Chapter 15
Design and Development of an Automated Voice Agent: Theory and Practice Brought Together

Pepi Stavropoulou
University of Athens, Greece

Dimitris Spiliotopoulos
University of Athens, Greece

Georgios Kouroupetroglou
University of Athens, Greece

ABSTRACT
Sophisticated, commercially deployed spoken dialogue systems capable of engaging in more natural human-machine conversation have increased in number over the past years. Besides employing advanced interpretation and dialogue management technologies, the success of such systems greatly depends on effective design and development methodology. There is, actually, a widely acknowledged, fundamentally reciprocal relationship between technologies used and design choices. In this line of thought, this chapter constitutes a more practical approach to spoken dialogue system development, comparing design methods and implementation tools highly suited for industry oriented spoken dialogue systems, and commenting on their interdependencies, in order to facilitate the developer’s choice of the optimal tools and methodologies. The latter are presented and assessed in the light of AVA, a real-life Automated Voice Agent that performs call routing and customer service tasks, employing advanced stochastic techniques for interpretation and allowing for free form user input and less rigid dialogue structure.

INTRODUCTION
Automated Voice Agents are systems capable of communicating with users by both understanding and producing speech within a specific domain. They engage in humanlike spoken dialogues, in order to route telephone calls, give traffic information, book flights, solve technical problems and provide access to educational material among others.
Depending on their design, the speech understanding and dialogue management technology involved, they may be of two basic types:

- **Directed Dialog Systems**: ranging from finite state-based to frame-based systems (McTear, 2004). The former systems are very simple and inflexible menu-driven interfaces, where the dialogue flow is static, specified in advance, no deviations from that flow are allowed, and only a limited number of words and phrases provided by the user can be understood. The latter systems are more advanced interfaces, where the interaction is not completely predetermined and a more elaborate vocabulary can be handled. While both types of systems are primarily system-directed, frame-based systems allow for a modest level of mixed-initiative by handling over-specification in user’s input; that is the user can provide more items of information than those requested by the system at each dialogue turn.

- **Open-ended natural language conversational systems**: mixed-initiative systems, where both system and user can take control of the dialogue introducing topics, changing goals, requesting clarifications, establishing common ground. Equipped with sophisticated speech and language processing modules, they can handle long, complex and variable user input in an attempt to approximate natural human-human interaction as close as possible.

The two types of systems to a significant extent reflect the differences in trends and directions followed by the spoken dialogue industry compared to spoken dialogue research during the last decades. As commercial dialogue systems aim primarily at usability and task completion (Pieraccini & Huerta, 2008), focus was placed on ways to restrict users’ input, in order to amend for speech technology limitations and reach industrial standards for useful applications. As a result, industry opted for more directed dialogue systems, which are the most commonly used on the market today.

Furthermore, the need for cost reduction, ease of development and maintenance has led to the development of reusable dialogue components and integration platforms promoting modularity and interoperability. Accordingly, VoiceXML (McGlashan et al., 2004; Larson, 2002) has become an industry standard for building voice applications, which exploits the existing and universally accepted web infrastructures eliminating the need for specific application protocol interfaces (APIs) designated to speech technology integration. Based on the Form Interpretation Algorithm it incorporates a frame-based architecture, providing an industry-feasible trade-off between naturalness and robustness.

Research, on the other hand, aims primarily at naturalness and freedom of communication (Pieraccini & Huerta, 2008). In an attempt to handle almost unrestricted user input and allow for a fully mixed initiative, conversational interface, focus has been on dialogue manager architectures exploiting inference and planning as part of a truly conversational agent. Speech act interpretation (Allen, 1995, Chapter 17; Cohen & Perrault, 1979; Core & Allen, 1997; Allen et al., 2007) and conversational games (Kowtko et al., 1993; Pulman, 2002), discourse structure (Grosz & Sidner, 1986; Stent et al., 1999; Fischer et al., 1994) and prosody manipulation (Hirschberg et al., 1995; Noth et al., 2002) are only some of the topics in an ongoing research for building natural language interfaces.

Furthermore, accessibility issues have gained attention, being important not only for the visual impaired (Freitas & Kouroupetroglou, 2008) but also to people with various disabilities (Fellbaum & Kouroupetroglou, 2008). In particular, spoken dialogue systems are considered as key technological factors for the universal accessibility strategies.