1

Introduction

The Media Resource Control Protocol (MRCP) is a new protocol designed to provide a standardised, uniform, and flexible interface to network-based media processing resources such as speech recognisers and speech synthesisers. The standard MRCP specification is being developed by the Internet Engineering Task Force (IETF) and has been designed to dovetail with existing IETF protocols and Web standards created by the World Wide Web Consortium (W3C). This chapter provides a background to MRCP by introducing some of the benefits of using speech technologies and the motivations behind MRCP. A brief history of the standardisation efforts that led to the development of MRCP is also covered.

1.1 Introduction to speech applications

Speech processing technologies facilitate conversational dialogues between human and machine by exploiting the most natural and intuitive communication modality available to man, namely speech. One powerful application of speech, and the focus of this book, is its use in interactive services delivered over telephony devices. Speech processing technologies can breathe new life into rigid interactive voice response (IVR) systems resulting in a more user friendly voice-user interface (VUI) that understands natural, conversational language. Speech allows the user to circumvent complex and confusing dual tone multifrequency (DTMF) touch-tone menu structures and instead navigate quickly and fluidly to the particular service or option they seek. The result is shortened call times, increased customer satisfaction, and enhanced automation levels. Speech-enabled VUIs possess several advantages over other human–computer interaction paradigms. For a start, conversational dialogues do not interfere with visual or manual tasks such as driving a car. Secondly, dynamic, timely information may be synthesised to the user through the use of text-to-speech (TTS) technologies. Finally, audio-based media can be easily incorporated, for example music or radio broadcasts.

Telephony applications incorporating speech typically rely on *network-based* speech processing technologies. Free from the constraints of limited processing power, memory, and finite battery life, network-based media processing resources can deliver high accuracy, large vocabulary, speaker-independent speech recognition services in conjunction with natural sounding speech

Speech Processing for IP Networks Dave Burke
© 2007 John Wiley & Sons, Ltd

synthesis. A further important advantage of network-based speech processing is that it obviates the need for specialist client-side software tailored for a limited set of devices; network-based speech services may be delivered uniformly to any telephony device without limitation. Given the ubiquity of telephones in comparison with desktop computers, and coupled with the attractiveness of providing a natural and intuitive interface, it is obvious why many companies are deploying telephony speech applications to deliver premium services to their customers. Network-based speech processing permits a plethora of advanced interactive services to be delivered by standard telephony, which are equally applicable across industries ranging from financial services to healthcare. Popular examples include automated directory enquiries, store locator, bill payment, account activation, password reset, flight information, ticket reservation and purchasing, biometric applications, and a wide variety of information services.

The speech-enabled IVR market is growing and expected to continue to grow. *DataMonitor* [1] reports that the total value of the North American IVR market (the largest IVR market in the world) will grow to $709 million in 2009. Within that statistic there is some further interesting information. The revenue from standards-based IVR built on technologies such as VoiceXML and MRCP will more than double from $90 million in 2004 to $196 million in 2009, at the expense of traditional, proprietary IVR. Further, by 2009, speech-enabled IVR port shipments will account for more than 50 % of total IVR port shipments.

1.2 The MRCP value proposition

The underlying technology of speech processing has not changed significantly in the last ten years but rather has undergone a steady evolution with increasing accuracy, performance, and scalability available year on year. The increase in scalability can be largely attributed to Moore's law – indeed most modern network-based speech processors run on general purpose hardware and operating systems. There is, however, a revolution taking place in the speech industry, and it is centred on the new and standardised mechanisms by which one can integrate and use speech technologies. Until recently, speech technologies remained costly and difficult to integrate into other network equipment. Once integrated, vendor lock-in was somewhat inevitable due to the complex and proprietary interfaces exposed by a particular speech media resource. In addition, maintenance costs were incurred as a result of the need to frequently update integrations as the speech technology evolved. Prior to MRCP, speech media resources typically exposed a client–server style interface. The client was shipped as a library exposing an interface implemented in low-level languages such as C, C++, and Java. The interfaces differed significantly across different vendors, and even versions from the same vendor, thus making support of multiple speech resources very difficult and costly to maintain. Compounding this is the fact that 'one size fits all' does not apply for speech technologies since different vendors' strengths lie in different spoken languages and dialects, for example.

