MRCPv2在电信智能语音识别业务中的应用

陈茂国
(华为技术有限公司 江苏南京 210012)

摘 要：随着自然语音识别技术的成熟，智能语音识别业务将会在传统电信行业迎来大发展，MRCPv2标准协议使得语音识别能力的集成变得更加方便快捷。该文详细介绍了MRCPv2协议的系统架构和控制流程，总结了MRCPv2协议的使用规范，并且通过MRCPv2在语音识别中状态机变迁、关键方法、事件以及重要消息头的讲解，对一次典型语音识别业务中MRCPv2协议的应用进行了详细的阐述。

关键词：MRCPv2 智能语音识别 系统结构 控制机制 电信
中图分类号：TN912.34 **文献标识码**：A **文章编号**：1674-098X(2014)01(c)-0057-04

Applications of MRCPv2 in Telecommunications Intelligent Speech Recognition Service

CHEN Mao guo
(Huawei Technologies Co.,Ltd.,Nanjing 210012,China)

Abstract: With natural language recognition technology matures, intelligent speech recognition service will get great development in the traditional telecommunications industry, MRCPv2 standard protocols enables the integration of speech recognition capabilities becoming more convenient. This paper describes the system architecture and control process of MRCPv2, summarizes MRCPv2 Use Agreement. Furthermore, by the detailed description of MRCPv2's state machine mechanism, key method, events and important message headers in speech recognition, it explains clearly the application of MRCPv2 in a typical speech recognition service.

Key words: MRCPv2 Intelligent Speech Recognition System Architecture Control Mechanism Telecommunication

1 MRCPv2协议简介

媒体资源控制协议(Media Resource Control Protocol，MRCP)是一种基于TCP/IP的通讯协议，用于客户端向媒体资源服务器请求提供各种媒体资源服务。此协议最初是由Cisco、Nuance等公司联合开发，由IETF作为Internet草案发布，经过不断的更新，目前最新的版本为RFC6787[1]，可以支持的媒体资源业务包括文语转换(Text to Speech，TTS)、自动语音识别(Automatic Speech Recognition，ASR)、录音(Recording)、声纹识别(Voiceprint Recognition，VPR)。

2 MRCPv2系统结构及协议控制

MRCP协议本身不是独立的，它不仅仅依赖于TCP/IP协议，还依赖于SIP、SDP、RTP、RTCP、RTSP等协议。

其系统结构如图1所示。[1]

(1) 控制面：它通过SIP协议在客户端(MRCP Client)和服务器(MRCP Server)之间建立和管理会话(注：MRCPv1就使用RTSP协议完成控制，MRCPv2改为SIP协议)。

