

# SPEECH PROCESSING FOR IP NETWORKS

MEDIA RESOURCE CONTROL  
PROTOCOL (MRCP)

DAVE BURKE

Companion Website

 WILEY

# Contents

|                                                         |           |
|---------------------------------------------------------|-----------|
| <i>About the author</i>                                 | <i>xi</i> |
| Preface                                                 | xiii      |
| Acknowledgements                                        | xv        |
| <b>PART I BACKGROUND</b>                                | <b>1</b>  |
| 1 Introduction                                          | 3         |
| 1.1 <i>Introduction to speech applications</i>          | 3         |
| 1.2 The MRCP value proposition                          | 4         |
| 1.3 History of MRCP standardisation                     | 5         |
| 1.3.1 Internet Engineering Task Force                   | 5         |
| 1.3.2 World Wide Web Consortium                         | 6         |
| 1.3.3 MRCP: from humble beginnings toward IETF standard | 6         |
| 1.4 Summary                                             | 8         |
| 2 Basic principles of speech processing                 | 9         |
| 2.1 <i>Human speech production</i>                      | 9         |
| 2.1.1 Speech sounds: phonemics and phonetics            | 12        |
| 2.2 Speech recognition                                  | 13        |
| 2.2.1 Endpoint detection                                | 16        |
| 2.2.2 Mel-cepstrum                                      | 16        |
| 2.2.3 Hidden Markov models                              | 19        |
| 2.2.4 Language modelling                                | 23        |
| 2.3 Speaker verification and identification             | 26        |
| 2.3.1 Feature extraction                                | 28        |
| 2.3.2 Statistical modelling                             | 28        |
| 2.4 Speech synthesis                                    | 31        |
| 2.4.1 Front-end processing                              | 33        |
| 2.4.2 Back-end processing                               | 34        |
| 2.5 Summary                                             | 37        |

|       |                                           |           |
|-------|-------------------------------------------|-----------|
| 3     | Overview of MRCP                          | 39        |
| 3.1   | Architecture                              | 39        |
| 3.2   | Media resource types                      | 41        |
| 3.3   | Network scenarios                         | 42        |
| 3.3.1 | VoiceXML IVR service node                 | 42        |
| 3.3.2 | IP PBX with voicemail                     | 44        |
| 3.3.3 | Advanced media gateway                    | 45        |
| 3.4   | Protocol operation                        | 45        |
| 3.4.1 | Establishing communication channels       | 45        |
| 3.4.2 | Controlling a media resource              | 47        |
| 3.4.3 | Walkthrough examples                      | 48        |
| 3.5   | Security                                  | 52        |
| 3.6   | Summary                                   | 53        |
|       | <b>PART II MEDIA AND CONTROL SESSIONS</b> | <b>55</b> |
| 4     | Session Initiation Protocol (SIP)         | 57        |
| 4.1   | Introduction                              | 57        |
| 4.2   | Walkthrough example                       | 58        |
| 4.3   | SIP URIs                                  | 63        |
| 4.4   | Transport                                 | 63        |
| 4.5   | Media negotiation                         | 64        |
| 4.5.1 | Session description protocol              | 64        |
| 4.5.2 | Offer/answer model                        | 66        |
| 4.6   | SIP servers                               | 67        |
| 4.6.1 | Registrars                                | 68        |
| 4.6.2 | Proxy servers                             | 69        |
| 4.6.3 | Redirect servers                          | 72        |
| 4.7   | SIP extensions                            | 72        |
| 4.7.1 | Querying for capabilities                 | 73        |
| 4.8   | Security                                  | 75        |
| 4.8.1 | Transport and network layer security      | 75        |
| 4.8.2 | Authentication                            | 75        |
| 4.8.3 | S/MIME                                    | 76        |
| 4.9   | Summary                                   | 77        |
| 5     | Session initiation in MRCP                | 79        |
| 5.1   | Introduction                              | 79        |
| 5.2   | Initiating the media session              | 79        |
| 5.3   | Initiating the control session            | 79        |
| 5.4   | Session initiation examples               | 81        |
| 5.4.1 | Single media resource                     | 81        |
| 5.4.2 | Adding and removing media resources       | 83        |
| 5.4.3 | Distributed media source/sink             | 88        |
| 5.5   | Locating media resource servers           | 91        |
| 5.5.1 | Requesting server capabilities            | 91        |
| 5.5.2 | Media resource brokers                    | 92        |

