Chapter 2
Modern Standards for VoiceXML in Pervasive Multimodal Applications

Dirk Schnelle-Walka  
Technische Universität Darmstadt, Germany

Stefan Radomski  
Technische Universität Darmstadt, Germany

Max Mühlhäuser  
Technische Universität Darmstadt, Germany

ABSTRACT
In this chapter, we will consider the language support of VoiceXML 2.1 to express flexible dialogs in pervasive environments. Missing information about the environment and the inability to react to external events lead to rigid and verbose dialogs. But building upon the recently defined W3C MMI architecture we present an approach where dialog authors can adapt their dialogs’ behavior with regard to the users’ surroundings and incorporate available information and devices from the pervasive environment. Adding these features extends the expressiveness of VoiceXML 2.1, and allows for an integration into a multimodal and mobile interaction as anticipated in the, as of now, dormant VoiceXML 3.0 standard.

INTRODUCTION
The modality of speech is widely considered to be a promising way to interact with systems in pervasive environments, e.g., in a smart-space or mobile setting (Sawhney & Schmandt, 2000; Turunen, 2004; Ranganathan, Chetan, Al-Muhtadi, Campbell, & Mickunas, 2005). Being surrounded by a plethora of information processing entities leaves a user with the problem on how to interact with them. While devices can employ dedicated interaction panels (i.e., touch interfaces and graphical displays), such interfaces still require the user to step up to them or carry a mobile device. With the modality of speech comes the promise of casual, effortless and hands-free interaction, suitable to command all the functionality of a pervasive environment.

DOI: 10.4018/978-1-4666-8583-3.ch002
Apart from the actual speech recognition technology, such an interaction paradigm needs a way to formally describe the spoken part of a multimodal dialog between the system and the user: VoiceXML is one such dialog description language. As part of the standardization efforts of the W3C for multimodal applications, version 2.1 of the language was given recommendation status in June 2007 with interest in towards version 3.0 dormant since four years. VoiceXML, as the most widespread and adopted language for voice centric application (Kolias, Kolias, Anagnostopoulos, Kambourakis, & Kayafas, 2010), has been proven to be useful as a modality specific markup language in mobile and embedded systems (Burkhardt, Schaeck, Henn, Hepper, & Rindtorff, 2001; Mueller, Schaefer, & Bleul, 2004; Park, Kim, Park, & Han, 2004; Fortier & Frost, 2011). Resource usage for mobile deployments was addressed by e.g. Bühler & Hamerich (2005) who developed a VoiceXML into ECMAScript compiler, allowing to author voice based applications for mobile computing. However, with VoiceXML’s roots in telephony applications and interactive voice response systems, common requirements regarding a multimodal application in pervasive environments remain unsatisfied.

Furthermore, even as speech as a modality is very suited for interaction in pervasive environments, technical issues still limit its applicability as the dominant means of interaction. While word-error-rates for limited domain vocabularies employing close-range microphones come close to the threshold of human understanding, distant speech recognition of spontaneous, overlapping speech, even when employing beam-forming microphone arrays is far from usable (see Figure 1). To overcome these issues, modern applications in pervasive environments employ multiple complementary modalities, allowing the user to fall back on more reliable, or switch to more suitable means of interaction.

Figure 1. Word-error-rates for speech recognition over time
(Fiscus et al., 2009)