(2) 媒体面：它通过SDP交换媒体能力以及通过RTP协议完成媒体的承载交换。

(3) 业务面：它通过MRCP协议来控制完成媒体资源服务的相关请求、响应和事件的传递，从而为客户端提供所需要的媒体资源服务。

其协议控制机制如图2所示。

(1) SIP协商过程中，MRCP Client在INVITE消息中携带自身用于传递MRCP协议以及RTP语音流的SDP(IP地址、端口号)[2-3]。

(2) 协商成功返回的200消息中会带上MRCP Server侧的SDP；

(3) MRCP Client发起TCP连接创建，并且通过在TCP连接上发送MRCP协议消息控制MRCP Server分配的资源。

(4) MRCP Client/Server通过在RTP连接上传输语音数据从而实现媒体资源业务。

(5) 当业务应用结束时，终止SIP会话的同时，还需要关闭TCP和RTP连接。

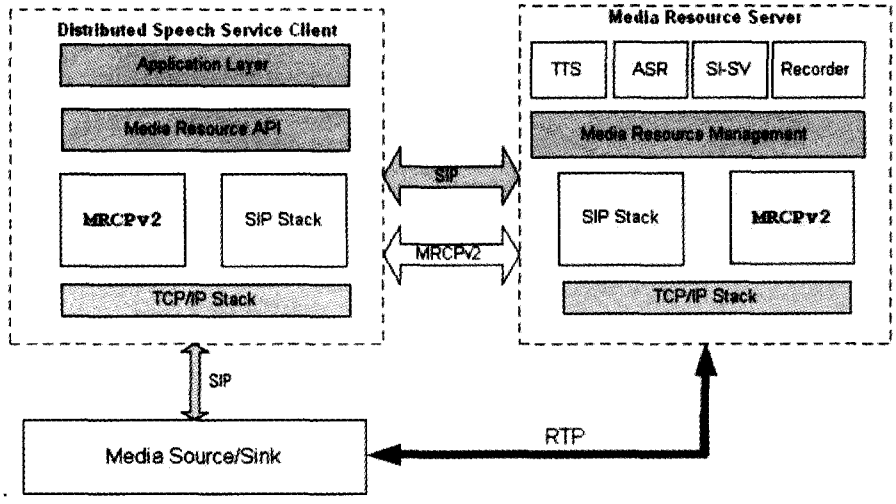

图1 MRCPv2系统结构

MRCPv2的使用规范可以总结如下：

(1) MRCP Client通过SIP&SDP建立与MRCP Server的MRCP控制通道(使用MRCP通道ID进行唯一标识，MRCP Server回200消息时通过a=channel属性指定)。

(2) 可以通过SIP的Re-INVITE消息添加或者删除一个会话中的MRCP控制通道，所以一个会话可以拥有多个MRCP控制通道(比如一个会话可以同时拥有ASR&TTS通道)。

(3) 多个MRCP控制通道可以共享同一个TCP连接。

(4) 一个MRCP消息只能携带一个MRCP通道ID。

(5) MRCP控制消息不能更改SIP会话的状态。

(6) 由于MRCP不保证传输的可靠性，所以必须使用TCP来保证其传输。

3 语音识别技术及其在电信智能语音识别业务中的应用

自动语音识别技术（Automatic Speech Recognition，ASR）是一种将人的语音转换为文本的技术，其广泛应用于语音通讯系统、声控电话交换、数据查询、订票系统、电信银行客服、计算机控制、工业控制等领域。

通常，我们说的语音识别可以分为固定词识别以及自然语音识别[4-5]，固定词语音识别只能识别已经指明的固定短语或词，而且用户也只能说这些固定的词，否则无法识别，而自然语音识别可以识别用户随意说的短语或者句子，很显然自然语音识别更易用，其技术难度也更大；

近几年来，自然语音识别相关的技术随着移动互联网的发展迎来了迅猛的发展。在Google引领下，互联网、通信公司纷纷把自然语音识别作为重要研究方向。

美国市场调查咨询公司Gartner于2013年发布的新兴技术成熟度曲线显示，语音识别技术已经走向成熟，在未来2～5年之内将会有大幅度的利用，而自然语音问答目前处于技术期望过热区，预计在未来的5～10年会有大幅度的利用，自然语音问答技术中就使用到自然语音识别。

