
- **NECA** (Gebhard et al., 2004): Its purpose was the creation of multi-user and multi-agent virtual spaces populated by affec-

tive conversational agents able to express themselves through synchronised emotional speech and non-verbal expression.

- **VICTEC** (Hall et al., 2005): Its purpose was the development of a toolkit that supports the creation of believable synthetic characters in a virtual environment who establish credible and empathic relations with children.
- **NICE** (Corradini et al., 2005): Its purpose was to foster universal natural interactive access, in particular for children and adolescents, by developing natural, fun and experientially rich communication between humans and embodied historical and literary characters.
- **HUMAINE** (Cowie & Schröder, 2005): Its purpose is to lay the foundations for European development of systems that can register, model and influence human emotional and emotion-related states coordinating efforts to come to a shared understanding of the issues involved.
- **COMPANIONS** (Wilks, 2006): Its purpose is the creation of companions: personalized, conversational interface to the Internet that knows its owner, on a range of platforms, indoor and nomadic, based on integrated high-quality research in multi-modal human-computer interfaces, intelligent agents, and human language technology.

4. EXAMPLES OF MULTIMODAL DIALOGUE SYSTEMS

In this section we describe the interactive multimodal interfaces which we have developed covering some of the issues described in the previous sections.

4.1. The Mayordomo Multimodal Dialogue System

Mayordomo (Ábalos et al., 2010) is a multimodal dialogue system developed in our laboratory which aims to centralize the control of appliances in a home. Specifically, users can employ either spontaneous speech or a traditional GUI interfaces based on keyboard and mouse. The system has been designed to operate in an AmI environment in order to ease the interaction. For example, Mayordomo can find out the room in which the user is at any time through RFID devices. This information is then used to optimize the dialogue with the user, thus ridding him off about providing unnecessary information. Mayordomo also allows parental control of some appliances in order to restrict the interaction with them. For instance, parents can forbid that children watch TV after 10 p.m. The system administrator has privileges to perform special actions, for example, installing and uninstalling appliances and handling the parental control. The system creates a log of all actions carried out within the environment by any user.

To provide spoken interaction, we used Windows Vista Speech Recognition (WVSR). This package includes both the engine for ASR and the engine for TTS. Windows Vista includes two development tools for programmers: SAPI 5.3 (Speech API) and System.Speech (.NET Framework 3.0 namespace). To implement the system we employed System.Speech as it is oriented mainly to programming languages for Microsoft .NET. Each appliance has an associated configuration file that allows the user to control it orally.

Speech understanding is based on what we have noted as an "action". In our application domain, an action consists of four fields of data: room, appliance, attribute, and value. Using these four elements, the system can execute a particular order on an appliance, or provide the information requested by the user. To implement the speech understanding process, we employed a method

that searches in the recognized sentence for the four fields of data in the action concept.

Once the semantic analysis of the sentence is finished, the dialogue manager must decide what will be the answer to be generated by the system. In particular it must determine whether to provide the information requested by the user or perform a specific action on an appliance. To do this it checks if there is information missing in the recognized sentence. If there is no data missing and the user is requesting information, the dialogue manager invokes the module *Provide Information*, which organizes the information to be provided to the user in well-formed sentences. The system uses speech synthesis (TTS) to communicate verbally with the user, employing as input the sentences in text format created by the module for sentence generation.

The GUI interface designed for Mayordomo (Figure 1) includes a status bar, and a text field which displays each system's response. The status bar can be very useful in case users want to interact with the system in noisy environments where understanding the messages generated by TTS might be difficult. The interface also provides a command prompt that allows users to communicate with the system in text format.

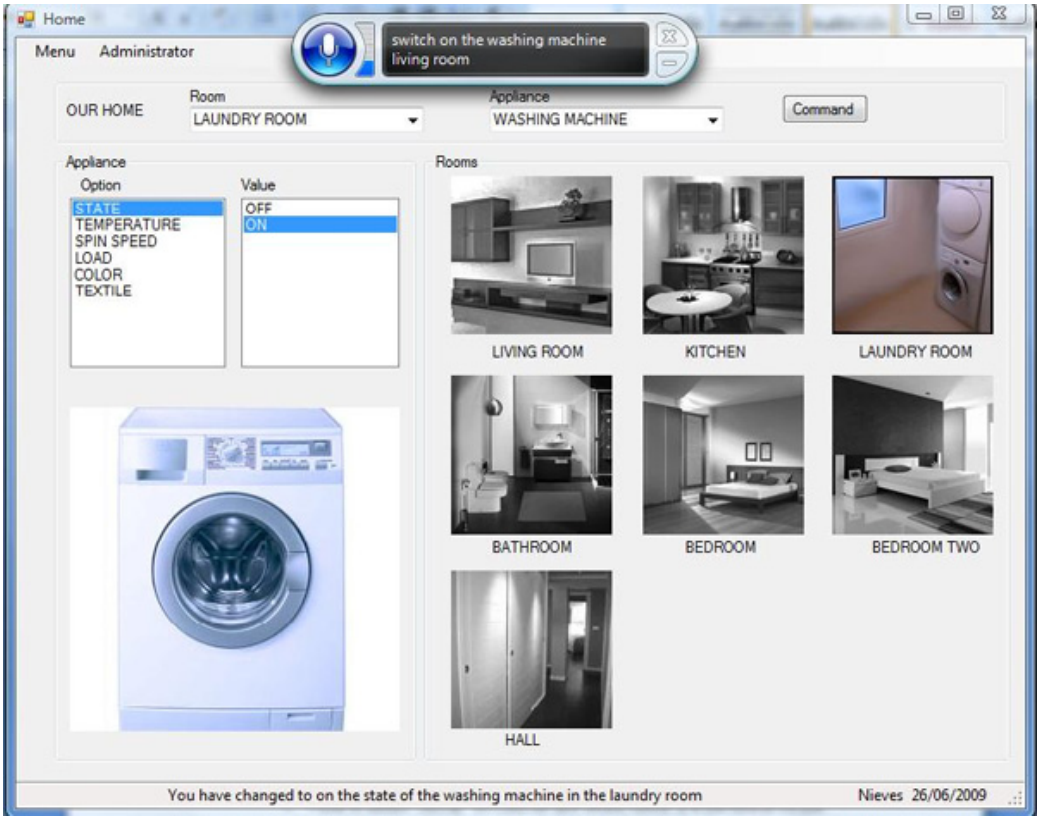
A set of tools have been developed to carry out the evaluation of the system (Ábalos et al., 2011). These tools include a corpus of spoken sentences, two transcribers, an orthographic and a semantic one, and a user simulator. The corpus is a set of audio files, spoken sentences recorded by users, which our system uses in order to perform the actions in the domain to which it is defined. Our current corpus design contains two scenarios: the scenario to switch on any appliances, called scenario 1, and a scenario designed to provide only names of rooms, appliances, attributes, values and actions, called scenario 2. Each of these scenarios contains 75 sentences divided into three sets of 25 sentences. Currently, our corpus contains 1500 sentences recorded: half of them, 750 sentences, are from scenario 1 and the other half are from scenario 2, that is, there have been recorded 10 whole set of sentences for each scenario. The collaboration of fifteen volunteers has been indispensable for recording and obtaining this corpus.