MRCP helps alleviate much of the headache and cost associated with the integration of speech processing technologies by delivering a standard, uniform, vendor-independent network protocol to control speech processing resources. MRCP itself is based on modern IP telephony and Web technologies, making it ideal for deployment in next generation IP networks. For example, MRCP leverages the Session Initiation Protocol (SIP), which is the core signalling protocol chosen by the 3GPP for its popular IP Multimedia Subsystem (IMS) architecture aimed at both mobile and fixed networks. As a standard, MRCP brings with it vendor independence, a proven interface, durability, and assurances of an evolution path for the protocol. Indeed, MRCP offers a win–win situation for network equipment providers and speech resource providers alike, since both parties may concentrate on their core business without expending substantial resources on designing, developing, and evolving different interfaces and APIs. Furthermore, by removing many of the integration burdens, MRCP serves to accelerate the adoption of speech technologies. MRCP allows network operators to procure and deploy network equipment from separate providers to those providing speech processing equipment and yet still retain assurances of interoperability.

1.3 History of MRCP standardisation

Before delving into the details of how MRCP came about, it is instructive to introduce briefly two of the most important standards organisations behind the Internet and Web, namely the IETF and W3C.

1.3.1 Internet Engineering Task Force

The IETF is an open international community of network designers, operators, vendors and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet. Founded in 1986, the IETF has grown to become the principal body engaged in the development of new Internet standard specifications.

Membership to the IETF is free and open to any interested person. The technical work within the IETF is done within working groups, which are focused on a particular topic or set of topics defined by their charter. Each working group has one or more working group chairs whose responsibility it is to ensure forward progress is made with a view to fulfilling the group's charter. Working groups maintain their own mailing list where the majority of the work is carried out. In addition, the IETF meets several times a year, enabling working groups to discuss topics face-to-face. Working groups belong to a particular Area (e.g. transport, network management, routing, etc). Each Area is overseen by its corresponding Area Director. The set of Area Directors collectively make up the Internet Engineering Steering Group (IESG). The IESG is responsible for the technical management of the IETF such as approving new working groups and giving final approval to Internet standards.

Early versions of an IETF specification are made available through the IETF website and mailing lists as an Internet-Draft. An Internet-Draft is released with the intention of receiving informal review and comment from a wide audience. Usually, during the development of a specification, the document is revised several times and republished. Internet-Drafts expire 6 months after publication at which time they are immediately removed from IETF repositories (though archives of older drafts can be readily found on the Web on so-called mirror websites). At any time, the current Internet-Drafts may be obtained via the Web from:

```
http://www.ietf.org/internet-drafts/
```

After a specification is deemed stable and has reached a sufficient level of technical quality, an area director may take it to the IESG for consideration for publication as a Request For Comments (RFC). Each RFC is given a unique number. A particular RFC is never updated but rather a new number is assigned for a revision. Many IETF protocols rely on centrally agreed identifiers for their operation. The Internet Assigned Numbers Authority (IANA) has been designated by the IETF to make assignments for its RFC document series. RFCs pertaining to Internet standards go through stages of development denoted by maturity levels. The entry level is called a Proposed Standard. Specifications that have demonstrated at least two independent and interoperable implementations from different code bases, and for which sufficient successful operational experience has been obtained, may be elevated to the Draft Standard level. Finally, a specification for which significant implementation and successful operational experience has been obtained may be elevated to the Internet Standard level. RFCs can be obtained via the Web from:

```
http://www.ietf.org/rfc.html
```

1.3.2 World Wide Web Consortium

The W3C is an international consortium focusing on its mission to lead the World Wide Web to its full potential by developing protocols and guidelines that ensure long-term growth of the web. Tim Berners-Lee, who invented the World Wide Web in 1989 by providing the initial specifications for URIs, HTML, and the HTTP (a protocol that runs over the Internet[1]), formed the W3C in 1994. Today, the W3C consists of a large number of member organisations and a full-time staff, which collectively publish open, non-proprietary standards for Web languages and protocols.