|                                                   |            |
|---------------------------------------------------|------------|
| 5.6 Security                                      | 94         |
| 5.7 Summary                                       | 94         |
| 6 The media session                               | 95         |
| 6.1 Media encoding                                | 95         |
| 6.1.1 Pulse code modulation                       | 95         |
| 6.1.2 Linear predictive coding                    | 97         |
| 6.2 Media transport                               | 102        |
| 6.2.1 Real-time protocol (RTP)                    | 102        |
| 6.2.2 DTMF                                        | 106        |
| 6.3 Security                                      | 107        |
| 6.4 Summary                                       | 109        |
| 7 The control session                             | 111        |
| 7.1 Message structure                             | 111        |
| 7.1.1 Request message                             | 113        |
| 7.1.2 Response message                            | 113        |
| 7.1.3 Event message                               | 114        |
| 7.1.4 Message bodies                              | 115        |
| 7.2 Generic methods                               | 117        |
| 7.3 Generic headers                               | 119        |
| 7.4 Security                                      | 126        |
| 7.5 Summary                                       | 126        |
| <b>PART III DATA REPRESENTATION FORMATS</b>       | <b>127</b> |
| 8 Speech Synthesis Markup Language (SSML)         | 129        |
| 8.1 Introduction                                  | 129        |
| 8.2 Document structure                            | 130        |
| 8.3 Recorded audio                                | 133        |
| 8.4 Pronunciation                                 | 134        |
| 8.4.1 Phonemic/phonetic content                   | 134        |
| 8.4.2 Substitution                                | 135        |
| 8.4.3 Interpreting text                           | 136        |
| 8.5 Prosody                                       | 136        |
| 8.5.1 Prosodic boundaries                         | 136        |
| 8.5.2 Emphasis                                    | 137        |
| 8.5.3 Speaking voice                              | 137        |
| 8.5.4 Prosodic control                            | 138        |
| 8.6 Markers                                       | 140        |
| 8.7 Metadata                                      | 141        |
| 8.8 Summary                                       | 142        |
| 9 Speech Recognition Grammar Specification (SRGS) | 143        |
| 9.1 Introduction                                  | 143        |
| 9.2 Document structure                            | 144        |
| 9.3 Rules, tokens and sequences                   | 146        |

|                                                           |                |
|-----------------------------------------------------------|----------------|
| 9.4 Alternatives                                          | 147            |
| 9.5 Rule references                                       | 148            |
| 9.5.1 Special rules                                       | 149            |
| 9.6 Repeats                                               | 150            |
| 9.7 DTMF grammars                                         | 151            |
| 9.8 Semantic interpretation                               | 152            |
| 9.8.1 Semantic literals                                   | 153            |
| 9.8.2 Semantic scripts                                    | 153            |
| 9.9 Summary                                               | 157            |
| <br>10 Natural Language Semantics Markup Language (NLSML) | <br>159        |
| 10.1 Introduction                                         | 159            |
| 10.2 Document structure                                   | 159            |
| 10.3 Speech recognition results                           | 160            |
| 10.3.1 Serialising semantic interpretation results        | 163            |
| 10.4 Voice enrolment results                              | 167            |
| 10.5 Speaker verification results                         | 169            |
| 10.6 Summary                                              | 173            |
| <br>11 Pronunciation Lexicon Specification (PLS)          | <br>175        |
| 11.1 Introduction                                         | 175            |
| 11.2 Document structure                                   | 176            |
| 11.3 Lexical entries                                      | 177            |
| 11.4 Abbreviations and acronyms                           | 178            |
| 11.5 Multiple orthographies                               | 179            |
| 11.6 Multiple pronunciations                              | 180            |
| 11.7 Summary                                              | 181            |
| <br><b>PART IV MEDIA RESOURCES</b>                        | <br><b>183</b> |
| <br>12 Speech synthesiser resource                        | <br>185        |
| 12.1 Overview                                             | 185            |
| 12.2 Methods                                              | 186            |
| 12.3 Events                                               | 199            |
| 12.4 Headers                                              | 201            |
| 12.5 Summary                                              | 206            |
| <br>13 Speech recogniser resource                         | <br>207        |
| 13.1 Overview                                             | 207            |
| 13.2 Recognition methods                                  | 212            |
| 13.3 Voice enrolment methods                              | 227            |
| 13.4 Events                                               | 233            |
| 13.5 Recognition headers                                  | 233            |
| 13.6 Voice enrolment headers                              | 244            |
| 13.7 Summary                                              | 247            |