The orthographic automatic transcriber is a module that receives an audio file (.wav) of our corpus and creates a text file (.txt) with the same name, whose content is the orthographic transcription of the audio spoken sentence. The text file created is extremely important in the evaluation of the dialogue system because it contains the sentence designed of the scenario, without any recognition errors, i.e., the correct sentence. The semantic automatic transcriber module, which works as a semantic analyser, receives as input the text file containing a sentence created with the corpus automatic transcriber, and creates a text file with the same name as the input text file.

The user simulator is basically an additional dialogue system which automatically interacts with the dialogue system to assess, representing in this way the behaviour of a real user. The simulator uses the files which contains the corpus in order to create dialogues and interacting with the dialogue system as if it were a user, with goals and answering to questions which are made by the dialogue system. Through this simulation, the dialogue system receives as input spoken sentences previously recorded by users, with the advantage of considering real phenomenon which appear in speech recognition. Several steps are followed to implement the interaction between the user simulator and the dialog system: i) Determine the purpose of the interaction with the dialogue system; ii) Choose the file to be used as a user's turn in the dialogue; iii) Speech recognition; iv) Dialogue management and sentence generation; v) Next dialogue turn or end of interaction.

These tools have been applied to the specific case of Mayordomo dialogue system. In particular, the contributions allow us to assess the dialogue system using two approaches. To perform an overall evaluation of the system, the user simulator tool

Figure 1. GUI interface of the Mayordomo system



has been developed. Meanwhile, to accomplish an evaluation of the individual components of the dialogue system (in our case, speech recognizer, natural language understanding and dialogue manager), the automatic transcribers and a number of secondary tools to calculate statistical measures have been elaborated. With these tools, measures obtained are Word Accuracy, Keyword Accuracy, Sentence Understanding, Sentence Recognition, Task Completion and Implicit Recovery. These measures allow us to obtain experimental results from which to draw conclusions, for instance, which components of the dialogue system must be improved.

In order to perform an overall evaluation of Mayordomo, the user simulator has been applied to our corpus. The simulation result is shown in Table 1. In this case, Microsoft Speech Recognizer 8.0 for Windows (American English) is the recognizer used in the simulation. 1000 dialogues have been generated with sentences recorded of our corpus of which 4751 sentences recorded have been correctly analyzed whereas there have been 263 sentences with recognition errors.

The number of completed dialogues is 274 so task completion is 27.4% which means that in three of every four cases the dialogue does not end satisfactorily. The reason is that sentence recognition rate (SR) and sentence understanding rate (SU) are not quite good because of the recognition errors. In fact, sentences from scenario 2 are names of appliances, rooms, attributes and values. Therefore, these spoken sentences are quite short (only a word or two) and in case there is any recognition error, the speech recognizer finds them difficult to understand. If we want to

Table 1. Simulation results obtained for the Mayordomo system

Analyzed Sentences	4751
Recognition Errors	263
Sentence Recognition Rate (SR)	
Sentence Recognition	28,23%
Correctly Recognized	1341
Sentence Understanding Rate (SU)	
Sentence Understanding Rate	17,7%
Correctly Understood	841
Implicit Recovery (IR)	
Implicit Recovery Rate	0,91%
Not recognized sentences but understood	43
Task Completion (TC)	
Total of dialogues	1000
Finished dialogues	274
Task completion	27,4%
Word Accuracy (WA)	24,86%
Keyword Accuracy (KWA)	24,86%

improve our dialogue system we must take into account these shortages. For example, we could change the speech recognizer engine and evaluate the dialogue system again to compare results and to find the speech recognizer which one is appropriate for our system.

4.2. An Academic Assistant in the SecondLife Virtual World

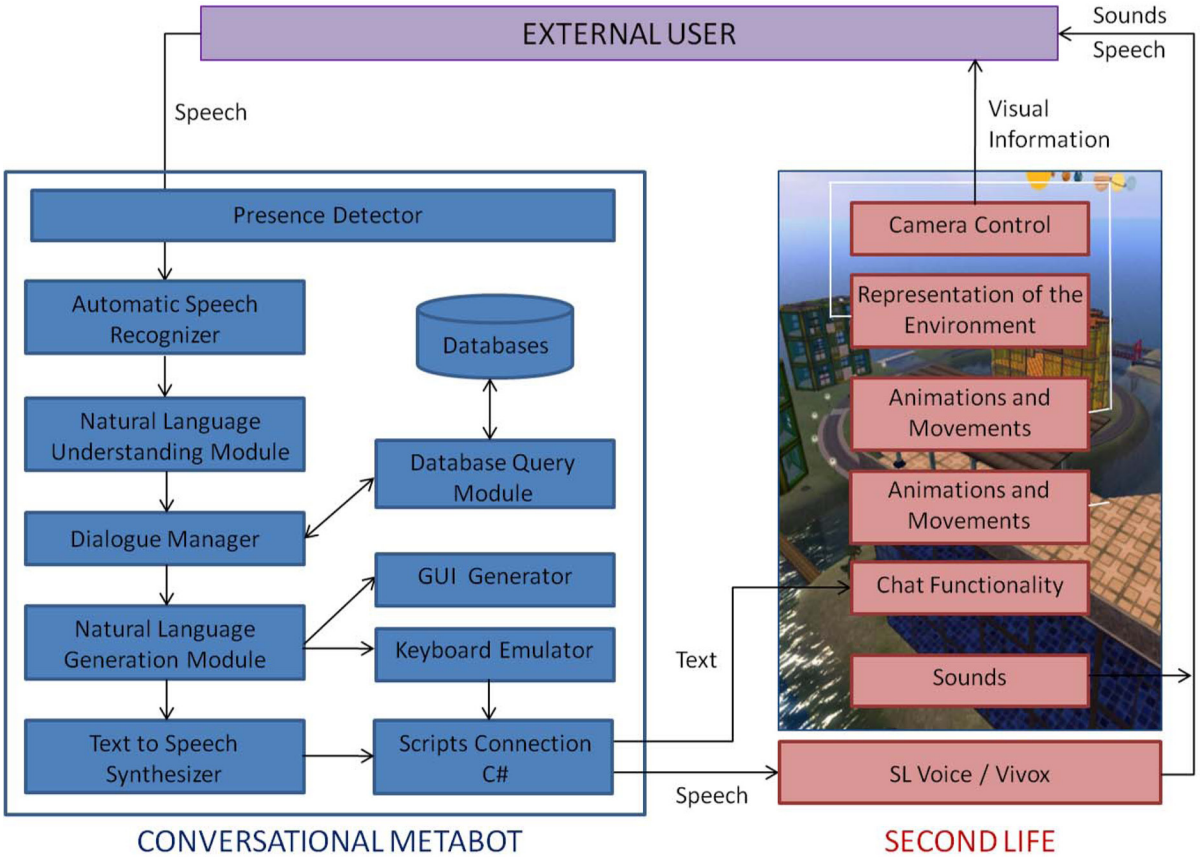
The stunning increase in the amount of time people are spending socializing online is creating new ways of communication and cooperation. With the advances in the so-called Web 2.0, virtual worlds have grown dramatically over the last decade. These worlds or "metaverses" are computer-simulated multimodal environments in which humans, through their avatars cohabit with other users. Traditionally, virtual worlds have had a predefined structure and fixed tasks that the user could carry out. However, social virtual worlds have emerged to emphasize the role of social interaction in these environments, allowing the users to determine their own experiences.

