Review of spoken dialogue systems

Article · February 2015

DOI: 10.3989/loquens.2014.012

CITATIONS
10

READS
448

4 authors, including:

Zoraida Callejas
University of Granada
128 PUBLICATIONS **718** CITATIONS
SEE PROFILE

Jose Quesada
Universidad de Sevilla
47 PUBLICATIONS **990** CITATIONS
SEE PROFILE

Loquens 1(2)
July 2014, e012
eISSN 2386-2637
doi: http://dx.doi.org/10.3989/loquens.2014.012

Review of spoken dialogue systems

Ramón López-Cózar[1], Zoraida Callejas[1], David Griol[2] and José F. Quesada[3]

[1] Universidad de Granada, [2] Universidad Carlos III de Madrid, [3] Universidad de Sevilla
e-mail: rlopezc@ugr.es, zoraida@ugr.es, dgriol@inf.uc3m.es, Jose-Francisco.Quesada@cs.us.es

Submitted: 18 de noviembre de 2013. Accepted: 30 de junio de 2014. Available on line: 11 de marzo de 2015

Citation / Cómo citar este artículo: López-Cózar, R., Callejas, Z., Griol, D., Quesada, J. F. (2014). Review of spoken dialogue systems. *Loquens, 1*(2), e012. doi: http://dx.doi.org/10.3989/loquens.2014.012

ABSTRACT: Spoken dialogue systems are computer programs developed to interact with users employing speech in order to provide them with specific automated services. The interaction is carried out by means of dialogue turns, which in many studies available in the literature, researchers aim to make as similar as possible to those between humans in terms of naturalness, intelligence and affective content.

In this paper we describe the fundaments of these systems including the main technologies employed for their development. We also present an evolution of this technology and discuss some current applications. Moreover, we discuss development paradigms, including scripting languages and the development of conversational interfaces for mobile apps.

The correct modelling of the user is a key aspect of this technology. This is why we also describe affective, personality and contextual models. Finally, we address some current research trends in terms of verbal communication, multimodal interaction and dialogue management.

Keywords: dialogue; language understanding; dialogue management; natural language generation; speech synthesis.

RESUMEN: *Sistemas de diálogo: una revisión.*– Los sistemas de diálogo son programas de ordenador desarrollados para interaccionar con los usuarios mediante habla, con la finalidad de proporcionarles servicios automatizados. La interacción se lleva a cabo mediante turnos de un tipo de diálogo que, en muchos estudios existentes en la literatura, los investigadores intentan que se parezca lo más posible al diálogo real que se lleva a cabo entre las personas en lo que se refiere a naturalidad, inteligencia y contenido afectivo.

En este artículo describimos los fundamentos de esta tecnología, incluyendo las tecnologías básicas que se utilizan para implementar este tipo de sistemas. También presentamos una evolución de la tecnología y comentamos algunas aplicaciones actuales. Asimismo, describimos paradigmas de interacción, incluyendo lenguajes de script y desarrollo de interfaces conversacionales para aplicaciones móviles.

Un aspecto clave de esta tecnología consiste en realizar un correcto modelado del usuario. Por este motivo, discutimos diversos modelos afectivos, de personalidad y contextuales. Finalmente, comentamos algunas líneas de investigación actuales relacionadas con la comunicación verbal, interacción multimodal y gestión del diálogo.

Palabras clave: diálogo; comprensión del lenguaje; gestión del diálogo; generación del lenguaje natural; síntesis del habla.

1. INTRODUCTION

Continuous advances in the development of information technologies have made it possible to access information, web applications and services from nearly anywhere, at anytime and almost instantaneously through wireless connections. Devices such as smartphones and tablets are widely used today to access the web. However, the contents are usually accessible only through web browsers, which are operated by means of traditional graphical user interfaces (GUIs).

Advanced paradigms on human-machine interaction, like the ones proposed by Ambient Intelligence and Smart Environments, emphasize greater user-friendliness, more

efficient services support, user-empowerment, and support for human interactions. In this vision, people will be surrounded by intelligent and intuitive interfaces embedded in everyday objects around us, and an environment that recognises and responds to the presence of individuals in a transparent way (Kovács & Kopacsi, 2006). This is why the systems proposed by these paradigms usually consist of a set of interconnected computing and sensing devices which surround the user pervasively in their environment and are invisible to them, providing a service that is dynamically adapted to the interaction context, so that users can interact naturally (De Silva, Morikawa, & Petra, 2012).

To ensure such a natural and intelligent interaction, it is necessary to provide an effective, easy, safe and transparent interaction between the user and the system. With this objective, as an attempt to enhance and ease human-to-computer interaction, in the last years there has been an increasing interest in simulating human-to-human communication, employing the so-called *Spoken Dialogue Systems* (SDSs; López-Cózar & Araki, 2005; McTear, 2004; Pieraccini, 2012). These systems have become a strong alternative to enhance computers with intelligent communicative capabilities employing speech, which is one of the most natural and flexible means of communication among humans.

SDSs can be defined as computer programs that accept speech as input and produce speech as output, engaging in a conversation with the user considering a given task. One goal of these systems is to make speech-based technologies more usable. Initially, they were used to ease interaction in simple tasks, such as provision of air travel information (Hempel, 2008). Nowadays, they are used in more complex scenarios, such as Intelligent Environments (Heinroth & Minker, 2013; Minker et al., 2006), in-car applications (Geutner, Steffens, & Manstetten, 2002), personal assistants (e.g., Siri, Google Now or Microsoft's Cortana; Janarthanam et al., 2013), smart homes (Krebber et al., 2004), and interaction with robots (Foster, Giuliani, & Isard, 2014). Another goal is it to make these technologies more accessible, especially for disabled and elderly people (Beskow et al., 2009; Vipperla, Wolters, & Renals, 2012), and to build assistants that are able to hold long-term relations with their users (Andrade et al., 2014; Bouakaz et al., 2014), which implies multifaceted research questions such as engagement and user modelling.

In this paper we present a review of the state of the art of this technology discussing its main advantages and pointing out some research trends. In Section 2 we discuss the fundaments of performance, addressing the main technologies employed. These technologies are used to implement several system modules, the characteristics of which vary depending on a number of factors, for example, the goal of the modules, the possibility of manually defining the behaviours of the modules, and the capability of automatically obtaining the modules from training samples.

In Section 3 we present an evolution of the technology, including some initial systems and research projects. Moreover, we discuss some sample applications in terms of health, education and embodied conversational agents.

In Section 4 we address current development paradigms to reduce the time and effort required in the processes of design, implementation and evaluation. More specifically, we focus on scripting languages and the development of conversational interfaces for mobile apps. The spoken dialogue industry has reached a maturity based on standards that pervade technology to provide high interoperability. This makes it possible to divide the market in a vertical structure of technology vendors, platform integrators, application developers, and hosting companies.

With regard to the evaluation of these systems, it is very difficult to define new procedures and measures that will be unanimously accepted by the scientific community (Lemon & Pietquin, 2012). This field can be considered to be in an initial phase of development. PARADISE (PARAdigm for DIalogue System Evaluation) is the most widely proposed methodology to perform a global evaluation of a dialogue system (Dybkjær, Bernsen, & Minker, 2004; Walker, Litman, Kamm, & Abella, 1998). This methodology combines different measures regarding task success, dialogue efficiency and dialogue quality in a single function that measures the yield of the system in direct correlation with user satisfaction. The EAGLES evaluation working group (Expert Advisory Group on Language Engineering Standards) proposes different quantitative and qualitative measures (EAGLES, 1996). In the same line, the DISC project (Spoken Language Dialogue Systems and Components) (Failenschmid, Williams, Dybkjær, & Bernsen, 1999) proposes different measures and criteria to be considered in the evaluation. More recent evaluation initiatives are focused on the assessment of usability and objective estimation of the quality of spoken dialogue interfaces (Möller, Engelbrecht, & Schleicher, 2008; Möller & Heusdens, 2013).

In Section 5 we discuss how to model the user to build more adaptive systems. Human speakers adapt their messages and the way they convey them to their interlocutors in a conversation, taking as well into account the context in which the dialogue takes place. The systems must be able to model this behaviour and try to replicate it.

Finally, in Section 6 we discuss how the specialists have recently envisioned future dialogue systems as being intelligent, adaptive, proactive, portable and multimodal. All these concepts are not mutually exclusive: for example, the system's intelligence can also be involved in the degree to which it can adapt to new situations, and this adaptiveness can result in better portability for use in different environments.

2. FUNDAMENTS

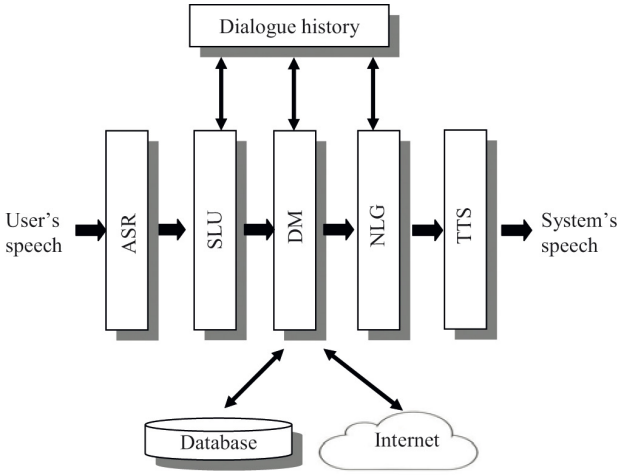
SDSs are complex to setup because the implementation requires employing a number of technologies to process the human language, which is a very complex task.