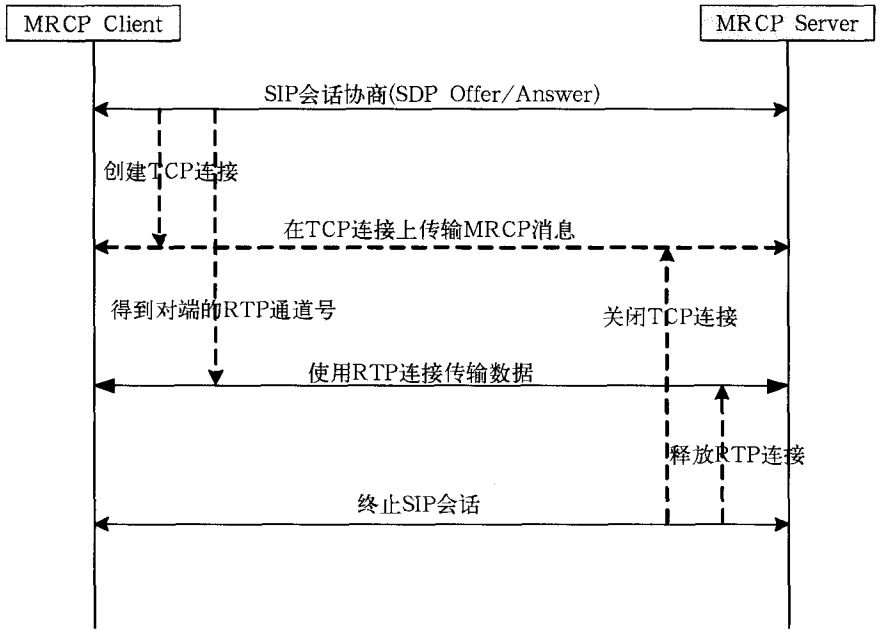

图2 MRCPv2协议控制

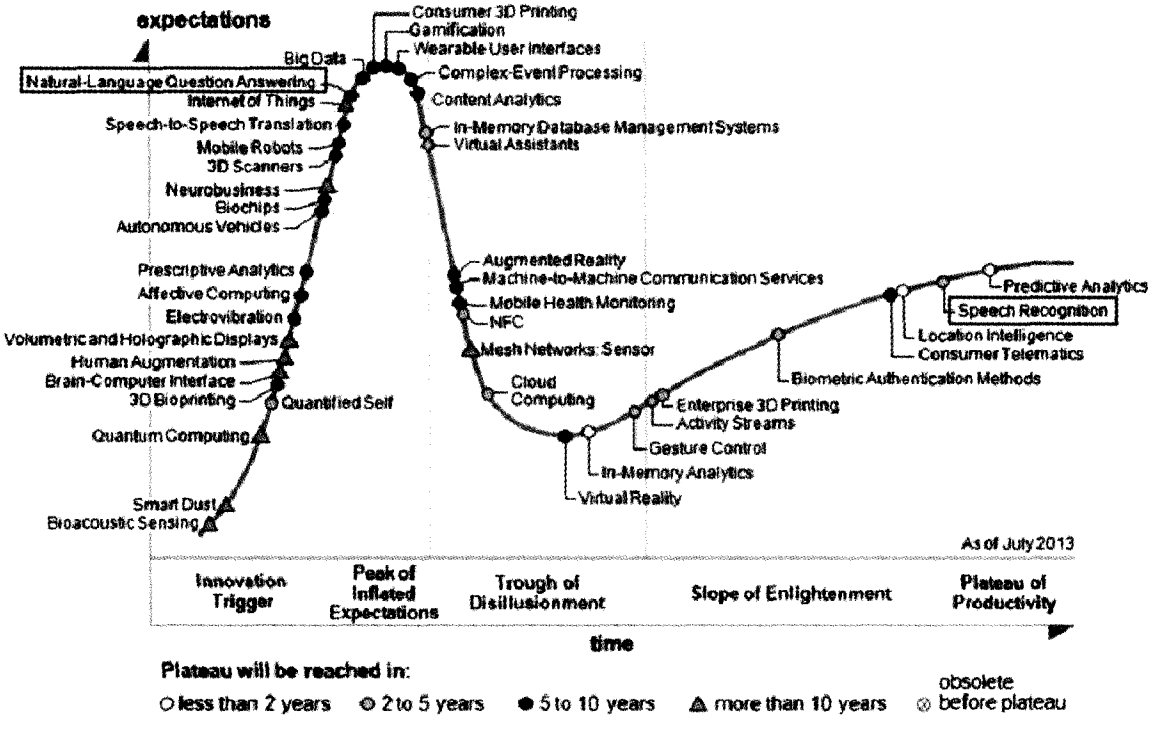

图3 2013新兴技术成熟度曲线

在电信领域，语音识别技术应用多年来一直停留在固定词识别上，基本限定在简单IVR领域，因为限制了用户的语音输入范围，易用性和可靠性受限导致应用实际并不广泛，从目前国内各大运营商的客服电话就很容易发现，采用按键式交互的IVR仍然是主流。

随着近几年语音识别技术的不断发展，自然语音识别技术也逐渐成熟，而且在移动互联网等可靠性要求不是太高的领域得到广泛的应用，iPhone的Siri、QQ的语音输入、Google的语音翻译、科大和移动合作的灵犀等智能语音识别业务都广泛应用到此技术。