We decided to use Second Life (SL) as a testbed for our research for several reasons. Firstly, because it is one of the most popular social virtual worlds available: its population is nowadays of millions of enthusiastic residents from around the world. Secondly, because it uses a sophisticated physics engine which generates very realistic simulations including collision detection, vehicle dynamics and animation look & feel, thus making the avatars and the environment more credible and similar to the real world. Thirdly, because SL's capacity for customization is extensive and encourages user innovation and participation, which increases the naturalness of the interactions that take place in the virtual world.

We have developed a conversational metabot (Griol et al., 2010) that facilitates academic information (courses, professors, doctoral studies and enrolment) in SL based on the functionalities provided by a previously developed dialogue system (Callejas & López-Cózar, 2008a). Figure 2 shows the architecture developed for the integration of conversational metabot both in the Second Life and OsGrid virtual worlds. The conversational agent that governs the metabot is outside the virtual world, using external servers that provide both data access and speech recognition and synthesis functionalities.

The speech signal provided by the text to speech synthesizer is captured and transmitted to the voice server module in Second Life (SLVoice) using code developed in Visual C#. NET and the SpeechLib library. This module is external to the client program used to display the virtual world and is based on the Vivox technology, which uses the RTP, SIP, OpenAL, TinyXPath, OpenSSL and libcurl protocols to transmit voice data. We also use the utility provided by Second Life lipsynch to synchronize the voice signal with the lip movements of the avatar. In addition, we have integrated a keyboard emulator that allows the transmission of the text transcription generated

Figure 2. Architecture designed for the development of the conversational metabot



by the conversational avatar directly to the chat in Second Life. The system connection with the virtual world is carried out by using the libOpenMetaverse library. This .Net library, based on the Client /Server paradigm, allows accessing and creating three-dimensional virtual worlds, and it is used to communicate with servers that control the virtual world of Second Life.

Speech recognition and synthesis are performed using the Microsoft Speech Application Programming Interface (SAPI), integrated into the Windows Vista operating system. To enable the interaction with the conversational bot in Spanish using the chat in Second Life, we have integrated synthetic voices developed by Loquendo. Using this architecture user's utterances can be easily recognized, the transcription of these utterances can be transcribed in the chat in Second Life, and the result of the user's query can be communicated using both text and speech modalities. To do this, we have integrated modules for the semantic understanding and dialogue management implemented for the original dialogue system, which are based on grammars and VoiceXML files. Figure 3 shows the developed metabot providing information about tutoring hours of a specific professor.

A set of 150 dialogs was acquired with the conversational metabot by means of its interaction with students and professors of our university. Figure 4 shows an example of a dialog extracted from this corpus and translated from Spanish to English. Turns with S refer to system turns, and turns with U refer to user turns.

Figure 3. Conversational metabot developed to interact in virtual worlds (dialogue in the Spanish original version)

We considered the following measures for the evaluation:

1. Dialog success rate (%success). This is the percentage of successfully completed dialogs in which the metabot provides the correct information to each one of the required questions.
2. Average number of turns per dialog (nT).
3. Confirmation rate (%confirm). It was computed as the ratio between the number of explicit confirmations turns (nCT) and the number of turns in the dialog (nCT/nT).
4. Average number of corrected errors per dialog (nCE). This is the average of errors detected and corrected by the dialog manager of the conversational metabot. We have considered only those errors that modify the values of the attributes and that could cause the failure of the dialog.
5. Average number of uncorrected errors per dialog (nNCE). This is the average of errors not corrected by the dialog manager. Again, only errors that modify the values of the attributes are considered.
6. Error correction rate (%ECR). The percentage of corrected errors, computed as nCE/(nCE + nNCE).

The results presented in Table 2 show that the developed conversational can interact correctly with the users in most cases, achieving a success rate of 94%. The dialog success depends on whether the system provides the correct data for every objective defined in the scenario. The analysis of the main problems detected in the acquired dialogs shows that, in some cases, the system did not detect that the user wanted to finish the dialog. A second problem was related to the introduction of data with a high confidence value

Figure 4. Example of a dialog acquired with the interaction of the system with real users

S1: Hi! You can ask me for information about subjects, lecturers, doctoral studies and registrations.
U1: I want to know information about subjects in Computer Science.
S2: Which subject do you want to consult?
U2: Language Processors I.
S3: Here you can consult the information of the subject. Anything else?
U3: Can you inform me about the schedules of Group D?
S4: I give the information for the Group D of Language Processors I: Monday 9:00-11:00, Wednesday 11:00-13:00. Anything else?
U4: Can you tell me the name of the lecturer?
S5: I give the information for the Group D of Language Processors I. The name of the lecturer is David Griol. Anything else?
U5: No, thank you.
S6: Bye!

due to errors generated by the automatic speech recognizer that were not detected by the dialog manager. However, the evaluation confirms a good operation of the approach since the information is correctly provided by the metabot in the majority of cases, as it is also shown in the value of the error correction rate.