Generally speaking, these systems are built employing five main technologies:

- Automatic Speech Recognition (ASR)
- Spoken Language Understanding (SLU)
- Dialogue Management (DM)
- Natural Language Generation (NLG)
- Text-to-Speech synthesis (TTS)

Additionally, the systems typically employ other technologies to store the dialogue history. Figure 1 shows a conceptual module structure of such systems, in which the flow of information between the modules can be observed.

Figure 1: Module architecture of a SDS.



2.1. Automatic Speech Recognition

The module that implements ASR is called the *speech recogniser*. Its goal is to receive the user's speech and generate as output a recognition hypothesis, which is the sequence of words that most likely corresponds to what the user has said (Rabiner & Huang, 1993). Unfortunately, in many cases the recognition hypothesis contains errors in the form of inserted, substituted or deleted words. For example, the user may say: "Please I want to book a flight from Boston to New York" and the speech recognition result might be: "want to book a flight from Denver to New York." Note that in this case, the words "Please" and "I" have been deleted in the speech recognition result, and that the word "Boston" has been replaced with the word "Denver." ASR errors can be due to a number of factors, including environmental conditions (e.g., noise), acoustic similarity between words, and phenomena concerned with spontaneous speech, such as false starts, filled pauses and hesitations.

2.1.1. Stochastic approach

Several approaches to ASR can be found in the literature but the most used today is the stochastic one, which is based on acoustic and language models corresponding to a given language, e.g., English (Huang, Acero, & Hon, 2001).

On the one hand, the acoustic models represent the basic speech units of which the words are comprised (e.g., phonemes), and usually are represented using Hidden Markov Models (HMMs). Mostly, Gaussian Mixture Models (GMMs) are used to determine how well each state of each HMM fits a frame or a short window of frames of coefficients that represent the acoustic input. However, there are more recent methods for carrying out the fit. For example, Hinton et al. (2012) proposed to use Artificial Neural Networks (ANNs) to take into account several frames of coefficients and produce posterior probabilities over HMM states.

On the other hand, the language models determine the sentences that are expected to be uttered by the user. In most systems available in the literature, these models are compiled automatically from an analysis of a corpus of sentences in text format regarding the system's application domain. The goal of this process is to obtain statistical information regarding the appearance of a word in a sentence, given a previous history of words. Thus, the corpus of sentences must be large enough to enable obtaining significant statistics, typically at least several thousand words. Usually, the set of all the words considered by these models are stored in the so-called *dictionary*. However, recent approaches to ASR employ wide-coverage models that do not require a dictionary. For example, the approach used in the Google speech API employs a knowledge graph that provides vocabulary related to more than several million entities, such as people, places and things.

2.1.2. N-best recognition

In the previous section we mentioned that the goal of the speech recogniser is to receive the user's speech and generate as output a recognition hypothesis, which is the sequence of words that most likely corresponds to what the user has said. However, many SDSs use a method called N-best recognition, in which case the recogniser generates a list of N recognition hypotheses as the maximum, instead of just one. Typically, this list is ranked in terms of likelihood, in such a way that the first hypothesis in the list is the one with highest likelihood, the second hypothesis in the list is second with highest likelihood and so forth.

This method is commonly used by SDSs because sometimes the correct recognition hypothesis is not the top-ranked one, but one of the lower-ranked hypotheses. Hence, it is possible for the speech recogniser to consider additional information provided by other modules of the dialogue system to re-score the hypotheses in the N-best list, thus replacing the initially top-ranked hypothesis with a different one. For example, the recogniser can employ semantic information provided by the SLU module to discard hypotheses in the list or re-score them if they

do not have a correct semantic meaning. Moreover, the recogniser can take into account contextual information to re-score the hypotheses.

2.1.3. Confidence scores

Many SDSs employ techniques to process the ASR results and obtain scores regarding the speech recogniser's confidence on the recognised words. These scores are typically real numbers in the range 0.0 - 1.0, which are attached to the words. A low value of the confidence score attached to a given word represents low confidence in the correct recognition of the word, whereas a high score denotes the opposite. These scores can be very important for the performance of a SDS, since by using them the system can decide to confirm a word if its confidence score is under a certain confidence threshold.

A method to compute the confidence scores followed by several researchers takes into account the N-best list of recognition hypotheses, and assigns a higher (or lower) score to each word considering whether the word appears in a large (or small) number of hypotheses (Cox & Cawley, 2003; Liu & Fung, 2003).

N-best lists can also be used to store the possible outputs of the SLU module given the ASR result, which will be discussed in the next section. For example, this can be useful for dialogue state tracking, whose goal is to estimate the user's goal as the dialogue progresses (Wang & Lemon, 2013). Recent work on the processing of N-best lists and confidence scores can be found in the Belief Tracking approach embodied in the Dialogue State Tracking Challenges (DSTC; Williams, 2012).

2.2. Spoken Language Understanding

As can be observed in Figure 1, the output of the speech recogniser is the input to the Spoken Language Understanding (SLU) module. The goal of this module is to obtain a semantic representation of the input, which typically is stored in the form of one or more *frames* (Allen, 1995). Essentially, a frame is a kind of record comprising several fields, which are called *slots*. For example, a SDS developed to provide flight information and register flight bookings might use a simple frame comprised of the following slots to *understand* the data in the ASR results:

```
speechActType
departureCity
destinationCity
departureDate
arrivalDate
airLine
```

Thus, if we consider again the example on flight booking mentioned in Section 2.1, an ASR result could be as follows (confidence scores are noted within brackets):

WANT (0.8676) GO (0.6745) BOOK (0.7853) A (0.7206) FLIGHT (0.6983) FROM (0.6205) DENVER (0.3935) TO (0.6874) NEW (0.8562) YORK (0.9876)

Thus, the frame obtained from the analysis of this sentence might be:

```
speechActType: flightBooking (0.6745)
departureCity: Denver (0.3935)
destinationCity: New York (0.8562)
```

In this frame, the confidence scores have been attached to the values of the slots. According to the frame, the dialogue system has correctly understood that the user wants to make a flight booking, and that the destination city is New York. However, it has incorrectly understood the departure city due to an ASR error.

The task to be performed by the SLU module is very challenging due to the specific difficulties inherent in the processing of natural language, such as ambiguity, anaphora and ellipsis. To carry out SLU, this module typically employs grammar rules or statistical approaches, or some combination of both (Griol, Callejas, López-Cózar, & Riccardi, 2014). Also, it can employ the information in the *dialogue history* module (see Figure 1), which keeps track of previous system and user turns in the current dialogue. The goal is to find out whether the user has recently provided specific words which could be considered implicit in the context and thus available for sentence understanding.

Moreover, in many cases the SLU module must deal with the errors made by the ASR module, which can make the sentences ungrammatical. To deal with these problems, a number of techniques can be employed, such as relaxing the grammars, focusing the analysis on keywords, carrying out partial analyses of the recognised sentences, and employing statistical approaches (He & Young, 2005; Lemon & Pietquin, 2012).

2.3. Dialogue Management

As can be observed in Figure 1, the output of the SLU module is the input to the module that implements the Dialogue Management (DM), which is typically termed *dialogue manager*. The goal of this module is to decide what the system must do next in response to the user's input (McTear, 2004), such as providing information to the user, prompting the user to confirm words that the system is uncertain of, and prompting the user to rephrase the sentence. For example, from an inspection of the frame shown in the previous section, the dialogue manager may decide to generate a confirmation request for the departure city given that its confidence score is very low (0.3935).

To provide information to the user, the dialogue manager usually queries a local database and/or looks for data in Internet. Moreover, it takes into account information about previous dialogue turns, which is kept in the dia-

logue history module. This information is important to guide the decision of the dialogue manager towards accomplishing its task. For example, from the information in this module the dialogue manager can notice that all the data regarding a flight booking but the departure date has already been obtained from the user. Hence, the dialogue manager may decide to prompt the user for the missing data.

A number of approaches can be found in the literature for carrying out dialogue management, such as rule-based, plan-based and based on statistical reinforcement learning (Frampton & Lemon, 2009).

2.4. Natural Language Generation

The dialogue manager's decision about what the system must do next is the input to the module that carries out the Natural Language Generation (NLG). As the decision is represented abstractly, the goal is to transform it into one or more sentences in text format that must be grammatically and semantically correct, as well as coherent with the current status of the dialogue (Lemon, 2011; López, Eisman, Castro, & Zurita, 2012). Several approaches can be found in the literature for this purpose. Many systems typically employ the simplest one, which is called *template-based*, and relies on the use of a number of templates to generate a number of sentence types (Baptist & Seneff, 2000). Some parts of the templates are fixed whereas others represent gaps that must be instantiated with data provided by the dialogue manager. For example, the following template can be used to generate sentences regarding available flights connecting two cities:

```
TTS_Template_1 ::=
I found <flightAmount> FLIGHT_S/P from
<departureCity> to <destinationCity>
leaving on <departureDate>
```

In this template, the gaps are represented by means of angle brackets (e.g., `<flightAmount>`) and `FLIGHT_S/P` is a function that returns either the singular or the plural form for the word "flight," depending on the value of the `<flightAmount>` gap. For example, a sentence that this template can generate is as follows: "I found three flights from Madrid to New York leaving on Friday."