|                                               |                                                    |            |
|-----------------------------------------------|----------------------------------------------------|------------|
| 14                                            | Recorder resource                                  | 249        |
| 14.1                                          | Overview                                           | 249        |
| 14.2                                          | Methods                                            | 249        |
| 14.3                                          | Events                                             | 255        |
| 14.4                                          | Headers                                            | 256        |
| 14.5                                          | Summary                                            | 260        |
| 15                                            | Speaker verification resource                      | 261        |
| 15.1                                          | Overview                                           | 261        |
| 15.2                                          | Methods                                            | 262        |
| 15.3                                          | Events                                             | 277        |
| 15.4                                          | Headers                                            | 277        |
| 15.5                                          | Summary                                            | 282        |
| <b>PART V PROGRAMMING SPEECH APPLICATIONS</b> |                                                    | <b>285</b> |
| 16                                            | <i>Voice Extensible Markup Language (VoiceXML)</i> | 287        |
| 16.1                                          | Introduction                                       | 287        |
| 16.2                                          | Document structure                                 | 288        |
| 16.2.1                                        | Applications and dialogs                           | 289        |
| 16.3                                          | Dialogs                                            | 290        |
| 16.3.1                                        | Forms                                              | 290        |
| 16.3.2                                        | Menus                                              | 293        |
| 16.3.3                                        | Mixed initiative dialogs                           | 294        |
| 16.4                                          | Media playback                                     | 296        |
| 16.5                                          | Media recording                                    | 297        |
| 16.6                                          | Speech and DTMF recognition                        | 299        |
| 16.6.1                                        | Specifying grammars                                | 299        |
| 16.6.2                                        | Grammar scope and activation                       | 300        |
| 16.6.3                                        | Configuring recognition settings                   | 301        |
| 16.6.4                                        | Processing recognition results                     | 301        |
| 16.7                                          | Flow control                                       | 306        |
| 16.7.1                                        | Executable content                                 | 306        |
| 16.7.2                                        | Variables, scopes and expressions                  | 306        |
| 16.7.3                                        | Document and dialog transitions                    | 308        |
| 16.7.4                                        | Event handling                                     | 312        |
| 16.8                                          | Resource fetching                                  | 314        |
| 16.9                                          | Call transfer                                      | 315        |
| 16.9.1                                        | Bridge                                             | 315        |
| 16.9.2                                        | Blind                                              | 317        |
| 16.9.3                                        | Consultation                                       | 317        |
| 16.10                                         | Summary                                            | 318        |
| 17                                            | VoiceXML -- MRCP interworking                      | 319        |
| 17.1                                          | Interworking fundamentals                          | 319        |
| 17.1.1                                        | Play prompts                                       | 320        |
| 17.1.2                                        | Play and recognise                                 | 321        |
| 17.1.3                                        | Record                                             | 324        |

|                          |     |
|--------------------------|-----|
| 17.2 Application example | 326 |
| 17.2.1 VoiceXML scripts  | 326 |
| 17.2.2 MRCP flows        | 333 |
| 17.3 Summary             | 341 |
| References               | 343 |
| Acronyms                 | 347 |
| Index                    | 349 |