In addition, we have already completed a preliminary evaluation of this functionality of the conversational metabot based on questionnaire to assess the students' subjective opinion about the metabot performance. The questionnaire had 10 questions: i) Q1: State on a scale from 1 to 5 your previous knowledge about new technologies for information access.; ii) Q2: How many times have you accessed virtual worlds like Second Life?; iii) Q3: How well did the metabot understand you?; iv) Q4: How well did you understand the messages generated by the metabot?; v) Q5: Was it easy for you to get the requested information?; vi) Q6: Was the interaction rate adequate?; vii) Q7: Was it easy for you to correct the metabot errors?; viii) Q8: Were you sure about what to say to the system at every moment?; ix) Q9: Do you believe the system behaved similarly as a human would do?; x) Q10: In general terms, are you satisfied with the metabot performance?

The possible answers for each one of the questions were the same: Never, Seldom, Sometimes, Usually, and Always. All the answers were

Table 2. Results of the objective evaluation of the conversational metabot

	%success	nT	%confirm	%ECR	nCE	nNCE
Conversational Metabot	94%	11.6	28%	93%	0.89	0.06

Table 3. Results of the subjective evaluation of the conversational metabot (1=worst, 5=best evaluation)

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Average Value	4.6	2.8	3.6	3.8	3.2	3.1	2.7	2.3	2.4	3.3
Maximum Value	5	3	4	5	5	4	3	3	4	4
Minimal Value	4	1	2	3	2	3	2	2	1	3

assigned a numeric value between one and five (in the same order as they appear in the questionnaire). Table 3 shows the average, minimal and maximum values for the subjective evaluation carried out by a total of 15 students from one of the groups in the subject.

From the results of the evaluation, it can be observed that students positively evaluates the facility of obtaining the data required to fulfill the complete set of objectives of the proposed in the exercises defined for the subject, the suitability of the interaction rate during the dialog. The sets of points that they mention to be improved include the correction of system errors and a better clarification of the set of actions expected by the platform at each time.

5. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

The development of multimodal systems is a very active research topic. The design and performance of these systems is very complex, not only because of the complexity of the different technologies involved, but also because of the required interconnection of very heterogeneous components. In this chapter we have provided an overview of the most representative architectures, techniques and toolkits available for the development of such systems. We have paid special attention to context adaptation as a key aspect of these systems, and provided examples of the multimodal dialogue systems developed in our lab that cover the issues discussed.

For future research additional work is needed in several directions to make these systems more usable by a wider range of potential users. For example, the development of emotional conversational agents represents a promising field of research, as emotions play a very important role in the rational decision-making, perception and human-to-human interaction. Also a very interesting trend are multimodal social systems which rely on the fact that in real settings people do not only speak about topics concerned with the task at hand, but also about other topics, especially at the beginning of the conversation, for example, weather conditions, family or current news. Hence, additional efforts must be made by the research community in order to make conversational agents more human-like employing dialogue strategies based on this kind of very genuine human behaviour.

REFERENCES

Ábalos, N., Espejo, G., López-Cózar, R., Callejas, Z., & Griol, D. (2010). A multimodal dialogue system for an ambient intelligent application in home environments. *In Proceedings of the 13th International Conference on Text, Speech and Dialogue* (pp. 491-498).

Ábalos, N., Espejo, G., López-Cózar, R., Callejas, Z., & Griol, D. (2011). A toolkit for the evaluation of spoken dialogue systems in ambient intelligence domains. In *Proceedings of the Second International Workshop on Human-Centric Interfaces for Ambient Intelligence*, Nottingham, UK.

Ai, H., Litman, D., Forbes-Riley, K., Rotaru, M., Tetreault, J., & Purandare, A. (2006). Using systems and user performance features to improve emotion detection in spoken tutoring dialogs. In *Proceedings of the International Conference on Spoken Language Processing*, Pittsburgh, PA (pp. 797-800).

Bailly, G., Raidt, S., & Elisei, F. (2010). Gaze, conversational agents and face-to-face communication. *Speech Communication*, *52*(6), 598–612. doi:10.1016/j.specom.2010.02.015

Balci, K. (2005). XfaceEd: Authoring tool for embodied conversational agents. In *Proceedings of the International Conference on Multimodal Interfaces* (pp. 208-213).

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russel, M., & Wong, M. (2004). Towards multilingual speech recognition using data driven source/target acoustical units association. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Montreal, QC, Canada (pp. 521-524).

Benesty, J., Sondhi, M. M., & Huang, Y. (2008). *Springer handbook of speech processing*. New York, NY: Springer. doi:10.1007/978-3-540-49127-9

Beveridge, M., & Fox, J. (2006). Automatic generation of spoken dialogue from medical plans and ontologies. *Biomedical Informatics*, *39*(5), 482–499. doi:10.1016/j.jbi.2005.12.008

Bickmore, T., & Giorgino, T. (2004). Some novel aspects of health communication from a dialogue systems perspective. In *Proceedings of the AAAI Fall Symposium on Dialogue Systems for Health Communication*, Washington, DC (pp. 275-291).

Bird, S., Klein, E., Loper, E., & Baldridge, J. (2008). Multidisciplinary instruction with the Natural Language Toolkit. In *Proceedings of the Third ACL Workshop on Issues in Teaching Computational Linguistics* (pp. 62-70).

Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human-Computer Studies*, *65*(4), 275–291. doi:10.1016/j.ijhcs.2006.11.016

Bos, J., Klein, E., & Oka, T. (2003). Meaningful conversation with a mobile robot. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics* (pp. 71-74).

Burkhardt, F., van Ballegooy, M., Englert, R., & Huber, R. (2005). An emotion-aware voice portal. In *Proceedings of the Electronic Speech Signal Processing Conference*, Prague, Czech Republic (pp. 123-131).

Callejas, Z., & López-Cózar, R. (2008a). Relations between de-facto criteria in the evaluation of a spoken dialogue system. *Speech Communication*, *50*(8-9), 646–665. doi:10.1016/j.specom.2008.04.004

Callejas, Z., & López-Cózar, R. (2008b). Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication*, *50*(5), 416–433. doi:10.1016/j.specom.2008.01.001

Camurri, A., Mazzarino, B., & Volpe, G. (2004). Expressive interfaces. *Cognition Technology and Work*, *6*(1), 15–22. doi:10.1007/s10111-003-0138-7

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhálmsson, H., & Yan, H. (1999). Embodiment in conversational interfaces: Rea. In *Proceedings of the Conference on Computer-Human Interaction* (pp. 520-527).

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. F. (2000). *Embodied conversational agents*. Cambridge, MA: MIT Press.

Catizone, R., Setzer, A., & Wilks, Y. (2003). Multimodal dialogue management in the COMIC Project. *Proceedings of the European Chapter of the Association for Computational Linguistics Workshop on Dialogue Systems*: *Interaction, Adaptation, and Styles of Management*, Budapest, Hungary (pp. 25-34).