In order to be coherent with the current status of the dialogue, the NLG module must generate sentences that consider what has already been said in the dialogue. This implies omitting some words in the sentences if these have been already mentioned (ellipsis) and using pronouns instead of nouns (anaphora). To accomplish this task, this module uses the dialogue history module, which stores recently used words. This module must also avoid redundant information in the output, as well as information that is so closely related that the user could automatically infer one piece when hearing another. The process of removing such information is called *sentence aggregation* (Dalianis, 1999). It is possible to find in the literature much more sophisticated and recent approaches than template-based, such as statistical (Dethlefs, Hastie, Cuayáhuitl, & Lemon, 2013; Rieser, Lemon, & Keizer, 2014).

2.5. Text-to-Speech synthesis

Finally, the sentences in text format generated by the NLG module are the input to the last module shown in Figure 1. This module carries the Text-to-Speech synthesis (TTS), which means a transformation of the sentences into the dialogue system's speech (Dutoit, 1996). As opposed to other simple methods for speech synthesis based on concatenation of pre-recorded words, the TTS process allows transforming into speech any arbitrary text, thus avoiding the need for having the words in the sentences pre-recorded in advance.

TTS is very complex due to a number of reasons. One is the possible existence in the sentences of abbreviations (e.g., Mr., Mrs. and Ms.) and other sequences of words (e.g., numbers) that cannot be transformed into speech directly. Another reason is that the pronunciation of words is not always the same and depends on a number of factors, such the position in the sentence (e.g., beginning vs. ending) and the type of sentence (e.g., declarative vs. interrogative). Hence, the TTS process requires two steps. The first performs a transformation of the input to replace the abbreviations and other sequences of words with the corresponding words. The second does a linguistic analysis of the transformed input to include in it marks that indicate how to pronounce the words, for example, in terms of intonation and speed.

3. EVOLUTION OF THE TECHNOLOGY

Human beings have always wanted to be able to communicate with artificial companions. There are many examples in cinema and literature. Some of the most ancient examples can be found in Greek and Roman mythology in which heroes could communicate with statues of goddesses or warriors. The first serious attempts at building talking systems were initiated in the eighteenth and nineteenth centuries, when the first automata were built to imitate human behaviour. These first machines were mechanical, and it was not until the end of the nineteenth century that scientists concluded that speech could be produced electrically.

3.1. Initial systems and research projects

At the beginning of the twentieth century, Stewart (1922) built a machine that could generate vocalic sounds electrically. During the 30s, the first electric systems that could produce any type of sound were built. At the same time there appeared the first systems with very basic natural language processing capabilities for machine translation applications. During the 40s, the first computers were developed and some prominent scientists like Alan Tu-

ring pointed out their potential for applications demanding intelligence (Turing, 1950).

This was the starting point that fostered the research initiatives that in the 60s yielded the first language-based systems. For example, ELIZA (Weizenbaum, 1966) was based on keyword spotting and predefined templates to transform the user input into the system's answers.

Benefiting from the incessant improvements in the fields of ASR, natural language processing and speech synthesis, the first research initiatives related to SDSs appeared in the 80s. To some extent the origin of this research area is linked to two seminal projects: the DARPA Spoken Language Systems in the USA and the Esprit SUNDIAL in Europe. These projects were a starting point for the research in MIT and CMU, where some of the most important systems in the academia have been created.

The DARPA Communicator project stands out as one of the most important research projects in the 90s including multi-domain capabilities. This government-funded project aimed at the development of cutting-edge speech technologies, which could employ as an input not only speech but also other modalities.

Currently experts have proposed higher level objectives to develop SDSs, such as providing them with advanced reasoning, problem solving capabilities, adaptiveness, proactiveness, affective intelligence, multimodality and multilinguality (Heinroth & Minker, 2013). These new objectives are referred to the dialogue system as a whole, and represent major trends that in practice are achieved through the joint work in different areas and different components of the system.

3.2. Sample applications

There is a high variety of applications in which SDSs are currently used. One of the most widespread is information retrieval. Some sample applications are tourist and travel information (Glass et al., 1995; Os, Boves, Lamel, & Baggia, 1999), weather forecast (Zue et al., 2000), banking (Hardy et al., 2006; Melin, Sandell, & Ihse, 2001), and conference help (Andreani et al., 2006; Bohus, Raux, Harris, Eskenazi, & Rudnicky, 2007).

Spoken interaction can be the only way to access information in some cases, for example when the screen is too small to display information (e.g., hand-held devices) or when the user eyes are busy with other tasks (e.g., driving; Boves & Os, 2002; Jokinen, Kanto, & Rissanen, 2004). It is also useful for remote control of devices and robots, especially in smart environments (Krsmanovic, Spencer, Jurafsky, & Ng, 2006; Menezes, Lerasle, Dias, & Germa, 2007; Minker, Haiber, Heisterkamp, & Scheible, 2004).

3.2.1. Health

SDSs have also proven to be useful for providing the general public with access to telemedicine services, promoting patients' involvement in their own care, assisting in health care delivery, and improving patient outcome. Bickmore and Giorgino (2006) defined these systems as being "those automated systems whose primary goal is to provide health communication with patients or consumers primarily using natural language dialogue."

These systems offer an innovative mechanism for providing cost-effective healthcare services within reach of patients who live in isolated regions, have financial or scheduling constraints, or simply appreciate confidentiality and privacy. Also, as they are based on speech, they are suitable for users with a wide range of computer, reading and health literacy skills. In general healthcare, professionals can only dedicate a very limited amount of time to each patient. Thus, patients can feel intimidated to ask questions, or to ask for information to be rephrased or simply uncomfortable to provide confidential information on face to face interviews. Many studies have shown that patients are more honest with a computer than a human clinician when disclosing potentially stigmatizing behaviours such as alcohol consumption, depression, and HIV risk behaviour (Ahmad et al., 2009; Ghanem, Hutton, Zenilman, Zimba, & Erbelding, 2005).

During the last two decades, SDSs have been increasingly used in Ambient Assisted Living providing services such as interviews (Ghanem et al., 2005; Pfeifer & Bickmore, 2010), counseling (Hubal & Day, 2006), chronic symptoms monitoring (Black, McTear, Black, Harper, & Lemon, 2005; Migneault, Farzanfar, Wright, & Friedman, 2006), medication prescription assistance and adherence (Bickmore, Puskar, Schlenk, Pfeifer, & Sereika, 2010), changing dietary behaviour (Delichatsios et al., 2001), promoting physical activity (Farzanfar, Frishkopf, Migneault, & Friedman, 2005), helping cigarette smokers quit (Ramelson, Friedman, & Ockene, 1999), speech therapy (Saz et al., 2009), and prognosis and diagnosis using different techniques (Maglogiannis, Zafiropoulos, & Anagnostopoulos, 2009).

3.2.2. Education

Education is another important application domain for SDSs. According to Roda, Angehrn, and Nabeth (2001), educative technologies should accelerate the learning process, facilitate access, personalize the learning process, and supply a richer learning environment.

These aspects can be addressed by means of multimodal conversational agents by establishing a more engaging and human-like relationship between students and systems. This is why this kind of agents have been employed to develop a number of educational systems in very different domains, including tutoring (Pon-Barry, Schultz, Bratt, Clark, & Peters, 2006), conversation practice for language learners (Fryer & Carpenter, 2006), pedagogical agents and learning companions (Cavazza, de la Camara, & Turunen, 2010), dialogues to promote reflection and metacognitive skills (Kerly, Ellis, & Bull, 2008), or role-playing actors in simulated experiential learning

environments (Griol, Molina, Sanchis de Miguel, & Callejas, 2012).

They have also been used for education and training, particularly in improving phonetic and linguistic skills, including assistance and guidance to F18 aircraft personnel during maintenance tasks (Bohus & Rudnicky, 2003), training soldiers in proper procedures for requesting artillery fire missions (Roque et al., 2006), and dialogue applications for computer-aided speech therapy with different language pathologies (Rodríguez, Saz, & Lleida, 2012).

3.2.3. Embodied conversational agents

Some of the most demanding applications for fully natural and understandable dialogues are embodied dialogue agents and personal companions. For example, Collagen is an application for building conversational assistants and collaborative agents (Rich & Sidner, 1998). AVATALK provides natural, interactive dialogues with responsive virtual humans (Hubal & Day, 2006). COMIC is a system developed for bathroom design using speech and gesture input/output, in collaboration with an avatar with facial emotions (Catizone, Setzer, & Wilks, 2003). NICE provides embodied historical and literary characters capable of natural, fun and experientially rich communication with children and adolescents (Corradini et al., 2004).

4. DEVELOPMENT PARADIGMS

As can be observed in Section 2, the dialogue system domain is highly multidisciplinary and benefits from the advances in multiple directions related to different specific areas (Williams et al., 2012). This way, current SDSs are the consequence of the work on more reliable speech recognizers, more intelligible synthetized voices and more flexible conversational behaviours, among other achievements (McTear, 2011).