The technical work in the W3C is carried out by working groups consisting of participants from member organisations, W3C staff, and invited experts. Working groups conduct their work through mailing lists, regular teleconferences, and face-to-face meetings. Each working group has one or more working group chairs and is categorised into a particular activity (e.g. HTML Activity or Voice Browser Activity). Each activity falls within a domain (e.g. Architecture Domain or Interaction Domain). The W3C Team manages both the technical activities and operations of the consortium by providing overall direction, coordinating activities, promoting cooperation between members, and communicating W3C results to the members and press.

W3C technical reports or specifications follow a development process designed to maximise consensus about the content and to ensure high technical and editorial quality. Specifications start out as a Working Draft, which is a document published for review by the community, including W3C members, the public, and other technical organisations. A specification advances to the Candidate Recommendation level after it has received wide review and is believed to satisfy the working group's technical requirements. Next, a working group will usually seek to show two interoperable implementations of each feature at which time the specification may be elevated to the Proposed Recommendation level. Finally, a specification is published as a W3C Recommendation when the W3C believes that it is appropriate for widespread deployment and that it promotes the W3C's mission. W3C Technical Reports and publications may be obtained via the Web from:

```
http://www.w3.org/TR/
```

1.3.3 MRCP: from humble beginnings toward IETF standard

Cisco, Nuance, and SpeechWorks jointly developed MRCP version 1 (MRCPv1). MRCPv1 was first made publicly available through the IETF as an Internet-Draft in 2001 and was later published as an 'Informational' document under RFC 4463 [13]. Although it has enjoyed wide implementation – a testament to the popularity and benefit of the MRCP approach – this version of the protocol has not been, and will not be, developed into an actual IETF standard. MRCPv1 has several restrictions and deficiencies that ultimately prevented it from becoming a standard. The protocol leverages the Real Time Streaming Protocol (RTSP) for both setting up media streams and for transport of the

[1] The HTTP protocol is now maintained by the IETF while HTML continues to be evolved by the W3C.

MRCP messages. SIP has since become the preferred mechanism for media session initiation in MRCP architectures, and the tunnelling mechanism that leverages RTSP as a transport protocol is largely regarded as an inappropriate use of RTSP. MRCPv1 also suffers from interoperability problems due to a weakly defined data representation for recognition results returned from speech recognisers. This, coupled with many vendor-specific extensions that were required (because they were not included in MRCPv1 in the first place) meant that true interoperability was not achieved and platform vendors were often still forced to perform some level of specific integration work for each speech technology vendor. Finally, the scope of MRCPv1 is limited and there is a lack of support for speaker verification and identification engines and speech recording.

In March 2002, a formal IETF Birds-of-a-Feather (BOF) was held at the 53rd IETF meeting where like-minded vendors came together to express their interest in developing a new and significantly improved version of MRCP. Later that year, the IESG chartered the Speech Services Control (SpeechSC) Working Group comprising leading speech technology and protocol experts. The core of the charter was to develop a standard protocol to support distributed media processing of audio streams with the focus on speech recognition, speech synthesis, and speaker verification and identification. The SpeechSC Working Group's first deliverable was a requirements document [10] outlining the specific needs of a standardised successor to MRCPv1. Particularly salient requirements for the new protocol were the proposed reuse of existing protocols while, at the same time, the avoidance of redefining the semantics of an existing protocol (something which MRCPv1 did with its particular reuse of RTSP). The SpeechSC Working Group has subsequently developed MRCP version 2 (MRCPv2), a significantly improved, extended, and standardised version of MRCPv1. At the time of going to press, the MRCPv2 specification [3] had undergone wide peer review and was at 'Last-Call' – a stage in the standardisation process used to permit a final review by the general Internet community prior to publishing the specification as an RFC.