Chu, S.-W., O'Neill, I., Hanna, P., & McTear, M. (2005). An approach to multistrategy dialogue management. In *Proceedings of the Interspeech/Eurospeech Conference*, Lisbon, Portugal (pp. 865-868).

Clark, R., Richmond, K., & King, S. (2004). Festival 2 - build your own general purpose unit selection speech synthesizer. In *Proceedings of the 5th ISCA Workshop on Speech Synthesis* (pp. 173-178).

Cole, R., Mariani, J., Uszkoreit, H., Varile, G. B., Zaenen, A., Zampolli, A., & Zue, V. (Eds.). (1997). *Survey of the state of the art in human language technology*. Cambridge, UK: Cambridge University Press.

Cole, R., Van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., & Movellan, J. …Wade-stein, D. (2003). Perceptive animated interfaces: first steps toward a new paradigm for human-computer interaction. *Proceedings of the IEEE, 91*(9), 1391-1405.

Corradini, A., Mehta, M., Bernsen, N. O., & Charfuelán, M. (2005). Animating an interactive conversational character for an educational game system. In *Proceedings of the International Conference on Intelligent User Interfaces*, San Diego, CA (pp. 183-190).

Corradini, A., & Samuelsson, C. (2008). A generic spoken dialogue manager applied to an interactive 2D game. In E. André, L. Dybkjær, W. Minker, H. Neumann, R. Pieraccini, & M. Weber (Eds.), *Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal Dialogue Systems* (LNCS 5078, pp. 2-13).

Cowie, R., & Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1-2), 5–32. doi:10.1016/S0167-6393(02)00071-7

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80. doi:10.1109/79.911197

Cowie, R., & Schröder, M. (2005). Piecing together the emotion jigsaw. In S. Bengio & H. Bourlard (Eds.), *Proceedings of the First International Conference on Machine Learning for Multimodal Interaction* (LNCS 3361, pp. 305-317).

Critchley, H. D., Rotshtein, P., Nagai, Y., O'Doherty, J., Mathias, C. J., & Dolana, R. J. (2005). Activity in the human brain predicting differential heart rate responses to emotional facial expressions. *NeuroImage*, *24*, 751–762. doi:10.1016/j.neuroimage.2004.10.013

Cuayáhuitl, H., Renals, S., Lemon, O., & Shimodaira, H. (2006). Reinforcement learning of dialogue strategies with hierarchical abstract machines. In *Proceedings of the IEEE/ACL Spoken Language Technology Workshop*, Palm Beach, Aruba (pp. 182-186).

Dey, A., & Abowd, G. (2000). Towards a better understanding of context and context-awareness. In *Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing* (pp. 304-307).

Dumas, B., Lalanne, D., & Ingold, R. (2009). Description languages for multimodal interaction: a set of guidelines and its illustration with SMUIML. *Journal on Multimodal User Interfaces*, *3*, 237–247. doi:10.1007/s12193-010-0043-3

Eckert, W., Levin, E., & Pieraccini, R. (1998). *Automatic evaluation of spoken dialogue systems* (Tech. Rep. No. TR98.9.1). Florham Park, NJ: ATT Labs Research.

Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication*, *50*(8-9), 630–645. doi:10.1016/j.specom.2008.04.002

Elhadad, M., & Robin, J. (1996). An overview of SURGE: A reusable comprehensive syntactic realization component. In *Proceedings of the Eighth International Natural Language Generation Workshop* (pp. 1-4).

Endrass, B., Rehm, M., & André, E. (2011). Planning small talk behavior with cultural influences for multiagent systems. *Computer Speech & Language*, *25*(2), 158–174. doi:10.1016/j.csl.2010.04.001

Faure, C., & Julia, L. (1993). Interaction homme-machine par la parole et le geste pour l'édition de documents. In *Proceedings of the International Conference on Real and Virtual Worlds* (pp. 171-180).

Flecha-García, M. L. (2010). Eyebrow raises in dialogue and their relation to discourse structure, utterance function and pitch accents in English. *Speech Communication*, *52*(6), 542–554. doi:10.1016/j.specom.2009.12.003

Forbes-Riley, K., & Litman, D. (2011). Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Computer Speech & Language*, *25*(1), 105–126. doi:10.1016/j.csl.2009.12.002

Forbes-Riley, K. M., & Litman, D. (2004). Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, New York, NY (pp. 264-271).

Fraser, N., & Gilbert, G. (1991). Simulating speech systems. *Computer Speech & Language*, *5*, 81–99. doi:10.1016/0885-2308(91)90019-M

Gebhard, P., Klesen, M., & Rist, T. (2004). Coloring multi-character conversations through the expression of emotions. In *Proceedings of the Tutorial and Research Workshop on Affective Dialogue Systems*, Kloster Irsee, Germany (pp. 128-141).

Georgila, K., Henderson, J., & Lemon, O. (2005). Learning user simulations for information state update dialogue systems. In *Proceedings of the Eurospeech Conference* (pp. 893-896).

Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., & Sakai, S. (1995). Multilingual spoken-language understanding in the MIT Voyager system. *Speech Communication*, *17*(1-2), 1–18. doi:10.1016/0167-6393(95)00008-C

Griol, D., Hurtado, L. F., Segarra, E., & Sanchis, E. (2008). A statistical approach to spoken dialog systems design and evaluation. *Speech Communication*, *50*(8-9), 666–682. doi:10.1016/j.specom.2008.04.001

Griol, D., Rojo, E., Arroyo, Á., Patricio, M. A., & Molina, J. M. (2010). A conversational academic assistant for the interaction in virtual worlds. *Advances in Soft Computing*, *79*, 283–290. doi:10.1007/978-3-642-14883-5_37

Gruenstein, A., McGraw, I., & Badr, I. (2008). The WAMI toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces. In *Proceedings of the International Conference on Multimodal Interfaces.*

Hall, L., Woods, S., Aylett, R., Paiva, A., & Newall, L. (2005). Achieving empathic engagement through affective interaction with synthetic characters. In J. Tao, T. Tan, & R. W. Picard (Eds.), *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, Beijing, China (LNCS 3784, pp. 731-738).

Haseel, L., & Hagen, E. (2005). Adaptation of an automotive dialogue system to users' expertise. In *Proceedings of the Interspeech/Eurospeech Conference*, Lisbon, Portugal (pp. 222-226).

Heim, J., Nilsson, E. G., & Skjetne, J. H. (2007). User profiles for adapting speech support in the opera Web browser to disabled users. In C. Stephanidis & M. Pieper (Eds.), *Proceedings of the 9th ECRIM Workshop on Universal Access in Ambient Intelligence Environments* (LNCS, 4397, pp. 154-172).