Considering this multidisciplinary nature, it is no surprise that the first hallmark in the development of these systems was the appearance of modular paradigms that allowed the developers to centre on their particular areas of interest, treating the other parts as black boxes. For instance, when the first speech recognizers and synthesizers were accessible, it was a huge advance for researchers and practitioners that centred on dialogue management, as they could focus on the aspects directly related to handling the conversation without worrying about the details of how to recognize the user input or synthesize the output.

Pieraccini and Huerta (2008) highlighted the importance of "reusable components" as one of the main trends for the industry of dialogue systems, as it was and still is an important aspect to build increasingly complex applications by taking advantage of already existing modules.

This new paradigm fostered a change from proprietary ad hoc architectures to others that fulfilled the purpose of reusability by means of modular "plug-and-play" models, such as different agent-based and modular architectures (Wilks, Catizone, Worgan, & Turunen, 2011), e.g., the Galaxy Communicator (O'Neill, Hanna, Liu, Greer, & McTear, 2005; Seneff, 2002) and Olympus, which is based on Galaxy (Bohus et al., 2007).

4.1. Scripting languages

The development of SDSs has also benefited from the appearance of scripting languages that are similar to other widespread general purpose languages. The most salient example is VoiceXML.[1] According to Levow (2012), this introduced some advantages including availability, robustness, ease of use, platform-independence, and flexibility. Soon other languages appeared to take advantage of the visual part of the web, for example SALT and X+V. However, speech-based web interaction with these languages has gradually lost support. Although they are still used to build some desktop systems (e.g., in Microsoft Speech API), most of the industrial platforms that hosted interpreters have disappeared. Nevertheless, now there seems to be an upsurge of voice navigation, and new initiatives have appeared, for example, the Web Speech API.[2]

4.2. Development of conversational interfaces for mobile apps

Also we can appreciate a big change in the SDSs community, a flourish due to the availability of large quantities of speech data (Williams et al., 2012), and the possibilities offered by mobile devices and their operating systems (Neustein & Markowitz, 2013).

Speech interaction with mobile assistants in smartphones is now more popular than ever, in part due to the pertinence of speech as an interaction modality with small-sized devices, the increasing accuracy of the recognizers offered to developers, and the democratization of their development.

Android and iOS offer specific libraries for ASR and speech synthesis that allow building conversational agents focusing on the interaction only (McTear & Callejas, 2013). The development is made in general purpose object-oriented languages (e.g., Java and C#), and thus is accessible to more developers.

Also robotics is starting to be increasingly relevant in the area, the appearance of open-hardware initiatives

[1] http://www.w3.org/TR/voicexml21/
[2] http://www.w3.org/community/speech-api/

have also brought more attention to this topic, and natural interaction is central in human-robot interaction studies (Graaf & Ben Allouch, 2013; Sekmen & Challa, 2013).

It is difficult to foresee how speech interfaces will be developed in the future. However, the access of an increasingly bigger number of developers to the community, the advance of statistical approaches, the increasing possibilities to access and share corpora, and the opportunities to reuse implementations of different developers establish a good basis for a promising future.

5. MODELLING THE USER

The advances in the field of SDSs described in the previous sections have provided an excellent opportunity to build richer user models. At the beginning, the capabilities of speech recognizers were limited to very small vocabularies, and so the developed applications were very simple and took into account very little information from the users. With the development of the technology started the study of how to adapt the vocabulary for recognition and the messages synthesised to enhance the user experience. That is, now the user was the centre of the system design, instead of the application domain.

Numerous publications provide hints for voice interaction design, including insights on how to specify the requirements of SDSs taking into account the users (Cohen, Giangola, & Balogh, 2004; Harris, 2004; Kortum, 2008). Some authors have focused on particular users, and particularize the guidelines to certain profiles, for example, age and familiarity with the new technologies (Callejas, Griol, Engelbrecht, & López-Cózar, 2014).

However, nowadays the information about the user is not only considered in design time, it is included in modules that allow the system to dynamically adapt to the users' state. Currently it is possible to obtain and manage a huge amount of information about the users, not only about what they say, but also about how they say it, where the say it and even predict why they said it and what they will say next, and these abilities will be increasingly more sophisticated in the future thanks to the multidisciplinary perspectives of different sciences including computer science, linguistics, psychology and sociology. In the next subsections we provide more details on some of these dynamic sources of information about the users.

5.1. Affective models

Affective computing deals with the recognition, management and synthesis of emotions (Picard, 2003). It is particularly relevant for SDSs to adapt to the user state and also to provide flexible emotionally-coloured responses for different purposes (Callejas, López-Cózar, Ábalos, & Griol, 2011).

It might seem obvious that the main use of emotional information in dialogue systems is to try to avoid negative user states and foster positive ones. Some examples of such behaviour are to avoid user negative emotions due to system errors (Callejas, Griol, & López-Cózar, 2011), to favour engagement by diminishing boredom (Baker, D'Mello, Rodrigo, & Graesser, 2010), to maximize satisfaction (Lebai Lufti, Fernández-Martínez, Lucas-Cuesta, López-Lebón, & Montero, 2013), or to foster positive emotions to adhere to healthy habits (Creed & Beale, 2012). However, in some application domains it might also be useful to render or provoke negative states, for example, for emotional mirroring, or to try to stress the users for a specific purpose, for example, for the treatment of different types of anxiety (Callejas, Ravenet, Ochs, & Pelachaud; 2014; Qu, Brinkman, Ling, Wiggers, & Heynderickx, 2014).

There exist many different ways in which emotions are defined, represented and managed within SDSs. Emotions can be represented as points in a space (usually with two dimensions: activation and evaluation), as discrete categories or with appraisal models that consider the cause and target of the emotional response (Hudlicka, 2014). The implementation of affective SDSs relies on the representation being used. If it follows the dimensional or discrete approach, the recognition is usually based on the manifestation of the user emotion, which can be processed considering linguistic (Balahur, Mihalcea, & Montoyo, 2014) and paralinguistic cues (Schuller & Batliner, 2013). When the appraisal model is used, a more sophisticated approach must be employed in order to consider as well the possible causes of the emotion (Moors, Ellsworth, Scherer, & Frijda, 2013).

Once a particular emotion is recognized, there are several ways how to consider it to adapt the system behaviour. Moreover, the approach selected also depends on the ultimate goal of the system, such as to optimize the selection of the answer, to lead the user to an optimal state for the interaction, to build a social interaction with the user, or a combination.

In the first case, the information about the user's emotional state can be employed as another source of information used to handcraft new rules or as a new input to a statistical dialogue manager (Callejas, Griol, et al., 2011). When the objective is to change the user's emotional state or build more social relations, the system must include complex models on how emotions vary over time, and how to sustain more complex forms of affect such as engagement and trust (Acosta & Ward, 2011). These same models can be used to generate a believable system's behaviour and to fine-tune the natural language generation and speech synthesis modules.

5.2. Personality models

Not only context and emotion determine our behaviour, they are also modulated by our personality (Callejas, López-Cózar, et al., 2011). Mairesse and Walker (2011) propose to tailor the system's personality according to the application domain. For example, in a tutoring system they suggest to render extrovert and agreeable pedagogic

agents, whereas it could be interesting for a psychotherapy system to be neurotic. They also point out that the personality rendered by telesales agents could match the company's brand.

Other studies focus on adapting the systems' personality to match users' personality. For example, Nass and Yen (2012) showed that users' perception of the system's intelligence and competence increases if the perceived agent's personality matches their own.

Also, having information about the user personality makes it possible to better adapt the system behaviour. This is very relevant to engage users in order to attain better performance and increase likeability, credibility, acceptance and overall user satisfaction. In Callejas, Griol, and López-Cózar (2014) we provide a discussion on these topics, as well as a framework for evaluating whether the system personality is perceived as intended by the users, and whether it matches the users' own personality.

5.3. Contextual models

Knowing the interaction context is very important for SDSs due to various reasons. Firstly, it allows obtaining a better system performance; for example, it is possible to use different noise models that allow increasing the speech recognition rates. Secondly, the location information can be used to deliver functionalities; for example, to find near spots, or to recognize the activities being carried out by the user to provide adequate services (Zhu & Sheng, 2011).

6. RESEARCH TRENDS

Language is one of the most pervasive and complex human capabilities. Developed over thousands of years, our abilities to get involved in long-term conversations, comprising multiple persons, on noisy environments, integrating multiple input/output modalities and covering multiple concurrent tasks is really amazing. This phenomenon has been pictured in fiction, for instance, in some thought-provoking films such as *2001*, *A.I.* and *Her*, among others.

Despite the extensive list of techniques created and applied in the field of human-computer interaction, language is still the most common, fastest and natural way of communication. However, the low-level connection between language and thought makes the work on natural language technologies both a critical challenge and a great opportunity for research and innovation.

SDSs constitutes one of the most demanding areas of work as it involves the majority of the language-related subfields, from ASR to speech synthesis going through natural language understanding, semantic representation, dialogue management, affective modelling, multimodal interfaces, etc. Nevertheless, improvements in this area have many direct social and economical impacts. A recent survey carried out by Grand View Research, Inc.[3] estimated the worldwide market for intelligent virtual assistants in 2012 at USD 352 million, and forecasts an annual growth of 31.7% from 2013 to 2020. According to this report, reduction of customer service operational costs is the most prominent area where the economical impact will take advantage of this technology.