拥有海量一手语音数据的电信行业也因为自然语音识别技术的成熟，智能语音识别业务将会迎来新的发展机遇。

4 MRCPv2协议在电信智能语音识别业务中的应用

由于识别技术的专一性，在电信领域，控制着语音接入的电信设备制造商，很少拥有扎实的语音识别技术，而提供语音识别技术的厂家很多。以前各电信设备集成商必须针对不同的语音识别厂家提供的API接口进行专门的集成开发，不同识别引擎的接口各不相同，从而导致了集成过程的复杂性和局限性。而利用MRCP协议提供的标准接口，电信设备集成商们不必再针对特定的识别引擎进行开发，而只需要满足MRCP协议即可与多个不同厂商的识别引擎对接。这样就为各种语音应用开发提供了更加灵活的选择，并有效地降低业务开发周期和成本。正

是由于具有以上优势，MRCP协议在推出以后得到了国外各电信设备制造商和语音识别提供商的广泛支持[6]，各电信设备制造商提供MRCP Client，语音识别提供商提供MRCP Server，通过对接完成语音识别业务。

MRCPv2为语音识别业务提供除了公共的SIP、SDP、RTP配合机制，公共的方法、消息头和事件之外，还包含如下两个部分：

(1) 语音识别业务中的状态变迁机制；

(2) 语音识别业务中的方法、事件以及配套的消息头、参数。

4.1 MRCPv2语音识别业务中的状态变迁机制介绍(图4)

在MRCPv2定义的语音识别应用中，MRCP Client和MRCP Server必须遵循如上图所示的状态机变迁机制，只有空闲、识别中、识别完三个状态。状态的变迁依靠方法和事件的驱动：

(1) 通过RECOGNIZE方法触发进入识别中状态；

(2) 通过RECOGNITION-COMPLETE事件触发进入识别完态；

(3) 通过STOP方法触发进入空闲态；

4.2 MRCPv2定义的语音识别应用中的方法、事件及重要消息头

支持语音识别业务的方法和事件主要如下：

(1) RECOGNIZE方法：启动识别命令，携带的主要消息头有No-Input-Timeout、Recognition-Timeout、Speech-Complete-Timeout、Start-Input-Timers、Confidence-Threshold，其含义分别如下：

No-Input-Timeout：无话超时时间，单位为毫秒，用于定义MRCP server启动识别后允许用户无声音输入的最大时长；

Recognition-Timeout：识别超时时间，单位为毫秒，用于定义MRCP server启动识别后允许返回识别结果的最大时长；

Speech-Complete-Timeout：说话完检测超时时间，单位为毫秒，用于定义MRCP Server判断用户一句话已说完的静默时长；

Start-Input-Timers：是否立即启动无话超时定时器，为"true"时立即启动，通常启动识别时同时伴随有提示音的情况下，可以置为"false"，即让MRCP Server暂时不要启动无话超时定时器；

Confidence-Threshold：识别置信度门槛，用于定义返回识别结果时必须满足的最小置信度；

典型的RECOGNIZE方法示例如下：

```
MRCP/2.0 ... RECOGNIZE 2
Channel-Identifier: 2ce5baab46401041@speechrecog
Content-Type: text/uri-list
Cancel-If-Queue: false
No-Input-Timeout: 10000
Recognition-Timeout: 15000
Speech-Complete-Timeout: 800
Start-Input-Timers: true
Confidence-Threshold: 0.0
Content-Length: 33

file://C:\tmp\analytics1.grxml
//一般是语法文件的URI
```

(2) START-INPUT-TIMERS方法：用于启动无话超时定时器，一般情况下，当RECOGNIZE方法中没有立即启动无话超时定时器的时候，通过这个方法通知MRCP Server启动；

典型的START-INPUT-TIMERS方法示例如下：

```
MRCP/2.0 ... START-INPUT-TIMERS 543260
Channel-Identifier:32AECB23433801@speakverify
```

(3) START-OF-INPUT事件：MRCP Server用于通知MRCP Client已经有用户语音输入，此时两边都会停止用户无话超时定时器。

典型的START-OF-INPUT事件示例如下：

```
MRCP/2.0 ... START-OF-INPUT 543260 IN-PROGRESS
Channel-Identifier:32AECB23433801@speechrecog
```