Henricksen, K., Indulska, J., & Rakotonirainy, A. (2002). Modeling context information in pervasive computing systems. In *Proceedings of the 1st International Conference on Pervasive Computing* (pp. 167-180).

Huang, C., Xu, P., Zhang, X., Zhao, S., Huang, T., & Xu, B. (1999). LODESTAR: A Mandarin spoken dialogue system for travel information retrieval. In *Proceedings of the Conference Eurospeech* (pp. 1159-1162).

Huang, H., Cerekovic, A., Pandzic, I., Nakano, Y., & Nishida, T. (2007). A script driven multimodal embodied conversational agent based on a generic framework. In *Proceedings of the 7th International Conference on Intelligent Virtual Agents* (pp. 381-382).

Ibrahim, A., & Johansson, P. (2002). Multimodal dialogue systems for interactive TV applications. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces* (pp. 117-122).

Jokinen, K. (2003). Natural interaction in spoken dialogue systems. In *Proceedings of the Workshop on Ontologies and Multilinguality in User Interfaces*, Crete, Greece (pp. 730-734).

Kang, H., Suh, E., & Yoo, K. (2008). Packet-based context aware system to determine information system user's context. *Expert Systems with Applications*, *35*, 286–300. doi:10.1016/j.eswa.2007.06.033

Keidl, M., & Kemper, A. (2004). A framework for context-aware adaptable Web services. In E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm, & E. Ferrari (Eds.), *Proceedings of the 9th International Conference on Advances in Database Technology* (LNCS 2992, pp. 826-829).

Langner, B., & Black, A. (2005). Using speech in noise to improve understandability for elderly listeners. In *Proceedings of the Conference on Automatic Speech Recognition and Understanding*, San Juan, Puerto Rico (pp. 392-396).

Lee, C., Yoo, S. K., Park, Y. J., Kim, N. H., Jeong, K. S., & Lee, B. C. (2005). Using neural network to recognize human emotions from heart rate variability and skin resistance. In *Proceedings of the Annual International Conference on Engineering in Medicine and Biology Society*, Shanghai, China (pp. 5523-5525).

Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, *13*(2), 293–303. doi:10.1109/TSA.2004.838534

Lee, K., Hon, H., & Reddy, R. (1990). An overview of the SPHINX speech recognition system. In Waibel, A., & Lee, K.-F. (Eds.), *Readings in speech recognition* (pp. 600–610). San Francisco, CA: Morgan Kaufmann. doi:10.1109/29.45616

Lemon, O., Georgila, K., & Henderson, J. (2006). Evaluating effectiveness and portability of reinforcement learned dialogue strategies with real users: the TALK TownInfo evaluation. In *Proceedings of the IEEE-ACL Spoken Language Technologies Conference*, Palm Beach, Aruba (pp. 178-181).

Lepri, B., Mana, N., Cappelletti, A., Pianesi, F., & Zancanaro, M. (2009). Modeling the personality of participants during group interactions. In G.-J. Houben, G. McCalla, F. Pianesi, & M. Zancanaro (Eds.), *Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization* (LNCS 5535, pp. 114-125).

Leßmann, N., & Wachsmuth, I. (2003). A cognitively motivated architecture for an anthropomorphic artificial communicator. In *Proceedings of the International Conference on Computing and Mission* (pp. 277-278).

Li, L., Cao, F., Chou, W., & Liu, F. (2006). XM-flow: An extensible micro-flow for multimodal interaction. In *Proceedings of the 8th Workshop on Multimedia Signal Processing* (pp. 497-500).

Li, L., Li, L., Chou, W., & Liu, F. (2007). R-Flow: An extensible XML based multimodal dialogue system architecture. In *Proceedings of the 9th Workshop on Multimedia Signal Processing* (pp. 86-89).

Litman, D. J., & Pan, S. (2002). Designing and evaluating an adaptive spoken dialogue system. *User Modeling and User-Adapted Interaction*, *12*, 111–137. doi:10.1023/A:1015036910358

López-Cózar, R., & Araki, M. (2005). *Spoken, multilingual and multimodal dialogue systems. development and assessment*. New York, NY: John Wiley & Sons.

López-Cózar, R., Callejas, Z., Kroul, M., Nouza, J., & Silovský, J. (2008). Two-level fusion to improve emotion classification in spoken dialogue systems. In P. Sojka, A. Horák, I. Kopecek, & K. Pala (Eds.), *Proceedings of the 11th International Conference on Text, Speech and Dialogue* (LNCS 5246, pp. 617-624).

Mahlke, S. (2006). Emotions and EMG measures of facial muscles in interactive contexts. In *Proceedings of the Conference on Human Factors in Computing Systems*, Montreal, QC, Canada.

Maragoudakis, M. (2007). MeteoBayes: Effective plan recognition in a weather dialogue system. *IEEE Intelligent Systems*, *22*(1), 66–77. doi:10.1109/MIS.2007.14

Martinovski, B., & Traum, D. (2003). Breakdown in human-machine interaction: the error is the clue. In *Proceedings of the ISCA Tutorial and Research Workshop on Error Handling in Dialogue Systems*, Chateau d'Oex, Vaud, Switzerland (pp. 11-16).

McGlashan, S., Burnett, D. C., Carter, J., Danielsen, P., Ferrans, J., & Hunt, A. …Tryphonas, S. (2004). *Voice Extensible Markup Language (VoiceXML)*. Retrieved from http://www.w3.org/TR/voicexml21/

McTear, M. F. (1998). Modelling spoken dialogues with state transition diagrams: experiences with the CSLU toolkit. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 1223-1226).

McTear, M. F. (2004). *Spoken dialogue technology*. New York, NY: Springer. doi:10.1007/978-0-85729-414-2

Minker, W. (1998). Stochastic versus rule-based speech understanding for information retrieval. *Speech Communication*, *25*(4), 223–247. doi:10.1016/S0167-6393(98)00038-7

Minker, W., Haiber, U., Heisterkamp, P., & Scheible, S. (2004). The Seneca spoken language dialogue system. *Speech Communication*, *43*(1-2), 89–102. doi:10.1016/j.specom.2004.01.005

Minsky, M. (1975). A framework for representing knowledge. In Winston, P. H. (Ed.), *The psychology of computer vision* (pp. 211–277). New York, NY: McGraw-Hill.

Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics*, *23*(2), 269–311.