In the last few years, the integration of speech-enabled technologies in mobile platforms has become a main target. The notion of personal assistant has entered the market through widespread applications like Siri, Google Now or Microsoft's Cortana.

The additional integration of Voice Search in these platforms opens new areas of applications. In this case, the speech recogniser is in charge of the transcription from speech to text (obtaining a text query), which is then used as the input to a traditional search engine. Accordingly, by the integration of ASR and search engines, Voice Search can help users in simple tasks as exemplified in queries like: "Is there any Japanese restaurant near here?" or "Show me the weather forecast for tomorrow in Paris." However, Voice Search lacks any complex dialogue capability as it usually focuses on just one single input that generates a single output.

To sum up, research and innovation on language technologies in general and on SDSs in particular constitute a major and prominent area of interest both in the public and private sectors.

In the previous sections of this paper, we have introduced the main ideas around the notion of SDS, its components and global architecture, some common areas of application of the whole technology as well as some key user-related aspects. In this section we focus on some of the most noticeable research trends in this field.

6.1. Verbal communication

The first and sometimes one of the most critical components of a SDS is the speech recogniser. Accordingly, ASR errors are the first problem that a SDS must be able to cope with. Despite the undeniable improvements of the technology irrespective of the task under consideration, it is quite evident that there is still significant room for improvements.

Some of the main lines of research at this level are: detection and cancellation of background noise, spontaneous speech where spoken disfluencies can considerably affect the recogniser, real-time recognition or even some kind of prediction or anticipation over the next input, the integration of affect and emotion recognition as part of ASR (Batliner, Seppi, Steidl, & Schuller, 2010), and the application of new techniques apart from HMMs, such as deep neural networks (Dahl, Yu, Deng, & Acero, 2012). Although some recognition errors do not prevent a rea-

[3] http://www.grandviewresearch.com/industry-analysis/intelligent-virtual-assistant-industry

sonable understanding of the user input (for instance, the detection of the main intent and keywords), there are still many cases in which the ASR's output leads to a complete semantic misunderstanding.

6.2. Multimodal interaction

Spoken language understanding (SLU) plays a crucial role in the design and implementation of SDSs. However, a natural user interaction not only requires reliable speech recognition but also the detection and analysis of additional nonverbal communication, such as facial expressions or emotional state and gesture, among others (Bui, 2006; López-Cózar & Araki, 2005).

The incorporation of multimodal interaction in Ambient Intelligence environments has become a basic goal in many research programs. For example, the first EU Call under the Horizon 2020 program in the area of language technologies (ICT-22-2014) has focused on multimodal and natural computer interaction.

Research over the current state of the art in multimodal SDSs includes topics such as semantic multimodal fusion (Russ et al., 2005). Additionally, some initial results demonstrate that using additional channels it is possible to reduce the ASR error rate employing multimodal disambiguation (Longé, Eyraud, & Hullfish, 2012).

Multimodal recognition of emotions has attracted the research community recently (Zeng, Pantic, Roisman, & Huang, 2009). For example, Calvo and D'Mello (2010) presented a survey on the combination of physiology, face, voice, text, body language and complex multimodal characterization.

6.3. Dialogue management

While introducing the global architecture and the main functional modules of a SDS, Section 2 has presented the Dialogue Manager as the component in charge of the coordination of the human-computer interaction. Different approaches for dialogue modelling have appeared in the last decades, each assuming a specific formalisation of the notion of dialogue. Taking into account their practical and theoretical aspects, some of the most prominent dialogue management approaches are the following (Jurafsky & Martin, 2009):

- Finite-state and dialogue-grammar based
- Frame-based
- Information State Update (ISU)
- Agent-based
- Markov Decision Processes (MDPs) and Partially Observable Markov Decision Processes (POMDPs).

Finite-state models conceive the dialogue as a sequence of steps over a state transition network. The nodes capture the implicit dialogue state and correspond to the system's utterances (answers, prompts, etc.), while the transitions between the nodes determine all the possible paths (Cohen, 1997). McTear (2002) described the Nuance automatic banking system implemented with this approach. Although simplicity can be mentioned as its main advantage, its lack of flexibility represents a crucial drawback. However, it is still a common strategy used to cope with basic operations in call centers.

Frame-based approaches have been introduced in Section 2. This dialogue management strategy is based on the idea that some components (called slots) of the dialogue often appear together and are required to complete a task. This approach incorporates flexibility as the order of filling the slots can be arbitrary, and even makes more natural the interactions as several slots can be filled in a single turn, or even it is possible to overwrite previous values of the slots, allowing correction and repair mechanisms. The frame-based framework originated some variations: schemas, agendas (used in the Carnegie Mellon Communicator system; Bohus & Rudnicky, 2003), task structure graphs, type hierarchies and blackboards (Rothkrantz, Wiggers, Flippo, Woei-A-Jin, & van Vark, 2004).

The Information State Update (ISU) approach models all the available information during the dialogue as an "Information State" (Larsson & Traum, 2000). Consequently, this state integrates information related to the state of all the participants in the dialogue. Basically, this state comprises all the information gathered during the previous contributions to the dialogue by the participants, and models the future actions to be taken by the dialogue manager. The ISU approach can be conceived as a declarative model of the dialogue.

All the approaches described so far require a computational linguist expert to formalize, design and implement the dialogue scheme itself. This hand-crafted strategy impacts on the global costs for the design, implementation and mainly on the maintainability of the dialogue system. In order to overcome these limitations, other approaches can be found in the literature, such as the agent-based and those focused on machine learning techniques.

The agent-based approach is particularly useful when it is necessary to execute and monitor operations in a dynamically changing application domain. It makes it possible to combine the benefits of different dialogue control models, such as finite-state based dialogue control and frame-based dialogue management (Chu, O'Neill, Hanna, & McTear, 2005). Similarly, it can benefit from alternative dialogue management strategies, such as system-initiative and mixed-initiative (Walker, Hindle, Fromer, Di Fabbrizio, & Mestel, 1997).

Recent research has applied machine learning techniques to automatically infer dialogue systems. Among these techniques, the use of MDPs and POMDPs are worth mentioning. Accordingly, the methodological motivation as well as the technical kernel relies on the possibility of inducing a statistical framework from a huge corpus of dialogues (Young, Gasic, Thomson, & Williams, 2013). Some advantages provided by this framework

need to be highlighted. Firstly, the incorporation on an explicit representation of uncertainty, which makes more robust the final system for verbal (speech) and non-verbal recognition in comparison to rule-based models. Secondly, the learning capability of the framework, which represents a significant reduction of developing costs. However, the tasks around data collection and annotation of the huge dialogue corpora that are required may jeopardise this second advantage.

6.3.1. Meta-cognition and incrementality

The human ability to get involved in complex interactions that create dialogues can be considered as a cognitive skill. This way, dialogue management is a technical sub-field which tends to mimic this cognitive skill using different approaches, as previously discussed. However, humans have the capability to reflect on their own behaviour and to use this reflection for improvement. The incorporation of metacognitive capabilities to the field of SDSs represents a challenging and promising research line (Alexandersson et al., 2014; EU-funded Metalogue project[4]). The turn-taking mechanism of standard Interaction Management architectures are based on complete sentences. However, human communication is intrinsically incremental. Some outstanding research is currently focusing on this topic (Schlangen & Skantze, 2011; EU-funded Parlance project[5]).

7. CONCLUSIONS

In this paper we have presented a short study on the state of the art of spoken dialogue systems, which are computer programs developed to interact with users employing speech in order to provide them with specific automated services. A key aspect with these systems is that the interaction is carried out by means of dialogue turns, which in many studies available in the literature, researchers aim to make as similar as possible to those between humans in terms of naturalness, intelligence and affective content.

The field is too broad to make a detailed study in just one paper. Thus, we have addressed a limited number of aspects to provide the reader with some basic knowledge on the core technologies employed for the development. Also, we have aimed at showing the technological challenges related to speech and language processing that limit the use of current systems for a wider range of potential users and applications.

In addition, we have presented an evolution of this technology and discussed some challenging applications, such as health, education and embodied conversational agents. As an outcome of the technological evolution, we have addressed the development paradigms, discussing specific scripting languages as well as development of conversational interfaces for mobile apps.

Given that the correct modelling of the user is a key aspect for this technology, we have addressed current models for affection, personality and contextual processing.

Finally, we have discussed some current research trends in terms of verbal communication, multimodal interaction and dialogue management.