(4) STOP方法：用于停止识别，一般应用于启动识别之后中途停止识别。

典型的STOP方法示例如下：

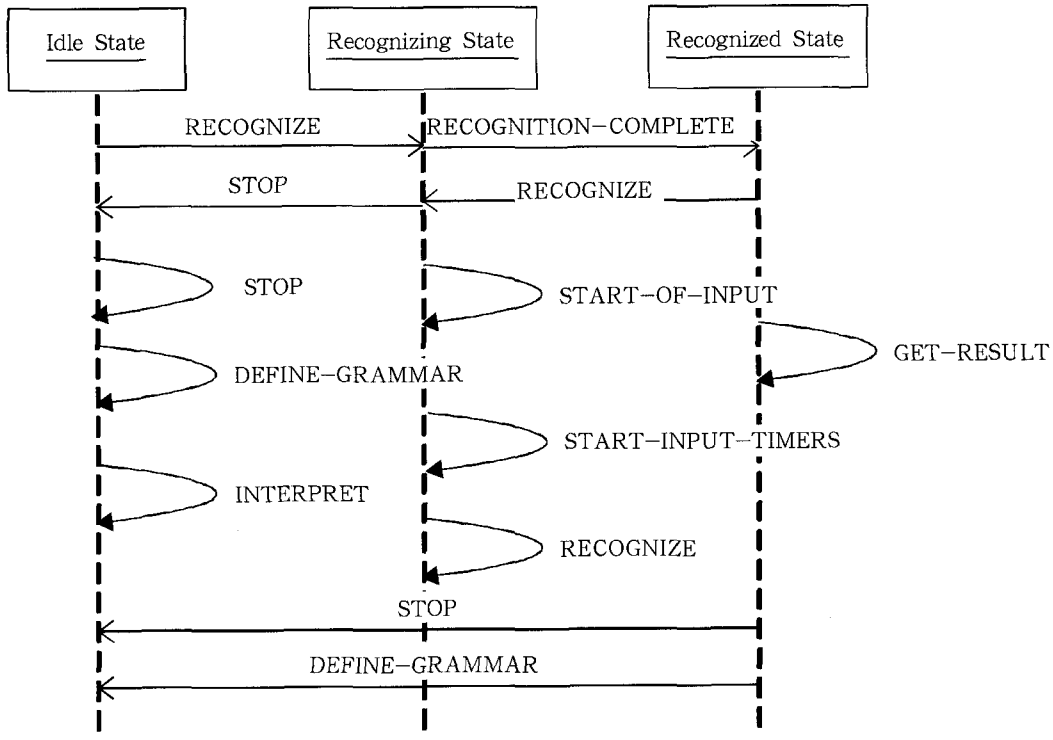

图4 语音识别中MRCP状态变迁

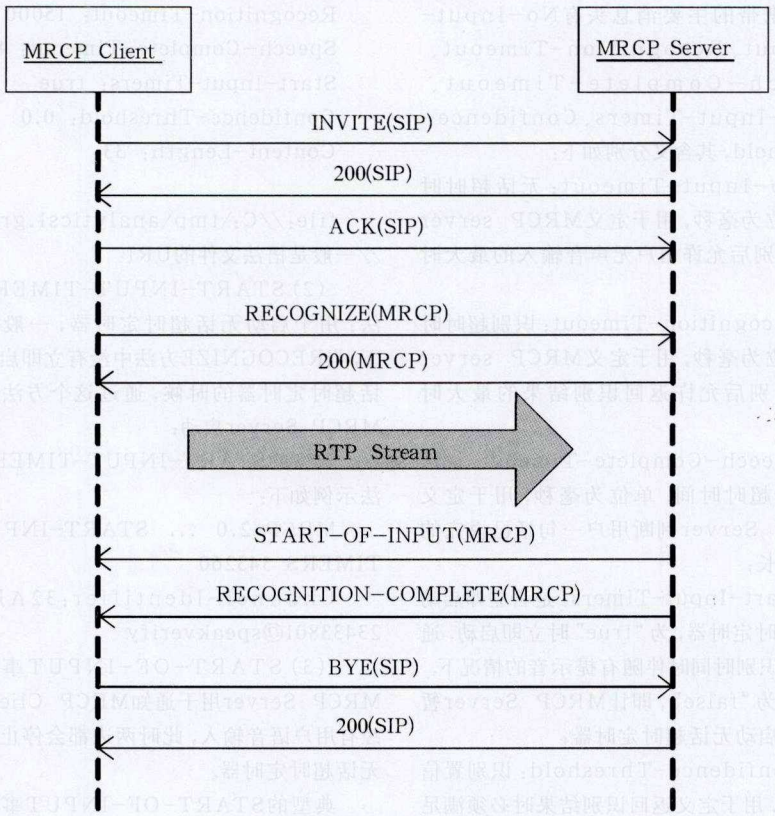

图5 一次完整语音识别消息交互

MRCP/2.0 ... STOP 314178

Channel-Identifier: 32AECB23433801@speakverify

(5) RECOGNITION-COMPLETE事件：应用于MRCP Server向MRCP Client反馈识别完成结果，携带的主要消息头有Completion-cause，识别成功情况下消息体里面还有识别的结果。

典型的RECOGNITION-COMPLETE事件示例如下：

MRCP/2.0 ... RECOGNITION-COMPLETE 2 COMPLETE

Channel-Identifier:2ce5baab46401041@speechrecog

Completion-Cause:000 success //描述识别完成原因

Content-Type: application/nlsml+xml

Content-Length: ...

```
<?xml version="1.0" encoding="utf-8"?>
<result>
<interpretation grammar="file://C:\tmp\callrouting1.grxml" confidence="97">          //置信度为97%
<instance>
<digit>one</digit>          //这里描述识别结果为one
</instance>
<input mode="speech">one</input>
</interpretation>
</result>
```

4.3 一次语音识别业务中完整MRCP交互

一次完整的语音识别业务交互如图5所示：

(1) MRCP Client发送INVITE消息给MRCP Server请求建立会话，携带MRCP Client侧的SDP；

(2) MRCP Server回复200表示请求已经成功接受处理，携带MRCP Server侧的SDP；

(3) MRCP Client随后发送ACK消息证实200消息已经收到，至此一个SIP会话成功建立；