Moran, D. B., Cheyer, A. J., Julia, L. E., Martin, D. L., & Park, S. (1997). Multimodal user interface in the open agent architecture. In *Proceedings of the 2nd International Conference on Intelligent User Interfaces* (pp. 61-68).

Morrison, D., Wang, R., & Silva, L. C. D. (2007). Ensemble methods for spoken emotion recognition in call-centers. *Speech Communication*, *49*(2), 98–112. doi:10.1016/j.specom.2006.11.004

Müller, C., & Runge, F. (1993). Dialogue design principles - key for usability of voice processing. In *Proceedings of the Eurospeech Conference* (pp. 943-946).

Naguib, H., Coulouris, G., & Mitchell, S. (2001). Middleware support for context-aware multimedia applications. In *Proceedings of the 3rd International Working Conference on New Developments in Distributed Applications and Interoperable Systems* (pp. 9-22).

Nielsen, P. B., & Baekgaard, A. (1992). Experience with a dialogue description formalism for realistic applications. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 719-722).

Nihei, K. (2004). Context sharing platform. *NEC Journal of Advanced Technology*, *1*(3), 200–204.

Oh, A., & Rudnicky, A. (2000). Stochastic language generation for spoken dialog systems. In *Proceedings of the ANLP North American Chapter of the Association for Computational Linguistics Workshop on Conversational Systems* (pp. 27-32).

Paek, T., & Horvitz, E. (2004). Optimizing automated call routing by integrating spoken dialogue models with queuing models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies* (pp. 41-48).

Pargellis, A., Kuo, H., & Lee, C. (2004). An automatic dialogue generation platform for personalized dialogue applications. *Speech Communication*, *42*, 329–351. doi:10.1016/j.specom.2003.10.003

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Picard, R. W., & Daily, S. B. (2005). Evaluating affective interactions: Alternatives to asking what users feel. In *Proceedings of the CHI Workshop on Evaluating Affective Interfaces-Innovative Approaches*, Portland, OR.

Pitterman, J., & Pitterman, A. (2006). Integrating emotion recognition into an adaptive spoken language dialogue system. In *Proceedings of the 2nd IEEE International Conference on Intelligent Environments* (pp. 213-219).

Poslad, S., Laamanen, H., Malaka, R., Nick, A., Buckle, P., & Zipf, A. (2001). Crumpet: Creation of user-friendly mobile services personalized for tourism. In *Proceedings of the 2nd International Conference on 3G Mobile* (pp. 28-32).

Prendinger, H., Mayer, S., Mori, J., & Ishizuka, M. (2003). Persona effect revisited. using biosignals to measure and reflect the impact of character-based interfaces. In *Proceedings of the Intelligent Virtual Agents*, Kloster Irsee, Germany (pp. 283-291).

Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Radford, L. (2003). Gestures, speech, and the sprouting of signs: A semiotic-cultural approach to students' types of generalization. *Mathematical Thinking and Learning*, *5*(1), 37–70. doi:10.1207/S15327833MTL0501_02

Raux, A., & Eskenazi, M. (2007). A multi-layer architecture for semi-synchronous event-driven dialogue management. In *Proceedings of the International Conference on Automatic Speech Recognition and Understanding* (pp. 514-519).

Raux, A., Langner, B., Black, A. W., & Eskenazi, M. (2003). LET'S GO: Improving spoken dialog systems for the elderly and non-natives. In *Proceedings of the Eurospeech Conference*, Geneva, Switzerland (pp. 753-756).

Reiter, E. (1995). NLG vs. templates. In *Proceedings of the Fifth European Workshop in Natural Language Generation* (pp. 95-105).

Rosenfeld, R. (1995). The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation. In *Proceedings of the ARPA Spoken Language Systems Technology Workshop*.

Salber, D., & Coutaz, J. (1993). Applying the wizard of oz technique to the study of multimodal systems. In *Proceedings of the Selected papers from the Third International Conference on Human-Computer Interaction* (pp. 219-230).

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, Cognition and Personality*, *9*, 185–211.

Schatzmann, J., Thomson, B., Weilhammer, K., Ye, H., & Young, S. (2007). Agenda-based user simulation for bootstrapping a POMDP dialogue system. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies* (pp. 149-152).

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, *53*, 1062–1087. doi:10.1016/j.specom.2011.01.011

Schultz, T., & Kirchhoff, K. (2006). *Multilingual speech processing*. Amsterdam, The Netherlands: Elsevier.

Sebe, N., Sun, Y., Bakker, E., Lew, M. S., Cohen, I., & Huang, T. S. (2004). Towards authentic emotion recognition. In *Proceedings of the IEEE Conference on Systems, Man and Cybernetics* (pp. 623-628).

Seneff, S. (1989). TINA: A probabilistic syntactic parser for speech understanding systems. In *Proceedings of ACL Workshop on Speech and Natural Language* (pp. 168-178).

Seneff, S., Adler, M., Glass, J., Sherry, B., Hazen, T., Wang, C., & Wu, T. (2007). Exploiting context information in spoken dialogue interaction with mobile devices. In *Proceedings of the International Workshop on Improved Mobile User Experience* (pp. 1-11).

Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., & Zue, V. (1998). Galaxy-II: A reference architecture for conversational system development. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 931-934).

Seron, F., Baldassarri, S., & Cerezo, E. (2006). MaxinePPT: Using 3D virtual characters for natural interaction. In *Proceedings of the 2nd International Workshop on Ubiquitous Computing and Ambient Intelligence* (pp. 241-250).

Shin-ichi, K., Shimodaira, H., Nitta, T., Nishimoto, T., Nakamura, S., & Itou, K. …Sagayama, S. (2003). Galatea: Open-source software for developing anthropomorphic spoken dialog agents. In H. Prendinger & M. Ishizuka (Eds.), *Life-like characters: Tools, affective functions, and applications* (pp. 187-212). Berlin, Germany: Springer-Verlag.

Stent, A., Dowding, J., Gawron, J. M., Bratt, E., & Moore, R. (1999). The CommandTalk spoken dialogue system. In *Proceedings of the Association for Computational Linguistics* (pp. 183-190).

Stern, A. (2003). Creating emotional relationships with virtual characters. In Trappl, R., Petta, P., & Payr, S. (Eds.), *Emotions in humans and artifacts* (pp. 333–362). Cambridge, MA: MIT Press.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Ess-Dykema, C. V., & Ries, K. (2000). Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26*(3), 339–373. doi:10.1162/089120100561737

Strauss, P., & Minker, W. (2010). *Proactive spoken dialogue interaction in multi-party environments*. New York, NY: Springer. doi:10.1007/978-1-4419-5992-8

TRINDIConsortium. (2001). *Task Oriented Instructional Dialogue Book Draft*. Retrieved from http://www.ling.gu.se/projekt/trindi/book.ps

Truong, H. L., & Dustdar, S. (2009). A survey on context-aware web service systems. *International Journal of Web Information Systems*, *5*(1), 5–31. doi:10.1108/17440080910947295

Truong, H. L., Dustdar, S., Baggio, D., Corlosquet, S., Dorn, C., Giuliani, G., & Gombotz, R. (2008). inContext: A pervasive and collaborative working environment for emerging team forms. In *Proceedings of the International Symposium on Applications and the Internet* (pp. 118-125).