REFERENCES

Acosta, J. C., & Ward, N. G. (2011). Achieving rapport with turn-by-turn, user-responsive emotional coloring. *Speech Communication, 53*(9–10), 1137–1148. http://dx.doi.org/10.1016/j.specom.2010.11.006

Ahmad, F., Hogg-Johnson, S., Stewart, D. E., Skinner, H. A., Glazier, R. H., & Levinson, W. (2009). Computer-assisted screening for intimate partner violence and control: A randomized trial. *Annals of Internal Medicine, 151(2)*, 93–102. http://dx.doi.org/10.7326/0003-4819-151-2-200907210-00124

Alexandersson, J., Girenko, A., Spiliotopoulos, D., Petukhova, V., Klakow, D., Koryzis, D., … & Gardner, M. (2014). Metalogue: A multiperspective multimodal dialogue system with metacognitive abilities for highly adaptive and flexible dialogue management. *Proceedings of 10th International Conference on Intelligent Environments (IE '14)*, 365–368. http://dx.doi.org/10.1109/IE.2014.67

Allen, J. (1995). *Natural language understanding*. Redwood City, CA: The Benjamin Cummings.

Andrade, A. O., Pereira, A. A., Walter, S., Almeida, R., Loureiro, R., Compagna, D., & Kyberd, P. J. (2014). Bridging the gap between robotic technology and health care. *Biomedical Signal Processing and Control, 10*, 65–78. http://dx.doi.org/10.1016/j.bspc.2013.12.009

Andreani, G., Di Fabbrizio, D., Gilbert, M., Gillick, D., Hakkani-Tur, D., & Lemon, O. (2006). Let's DISCOH: Collecting an annotated open corpus with dialogue acts and reward signals for natural language helpdesks. *IEEE 2006 Workshop on Spoken Language Technology*, 218–221. http://dx.doi.org/10.1109/SLT.2006.326794

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human–Computer Studies, 68*(4), 223–241. http://dx.doi.org/10.1016/j.ijhcs.2009.12.003

Balahur, A., Mihalcea, R., & Montoyo, A. (2014). Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. *Computer Speech and Language, 28*(1), 1–6. http://dx.doi.org/10.1016/j.csl.2013.09.003

Baptist, L., & Seneff, S. (2000). GENESIS-II: A versatile system for language generation in conversational system applications. *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP '00)*, *3*, 271–274.

Batliner, A., Seppi, D. Steidl, S., & Schuller, B. (2010). Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach. *Advances in Human Computer Interaction, 2010*. http://dx.doi.org/10.1155/2010/782802

Beskow, J., Edlund, J., Granström, B., Gustafson, J., Skantze, G., & Tobiasson, H. (2009). The MonAMI reminder: A spoken dialogue system for face-to-face interaction. *Proceedings of the 10th INTERSPEECH Conference 2009*, 296–299.

[4] www.metalogue.eu

[5] https://sites.google.com/site/parlanceprojectofficial/

Bickmore, T., & Giorgino, T. (2006). Health dialog systems for patients and consumers. *Journal of Biomedical Informatics, 39*(5), 556–571. http://dx.doi.org/10.1016/j.jbi.2005.12.004

Bickmore, T. W., Puskar, K., Schlenk, E. A., Pfeifer, L. M., & Sereika, S. M. (2010). Maintaining reality: Relational agents for antipsychotic medication adherence. *Interacting with Computers, 22*(4), 276–288. http://dx.doi.org/10.1016/j.intcom.2010.02.001

Black, L. A., McTear, M. F., Black, N. D., Harper, R., & Lemon, M. (2005). Appraisal of a conversational artefact and its utility in remote patient monitoring. *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, 506–508. http://dx.doi.org/10.1109/CBMS.2005.33

Bohus, D., Raux, A., Harris, T. K., Eskenazi, M., & Rudnicky, A. I. (2007). Olympus: An open-source framework for conversational spoken language interface research. Computer Science Department, Carnegie Mellon University. Retrieved from http://www.cs.cmu.edu/~max/mainpage_files/bohus%20et%20al%20olympus_hlt2007.pdf

Bohus, D., & Rudnicky, A. I. (2003). RavenClaw: Dialog management using hierarchical task decomposition and an expectation agenda. *Proceedings of the 8th European Conference on Speech Communication and Technology. EUROSPEECH 2003–INTERSPEECH 2003*, 597–600.

Bouakaz, S., Vacher, M., Bobillier Chaumon, M.-E., Aman, F., Bekkadja, S., Portet, F., … & Chevalier, T. (2014). CIRDO: Smart companion for helping elderly to live at home for longer. *IRBM, 35*(2), 100–108. http://dx.doi.org/10.1016/j.irbm.2014.02.011

Boves L., & Os, E. den (2002). *Multimodal services–A MUST for UMTS* (Tech. Rep.). EURESCOM 2002.

Bui, T. H. (2006). Multimodal dialogue management - State of the art. Human Media Interaction Department, University of Twente (Vol. 2).

Callejas, Z., Griol, D., Engelbrecht, K.-P., & López-Cózar, R. (2014). A clustering approach to assess real user profiles in spoken dialogue systems. In J. Mariani, S. Rosset, M. Garnier-Rizet & L. Devillers (Eds.), *Natural interaction with robots, knowbots and smartphones* (pp. 327–334). New York: Springer. http://dx.doi.org/10.1007/978-1-4614-8280-2_29

Callejas, Z., Griol, D., & López-Cózar, R. (2011). Predicting user mental states in spoken dialogue systems. *EURASIP Journal on Advances in Signal Processing*, *2001*, 6. http://dx.doi.org/10.1186/1687-6180-2011-6

Callejas, Z., Griol, D., & López-Cózar, R. (2014). A framework for the assessment of synthetic personalities according to user perception. *International Journal of Human–Computer Studies, 72*(7), 567–583. http://dx.doi.org/10.1016/j.ijhcs.2014.02.002

Callejas, Z., López-Cózar, D., Ábalos, N., & Griol, D. (2011). Affective conversational agents: The role of personality and emotion in spoken interactions. In D. Pérez-Marín & I. Pascual-Nieto (Eds.), *Conversational agents and natural language interaction: Techniques and effective practices* (pp. 203–222). IGI Global. http://dx.doi.org/ 10.4018/978-1-60960-617-6.ch009

Callejas, Z., Ravenet, B., Ochs, M., & Pelachaud, C. (2014). A model to generate adaptive multimodal job interviews with a virtual recruiter. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, 3615–3619.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18–37. http://dx.doi.org/10.1109/T-AFFC.2010.1

Catizone, R., Setzer, A., & Wilks, Y. (2003). Multimodal dialogue management in the COMIC project. *Proceedings of the EACL-03 Workshop on 'Dialogue Systems: Interaction, Adaptation and Styles of Management'. European Chapter of the Association for Computational Linguistics*, 25–34.

Cavazza, M., de la Camara, R. S., & Turunen, M. (2010). How was your day? A Companion ECA. *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, 1629–1630.

Chu, S.-W., O'Neill, I., Hanna, P., & McTear, M. (2005) An approach to multi-strategy dialogue management. *Proceedings of the 9th European Conference on Speech Communication and Technology. EUROSPEECH 2005–INTERSPEECH 2005,* 865–868.

Cohen, P. (1997). Dialogue modeling. In R. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, & V. Zue (Eds.), *Survey of the state of the art in human language technology* (pp. 204–210). New York: Cambridge University Press.

Cohen, M. H., Giangola, J. P., & Balogh, J. (2004). *Voice user interface design*. Boston, MA: Addison-Wesley.

Corradini, A., Fredriksson, M., Mehta, M., Königsmann, J., Bernsen, N. O., & Johanneson, L. (2004). Towards believable behavior generation for embodied conversational agents. *Proceedings of the Workshop on Interactive Visualisation and Interaction Technologies (IV&IT)*, 946–953.

Cox, S. J., & Cawley, G. (2003). The use of confidence values in vector-based call-routing. *Proceedings of the 8th European Conference on Speech Communication and Technology. EUROSPEECH 2003–INTERSPEECH 2003*, 633–636.

Creed, C., & Beale, R. (2012). User interactions with an affective nutritional coach. *Interacting with Computers, 24*(5), 339–350. http://dx.doi.org/10.1016/j.intcom.2012.05.004

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech and Language Processing, 20*(1), 30–42. http://dx.doi.org/10.1109/TASL.2011.2134090

Dalianis, H. (1999). Aggregation in natural language generation. *Computational Intelligence, 15*(4), 384–414. http://dx.doi.org/10.1111/0824-7935.00099

De Silva, L. C., Morikawa, C., & Petra, I. M. (2012). State of the art of smart homes. *Engineering Applications of Artificial Intelligence, 25*(7), 1313–1321. http://dx.doi.org/10.1016/j.engappai.2012.05.002

Delichatsios, H., Friedman, R. H., Glanz, K., Tennstedt, S., Smigelski, C., Pinto, B., … & Gillman, M. W. (2001). Randomized trial of a "talking computer" to improve adults' eating habits. *American Journal of Health Promotion, 15*(4), 215–224. http://dx.doi.org/10.4278/0890-1171-15.4.215

Dethlefs, N., Hastie, H., Cuayáhuitl, H., & Lemon, O. (2013). Conditional random fields for responsive surface realisation using global features. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, 1254–1263.

Dutoit, T. (1996). *An introduction to Text-to-Speech synthesis*. Dordrecht: Kluwer Academic.

Dybkjær, L., Bernsen, N. O., & Minker, W. (2004). Evaluation and usability of multimodal spoken language dialogue systems. *Speech Communication, 43*(1–2), 33–54. http://dx.doi.org/10.1016/j.specom.2004.02.001

Expert Advisory Group on Language Engineering Standards (EAGLES) (1996). *Evaluation of natural language processing systems* (Tech. Rep.). EAGLES Document EAG-EWG-PR2. Center for Sprogteknologi, Copenhagen.