(4) MRCP Client发送RECOGNIZE消息给MRCP Server，请求语音识别，按照MRCP协议规定的格式携带相关的语音识别控制消息头，并且指定语法文件路径；

(5) MRCP Server接收RECOGNIZE请求，编译语法文件，回复200消息给MRCP Client，此时两侧进入识别中状态；

(6) MRCP Client此时开始根据之前协商好的SDP，开始源源不断的发送RTP语音流给MRCP Server；

(7) MRCP Server接收RTP语音流，当检测到用户开始说话时，发送START-OF-INPUT事件；

(8) 当MRCP Server根据语法文件定义得到识别结果时，通过RECOGNITION-COMPLETE事件返回识别结果，两侧进入识别完状态；

(9) MRCP Client发送BYE消息给MRCP Server结束会话；

(10) MRCP Server发送200消息给MRCP Client确认结束；

MRCP Client通过上述消息交互获得MRCP Server提供的一次完整语音识别能力。

5 MRCPv2在电信实时智能语音识别业务中的应用展望

当前，MRCPv2协议已经能够很好的解决单次语音识别问题，各大电信运营商正火热上线的智能语音导航、机器人客服等业务都基于MRCPv2协议，但是这些都是IVR性质，其特定都是要识别的语音内容不长。在人工业务辅助识别等大量连续识别场景中应用仍然受限，比如说话内容实时回显，此时需要完成不间断的语音识别结果上报，而当前的MRCPv2协议只支持每次上报一个结果，所以需要进一步扩展MRCPv2协议才能完成。

参考文献

[1] MRCPv2 RFC 6787: Media Resource Control Protocol Version 2.

[2] SDP RFC 2327: Session Description Protocol.

[3] SIP RFC 3261: Session Initiation Protocol.

[4] 薛德黔.交互式自然口语语音识别关键技术[J].计算机应用,2002,22(7):45-47.

[5] 冯俊兰,杜利民.自然口语语音识别研究概况[J].电子商务,1999(9):3-7.

[6] 史俊波,詹舒波.MRCPv2协议及其在分布式语音资源解决方案中的应用,2010.

[7] Gartner.Hype Cycle for Emerging Technologies 2013[R].2013.