Van de Burgt, S. P., Andernach, T., Kloosterman, H., Bos, R., & Nijholt, A. (1996). Building dialogue systems that sell. In *Proceedings of the NLP and Industrial Applications Conference* (pp. 41-46).

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: resources, features and methods. *Speech Communication*, *48*, 1162–1181. doi:10.1016/j.specom.2006.04.003

Wahlster, W. (2001). SmartKom: Multimodal dialogues with mobile Web users. In *Proceedings of the International Cyber Assist Symposium* (pp. 33-40).

Wahlster, W. (2003) Towards symmetric multimodality: Fusion and fission of speech, gesture, and facial expression. In *Proceedings of the 26th German Conference on Artificial Intelligence* (pp. 1-18).

Wahlster, W. (Ed.). (2006). *SmartKom: Foundations of multimodal dialogue systems*. New York, NY: Springer. doi:10.1007/3-540-36678-4

Walker, W., Lamere, P., & Kwok, P. (2002). *FreeTTS: A performance case study*. Santa Clara, CA: Sun Microsystems.

Walsh, P., & Meade, J. (2003). Speech enabled e-learning for adult literacy tutoring. In *Proceedings of the International Conference on Advanced Learning Technologies* (pp. 17-21).

Ward, W., & Issar, S. (1994). Recent improvements in the CMU spoken language understanding system. In *Proceedings of the ACL Workshop on Human Language Technology* (pp. 213-216).

Wei, X., & Rudnicky, A. (2000). Task-based dialogue management using an agenda. In *Proceedings of the ANLP/NAACL Workshop on Conversational Systems* (pp. 42-47).

Wilks, Y. (2006). *Artificial companions as a new kind of interface to the future internet* (Tech. Rep. No. 13). Oxford, UK: Oxford Internet Institute.

Williams, J., & Young, S. (2007). Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, *21*(2), 393–422. doi:10.1016/j.csl.2006.06.008

Xiao, H., Reid, D., Marriott, A., & Gulland, E. K. (2005). An adaptive personality model for ECAs. In J. Tao, T. Tan, & R. W. Picard (Eds.), *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction* (LNCS 3784, pp. 637-645).

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2000). *The HTK book*. Redmond, WA: Microsoft Corporation.

Zhu, Z., & He, K. (2008). A novel approach of emotion recognition based on selective ensemble. In *Proceedings of the 3rd International Conference on Intelligent Systems and Knowledge Engineering* (pp. 695-698).

Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T., & Hetherington, L. (2000). JUPITER: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, *8*(1), 85–96. doi:10.1109/89.817460

ADDITIONAL READING

Bernsen, N. O., & Dybkjaer, L. (2010). *Multimodal usability*. New York, NY: Springer.

Bezold, M., & Minker, W. (2011). *Adaptive multimodal interactive systems*. New York, NY: Springer. doi:10.1007/978-1-4419-9710-4

Grifoni, P. (Ed.). (2009). *Multimodal human-computer interaction and pervasive services*. Hershey, PA: IGI Global. doi:10.4018/978-1-60566-386-9

Jokinen, K. (2009). *Constructive dialogue modelling: speech interaction and rational agents*. New York, NY: John Wiley & Sons.

Kurkovsky, S. (Ed.). (2009). *Multimodality in mobile computing and mobile devices: methods for adaptable usability*. Hershey, PA: IGI Global. doi:10.4018/978-1-60566-978-6

Macías, J. A., Granollers, A., & Latorre, P. M. (Eds.). (2009). *New trends on human-computer interaction*. New York, NY: Springer. doi:10.1007/978-1-84882-352-5

Maragos, P., Potamianos, A., & Graos, P. (Eds.). (2010). *Multimodal processing and interaction: Audio, video, text*. New York, NY: Springer.

Tzovaras, D. (Ed.). (2010). *Multimodal user interfaces: from signals to interaction*. New York, NY: Springer.

KEY TERMS AND DEFINITIONS

Affective Computing: Interdisciplinary field of study concerned with developing computational systems which are able to understand, recognize, interpret, synthesize, predict and/or respond to human emotions.

Automatic Speech Recognition (ASR): Technique to determine the word sequence in a speech signal. To do this, this technology first detects basic units in the signal, e.g., phonemes, which are then combined to determine words.

Context Information: Any information that can be used to characterize the situation of an entity relevant to the interaction between a user and an application, including the user and the application themselves (Dey & Abowd, 2000).

Dialogue Management (DM): Iimplementation of the "intelligent" behaviour of the conversational system. It receives some sort of internal representation obtained from the user input and decides the next action the system must carry out.

Fission of Multimodal Information: Opposite to the *fusion* operation, chooses the output to be produced through each output modality and coordinates the output across the modalities in order to generate a system response appropriately for the user.

Fusion of Multimodal Information: Operation that combines the information chunks provided by the diverse input modules of the conversational agent in order to obtain a better understanding of the intention of the user.

Natural Language Generation (NLG): Creation of messages in text mode, grammatical and semantically correct, which will be either displayed on screen or converted into speech by means of text-to-speech synthesis.

Second Life: A three dimensional virtual world developed by Linden Lab in 2003 and accessible via the Internet.

Speech Synthesis: Artificial generation of human-like speech. A particular kind of speech synthesis technique is called Text-To-Speech synthesis (TTS), the goal of which is to transform into speech of any input sentence in text format.

Spoken Language Understanding (SLU): Technique to obtain the semantic content of the sequence of words provided by the ASR module. It must face a variety of phenomena, for example, ellipsis, anaphora and ungrammatical structures typical of spontaneous speech.

Virtual World/Environment: Synthetic environment which resembles real world or can be perceived as a real world by their users.

VoiceXML: Standard XML-based language to access web applications by means of speech.

Wizard of Oz (WOz): Technique that uses a human called *Wizard* to play the role of the computer in a human-computer interaction. The users are made to believe that they interact with a computer but actually they interact with the Wizard.

XHTML+Voice (X+V): XML-based language that combines traditional web access using XHTML and speech-based access to web pages using VoiceXML.