Failenschmid, K., Williams, D., Dybkjær, L., & Bernsen, N. (1999). *DISC Deliverable D3.6* (Tech. Rep.). NISLab, University of Southern Denmark.

Farzanfar, R., Frishkopf, S., Migneault, J., & Friedman, R. (2005). Telephone-linked care for physical activity: A qualitative evaluation of the use patterns of an information technology program for patients. *Journal of Biomedical Informatics, 38*(3), 220–228. http://dx.doi.org/10.1016/j.jbi.2004.11.011

Foster, M. E., Giuliani, M., & Isard, A. (2014). Task-based evaluation of context-sensitive referring expressions in human-robot dialogue. *Language, Cognition and Neuroscience, 29*(8), 1018–1034. http://dx.doi.org/10.1080/01690965.2013.855802

Frampton, M., & Lemon, O. (2009). Recent research advances in reinforcement learning in spoken dialogue systems. *Knowledge Engineering Review, 24*(4), 375–408. http://dx.doi.org/10.1017/S0269888909990166

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. *Language Learning and Technology, 10*(3), 8–14.

Geutner, P., Steffens, F., & Manstetten, D. (2002). Design of the VICO spoken dialogue system: Evaluation of user expectations by Wizard-of-Oz experiments. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC '02),* Canary Islands.

Ghanem, K. G., Hutton, H. E., Zenilman, J. M., Zimba, R., & Erbelding, E. J. (2005). Audio computer assisted self interview and face to face interview modes in assessing response bias among STD clinic patients. *Sexually Transmitted Infections, 81*(5), 421–425. http://dx.doi.org/10.1136/sti.2004.013193

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

Graaf, M. M. A. de, & Ben Allouch, S. (2013). Exploring influencing variables for the acceptance of social robots. *Robotics and Autonomous Systems, 61*(12), 1476–1486. http://dx.doi.org/10.1016/j.robot.2013.07.007

Griol, D., Callejas, Z., López-Cózar, R., & Riccardi, G. (2014). A domain-independent statistical methodology for dialog management in spoken dialog systems. *Computer Speech and Language, 28*(3), 743–768. http://dx.doi.org/10.1016/j.csl.2013.09.002

Griol, D., Molina, J. M., Sanchis de Miguel, A., & Callejas, Z. (2012). A proposal to create learning environments in virtual worlds integrating advanced educative resources. *Journal of Universal Computer Science, 18*(18), 2516–2541. http://dx.doi.org/10.3217/jucs-018-18-2516

Hardy, H., Biermann, A., Bryce Inouye, R., McKenzie, A., Strzalkowski, T., Ursu, C., … & Wu, M. (2006). The AMITIÉS system: Data-driven techniques for automated dialogue. *Speech Communication, 48*(3–4), 354–373. http://dx.doi.org/10.1016/j.specom.2005.07.006

Harris, R. A. (2004). *Voice interaction design: Crafting the new conversational speech systems*. Morgan Kaufmann.

He, Y., & Young, S. (2005). Semantic processing using the Hidden Vector State Model. *Computer Speech and Language, 19*(1), 85–106. http://dx.doi.org/10.1016/j.csl.2004.03.001

Heinroth, T., & Minker, W. (2013). *Introducing spoken dialogue systems into Intelligent Environments*. New York: Springer. http://dx.doi.org/10.1007/978-1-4614-5383-3

Hempel, T. (2008). *Usability of speech dialogue systems: Listening to the target audience*. Springer.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. Jaitly, N., … & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82–97. http://dx.doi.org/10.1109/MSP.2012.2205597

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken language processing: A guide to theory, algorithm and system development*. Prentice Hall.

Hubal, R., & Day, R. S. (2006). Informed consent procedures: An experimental test using a virtual character in a dialog systems training application. *Journal of Biomedical Informatics, 39*(5), 532–540. http://dx.doi.org/10.1016/j.jbi.2005.12.006

Hudlicka, E. (2014). Affective BICA: Challenges and open questions. *Biologically Inspired Cognitive Architectures, 7*, 98–125. http://dx.doi.org/10.1016/j.bica.2013.11.002

Janarthanam, S., Lemon, O., Liu, X., Bartie, P., Mackaness, W., & Dalmas, T. (2013). A multithreaded conversational interface for pedestrian navigation and question answering. *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 151–153.

Jokinen, K., Kanto, K., & Rissanen, J. (2004). Adaptative user modelling in AthosMail. *Lecture Notes on Computer Science, 3196*, 149–158. http://dx.doi.org/10.1007/978-3-540-30111-0_12

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics* (2nd ed.). Prentice Hall.

Kerly, A., Ellis, R., & Bull, S. (2008). CALMsystem: A conversational model for learner modelling. *Knowledge-Based Systems, 21*(3), 238–246. http://dx.doi.org/10.1016/j.knosys.2007.11.015

Kortum, P. (2008). *HCI beyond the GUI: Design for haptic, speech, olfactory, and other nontraditional interfaces*. Morgan Kaufmann.

Kovács, G. L., & Kopácsi, S. (2006). Some aspects of Ambient Intelligence. *Acta Polytechnica Hungarica, 3*(1), 35–60.

Krebber, J. Möller, S., Pegam, R., Jekosch, U., Melichar, M., & Rajman, M. (2004). *Wizard-of-Oz tests for a dialog system in smart homes*. Paper presented at the Joint Congress CFA/DAGA '04, Strasbourg.

Krsmanovic, F., Spencer, C., Jurafsky, D., & Ng. A. Y. (2006). Have we met? MDP based speaker ID for robot dialogue. *Proceedings of the 9th International Conference on Spoken Language Processing, INTERSPEECH 2006–ICSLP*, 461–464.

Larsson, S. & Traum, D. R. (2000). Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering, 6*(4), 323–340. http://dx.doi.org/10.1017/S1351324900002539

Lebai Lutfi, S., Fernández-Martínez, F., Lucas-Cuesta, J. M., López-Lebón, L., & Montero, J. M. (2013). A satisfaction-based model for affect recognition from conversational features in spoken dialog systems. *Speech Communication, 55*(7–8), 825–840. http://dx.doi.org/10.1016/j.specom.2013.04.005

Lemon, O. (2011). Learning what to say and how to say it: Joint optimisation of spoken dialogue management and natural language generation. *Computer Speech and Language, 25*(2), 210–221. http://dx.doi.org/10.1016/j.csl.2010.04.005

Lemon, O., & Pietquin, O. (Eds.) (2012). *Data-driven methods for adaptive spoken dialogue systems: Computational learning for conversational interfaces*. Springer. http://dx.doi.org/10.1007/978-1-4614-4803-7

Levow, G.-A. (2012). Bridging gaps for spoken dialog system frameworks in instructional settings. *Proceedings of NAACL–HLT Workshop on Future Directions and Needs in the Spoken Dialogue Community: Tools and Data*, 21–22.

Liu, Y., & Fung, P. (2003). Automatic phone set extension with confidence measure for spontaneous speech. *Proceedings of the 8th European Conference on Speech Communication and Technology. EUROSPEECH 2003–INTERSPEECH 2003*, 2741–2744.

Longé, M., Eyraud, R., & Hullfish, K.C. (2012). *Multimodal disambiguation of speech recognition. U.S. Patent No. 8095364 B2*. Retrieved from http://www.google.com/patents/US8095364

López, V., Eisman, E. M., Castro, J. L., & Zurita, J. M. (2012). A case based reasoning model for multilingual language generation in dialogues. *Expert Systems with Applications, 39*(8), 7330–7337. http://dx.doi.org/10.1016/j.eswa.2012.01.085

López-Cózar, & R., Araki, M. (2005). *Spoken, multilingual and multimodal dialogue systems: Development and assessment*. John Wiley.

Maglogiannis, I., Zafiropoulos, E., & Anagnostopoulos, I. (2009). An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. *Applied Intelligence, 30*(1), 24–36. http://dx.doi.org/10.1007/s10489-007-0073-z

Mairesse, F., & Walker, M. A. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics, 37*(3), 455–488. http://dx.doi.org/10.1162/COLI_a_00063

McTear, M. F. (2002). Spoken dialogue technology: Enabling the conversational user interface. *ACM Computing Surveys, 34*(1), 90–169. http://dx.doi.org/10.1145/505282.505285

McTear, M. F. (2004). *Spoken dialogue technology. Toward the conversational user interface*. Springer. http://dx.doi.org/10.1007/978-0-85729-414-2

McTear, M. F. (2011). Trends, challenges and opportunities in spoken dialogue research. In W. Minker, G. G. Lee, S. Nakamura, & J. Mariani (Eds.), *Spoken dialogue systems technology and design* (pp. 135–161). New York: Springer.

McTear, M. F., & Callejas, Z. (2013). *Voice application development for Android*. Packt.

Melin, H., Sandell, A., & Ihse, M. (2001). CTT-bank: A speech controlled telephone banking system–An initial evaluation. *TMH-QPSR 42*(1), 1–27.

Menezes, P., Lerasle, F., Dias, J., & Germa, T. (2007). Towards an interactive humanoid companion with visual tracking modalities. *International Journal of Advanced Robotic Systems*, 48–78.