MRCPv2 leverages SIP for establishing independent media and control sessions to speech media resources, adds support for speaker verification and identification engines, and includes many welcomed extensions and clarifications that improve flexibility and interoperability. The use of SIP allows MRCPv2 to leverage the many benefits that SIP offers, including resource discovery and reuse of existing network infrastructure such as proxies, location servers, and registrars. Significantly, MRCPv2 does not tunnel the control messages through SIP but rather uses SIP to establish a dedicated connection. MRCPv1 may be thought of as an early, proprietary version of the standard MRCPv2. MRCPv1 shares many similarities to MRCPv2 and is outlined in more detail in Appendix A. Throughout the rest of the book, unless otherwise stated, the term MRCP is used to refer to MRCPv2.

Independent of the IETF standardisation work, the W3C established the Voice Browser Working Group (VBWG) in 1999, with a charter to apply Web technology to enable users to access services from their telephone via a combination of speech and DTMF. The VBWG's deliverable is the W3C Speech Interface Framework, a collection of markup languages for developing Web-based speech applications. VoiceXML [4, 5] is a core component of that framework and is a standard, easy-to-learn, Web-based technology for authoring telephony applications that employ DTMF, speech recognition, and speech synthesis. VoiceXML depends on its sibling languages for specifying speech recognition and speech synthesis behaviours. The W3C Speech Recognition Grammar Specification (SRGS) [6] is a standard, XML-based markup approach for specifying speech grammars (the set of words and phrases a speech recogniser may recognise at a given point in a dialogue). A closely related language is the W3C Semantic Interpretation for Speech Recognition (SISR) [7]. This specification enables 'tagging' of semantic information to speech grammars to facilitate a basic form of natural language understanding. The W3C Speech Synthesis Markup Language (SSML) [8] is a standard, XML-based markup approach for specifying content for speech synthesis together with a mechanism for controlling aspects of speech production such as pronunciation, volume, pitch, and rate. Both SRGS and SSML can leverage the W3C Pronunciation Lexicon Specification (PLS) [9], which allows pronunciations for words or phrases to be specified using a standard pronunciation alphabet.

The VoiceXML language is a common ‘user’ of the MRCP protocol, that is, many VoiceXML platforms employ the MRCP protocol to interact with third party speech recognisers and speech synthesisers. In many ways (though not formally called out anywhere), VoiceXML capabilities were a primary driver behind a large number of functional additions to MRCP. Historically, the VBWG and SpeechSC Working Group have shared several participants, which has served to provide a healthy amount of cross-pollination. One obvious result of this is that MRCP leverages several W3C Speech Interface Framework specifications for specifying speech recogniser and speech synthesiser behaviours including SRGS, SISR, and SSML.

1.4 Summary

In this chapter, we have provided a brief introduction to speech applications by focusing particularly on network-based speech processing technologies and the many benefits that they bring to the world of IVR. The business case for MRCP was presented, including how it significantly helps to alleviate the burden and cost of integrating speech technologies, helps to open the market by allowing network operators to ‘mix and match’ the best technology for their particular purpose, and assists in accelerating the adoption of speech technologies for commercial applications. Finally, we presented a short history of MRCP standardisation by introducing the two standard bodies that were key to its development and discussing how MRCP and its related specifications came about.

In Chapter 2, we provide a background on how modern speech processing technologies function. This chapter is recommended for readers new to speech technologies; experienced readers may prefer to skip directly to Chapter 3, which introduces the basic principles of MRCP.