Migneault, J. P., Farzanfar, R., Wright, J. A., & Friedman, R. H. (2006). How to write health dialog for a talking computer. *Journal of Biomedical Informatics, 39*(5), 468–481. http://dx.doi.org/10.1016/j.jbi.2006.02.009

Minker, W., Albalate, A., Buhler, D., Pittermann, A., Pittermann, J., Strauss, P.-M., & Zaykovskiy, D. (2006). Recent trends in spoken language dialogue systems. *ITI 4th International Conference on Information Communications Technology* (*ICIT '06)*, 1–2. http://dx.doi.org/10.1109/ITICT.2006.358271

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

Möller, S., Engelbrecht, K.P., & Schleicher, R. (2008). Predicting the quality and usability of spoken dialogue services. *Speech Communication, 50*(8–9), 730–744. http://dx.doi.org/10.1016/j.specom.2008.03.001

Möller, S., & Heusdens, R. (2013). Objective estimation of speech quality for communication systems. *IEEE Transactions on Audio, Speech and Language Processing, 101*(9), 1955–1967. http://dx.doi.org/10.1109/JPROC.2013.2241374

Moors, A., Ellsworth, P. C., Scherer, K. R., & Frijda, N. H. (2013). Appraisal theories of emotion: State of the art and future development. *Emotion Review, 5*(2), 119–124. http://dx.doi.org/10.1177/1754073912468165

Nass, C., & Yen, C. (2012). *The man who lied to his laptop: What we can learn about ourselves from our machines*. Current Trade.

Neustein, A., & Markowitz, J. A. (2013). *Mobile speech and advanced natural language solutions* (2013 ed.). New York: Springer. http://dx.doi.org/10.1007/978-1-4614-6018-3

O'Neill, I., Hanna, P., Liu, X., Greer, D., & McTear, M. F. (2005). Implementing advanced spoken dialogue management in Java. *Science of Computer Programming, 54*(1), 99–124. http://dx.doi.org/10.1016/j.scico.2004.05.006

Os, E. den, Boves, L., Lamel, L, & Baggia, P. (1999). Overview of the ARISE project. *Proceedings of the 6th European Conference on Speech Communication and Technology, EUROSPEECH 1999*, 1527–1530.

Pfeifer, L. M., & Bickmore, T. (2010). Designing embodied conversational agents to conduct longitudinal health interviews. *Proceedings of Intelligent Virtual Agents*, 4698–4703.

Picard, R. W. (2003). Affective computing: Challenges. *International Journal of Human–Computer Studies, 59*(1–2), 55–64. http://dx.doi.org/10.1016/S1071-5819(03)00052-1

Pieraccini, R. (2012). *The voice in the machine: Building computers that understand speech*. Cambridge, MA: MIT Press.

Pieraccini, R., & Huerta, J. M. (2008). Where do we go from here? In L. Dybkjær & W. Minker (Eds.), *Recent trends in discourse and dialogue* (pp. 1–24). Springer Netherlands. http://dx.doi.org/10.1007/978-1-4020-6821-8_1

Pon-Barry, H., Schultz, K., Bratt, E.O., Clark, B., & Peters, S. (2006). Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education, 16*(2), 171–194.

Qu, C., Brinkman, W.-P., Ling, Y., Wiggers, P., & Heynderickx, I. (2014). Conversations with a virtual human: Synthetic emotions and human responses. *Computers in Human Behavior, 34*, 58–68. http://dx.doi.org/10.1016/j.chb.2014.01.033

Rabiner, L. R., & Huang, B. H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice Hall.

Ramelson, H. Z., Friedman, R. H., & Ockene, J. K. (1999). An automated telephone-based smoking cessation education and counseling system. *Patient Education and Counseling, 36*(2), 131–144. http://dx.doi.org/10.1016/S0738-3991(98)00130-X

Rich, C., & Sidner, C. L. (1998). COLLAGEN: A collaboration manager for software interface agents. *User Modeling and User-Adapted Interaction, 8*(3–4), 315–350. http://dx.doi.org/10.1023/A:1008204020038

Rieser, V., Lemon, O., & Keizer, S. (2014). Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems. *IEEE/ACM Transactions on Audio, Speech and Language Processing, 22*(5), 979–994. http://dx.doi.org/10.1109/TASL.2014.2315271

Roda, C., Angehrn, A., & Nabeth, T. (2001). Conversational agents for advanced learning: Applications and research. *Proceedings of BotShow 2001*, 1–7.

Rodríguez, W. R., Saz, O., & Lleida, E. (2012). A prelingual tool for the education of altered voices. *Speech Communication, 54*(5), 583–600. http://dx.doi.org/10.1016/j.specom.2011.05.006

Roque, A., Leuski, A., Rangarajan, V., Robinson, S., Vaswani, A., Narayanan, S., & Traum, D. (2006). Radiobot-CFF: A spoken dialogue system for military training. *Proceedings of the 9th International Conference on Spoken Language Processing, INTERSPEECH 2006–ICSLP*, 477–480.

Rothkrantz, L. J. M., Wiggers, P., Flippo, F., Woei-A-Jin, D., & van Vark, R. J. (2004). Multimodal dialogue management. *Lecture Notes in Computer Science*, *3206*, 621–628. http://dx.doi.org/10.1007/978-3-540-30120-2_78

Russ, G., Sallans, B., & Hareter, H. (2005). Semantic based information fusion in a multimodal interface. *Proceedings of the International Conference on Human–Computer Interaction, HCI '05*, Las Vegas. Lawrence Erlbaum.

Saz, O., Yin, S. C., Lleida, E., Rose, R., Vaquero, C., & Rodríguez, W. R. (2009). Tools and technologies for computer-aided speech and language therapy. *Speech Communication, 51*(10), 948–967. http://dx.doi.org/10.1016/j.specom.2009.04.006

Schlangen, D., & Skantze, G. (2011). A general, abstract model of incremental dialogue processing. *Dialogue & Discourse, 2*(1), 83–111. http://dx.doi.org/10.5087/dad.2011.105

Schuller, B. W., & Batliner, A. (2013). *Computational paralinguistics: Emotion, affect and personality in speech and language processing*. John Wiley & Sons. http://dx.doi.org/10.1002/9781118706664

Sekmen, A., & Challa, P. (2013). Assessment of adaptive human–robot interactions. *Knowledge-Based Systems, 42*, 49–59. http://dx.doi.org/10.1016/j.knosys.2013.01.003

Seneff, S. (2002). Response planning and generation in the MERCURY flight reservation system. *Computer Speech and Language, 16*(3–4), 283–312. http://dx.doi.org/10.1016/S0885-2308(02)00011-6

Stewart, J. Q. (1922). An electrical analogue of the vocal organs. *Nature, 110*, 311–312. http://dx.doi.org/10.1038/110311a0

Turing, A. (1950). Computing machinery and intelligence. *Mind, 236*, 433–460. http://dx.doi.org/10.1093/mind/LIX.236.433

Vipperla, R., Wolters, M., & Renals, S. (2012). Spoken dialogue interfaces for older people. In K. J. Turner (Ed.), *Advances in home care technologies* (pp. 118–137). IOS Press.

Walker, M., Hindle, D., Fromer, J., Di Fabbrizio, G., & Mestel, C. (1997). Evaluating competing agent strategies for a voice email agent. *Proceedings of the 5th European Conference on Speech Communication and Technology, EUROSPEECH 1997*, 2219–2222.

Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1998). Evaluating spoken dialogue agents with PARADISE: Two case studies. *Computer Speech and Language, 12*(4), 317–347. http://dx.doi.org/10.1006/csla.1998.0110

Wang, Z., & Lemon, O. (2013). A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. *Proceedings of 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 423–432.

Weizenbaum, J. (1966). ELIZA–A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45. http://dx.doi.org/10.1145/365153.365168

Wilks, Y., Catizone, R., Worgan, S., & Turunen, M. (2011). Some background on dialogue management and conversational speech for dialogue systems. *Computer Speech and Language, 25*(2), 128–139. http://dx.doi.org/10.1016/j.csl.2010.03.001

Williams, J. D. (2012). A belief tracking challenge task for spoken dialog systems. *Proceedings of NAACL HLT 2012 Workshop on future directions and Needs in the Spoken Dialog Community: Tools and Data*.

Williams, J. D., Yu, K., Chaib-draa, B., Lemon, O., Pieraccini, R., Pietquin, O., … & Young, S. (2012). Introduction to the issue

on advances in spoken dialogue systems and mobile interface. *IEEE Journal of Selected Topics in Signal Processing, 6*(8), 889–890. http://dx.doi.org/10.1109/JSTSP.2012.2234401

Young, S., Gasic, M., Thomson, B., & Williams, J. D. (2013). POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE, 101*(5), 1160–1179. http://dx.doi.org/10.1109/JPROC.2012.2225812

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T.S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39–58. http://dx.doi.org/10.1109/TPAMI.2008.52

Zhu, C., Sheng, W. (2011). Motion- and location-based online human daily activity recognition. *Pervasive and Mobile Computing, 7*(2), 256–269. http://dx.doi.org/10.1016/j.pmcj.2010.11.004

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