基于智能语音的人机交互方案设计与实现*

文／深圳数字电视国家工程实验室股份有限公司　马亚飞
康佳集团股份有限公司　邓忠平
深圳数字电视国家工程实验室股份有限公司　黄华松　常林

摘要：智能电视作为家庭娱乐的中心，正逐步成为电脑、手机之外的第三种信息访问沟通平台，其智能化的定义决定了智能电视的复杂性和多样性，智能化定义中最为关键的环节就是人机交互。从某种意义上讲，智能电视人机交互方式决定着智能电视方案的成败与否。本文提出了一种基于Android智能电视的语音交互系统，该系统使得用户可以通过语音信息的输入来对智能电视进行控制操作，极大地降低了用户操作的复杂性，丰富了智能电视的人机交互方式。

关键字：智能电视　人机交互　语音识别

1 引言

随着社会的不断发展和互联网技术的不断更新，传统的电视已经不能满足人们对于高智能、高水平电视的需求。Android作为一种优异的开源操作系统解决方案，慢慢地从手机、平板电脑领域推广到智能电视领域，如何利用其开发出功能多样、操作便捷、稳定的智能电视，成为了广大电视厂商的研发热点。

一般来说，智能电视有着距离远、光线暗、用户操作心态比较被动等特点，因此不能生搬硬套传统的人机交互方案。针对智能电视的这些特点，目前业界提出了多种多样的人机交互方式，其中通过语音技术来对智能电视进行操控，成为智能电视人机交互的一大突破点。采用语音的人机交互方式有着更趋于自然，符合用户习惯，输入信息方便、迅速等优点。

本文依据智能语音技术提出了一种基于智能语音识别的智能电视人机交互解决方案，该方案将采集到的用户语音输入数据，通过模数转换后发送至云端进行解析处理，然后返回对应的控制命令到智能电视再进行处理。

2 系统设计架构

鉴于智能语音识别引擎中本地库设计的复杂性和不完备性，该交互系统采用云端的智能语音识别引擎来处理采集到的数据，系统设计架构如图1所示。从使用层次上分为3层：语音数据采集层、智能电视系统层、语音数据解析处理层。通过声音采样设备，如麦克风，进行语音数据的录入；然后传递给智能电视系统层进行编码；再将编码后的语音数据经由Internet传输至位于云端的智能语音识别引擎进行语音数据的解析工作；接着在智能语音识别引擎解析处理完成后，将语音数据中包含的文字信息解析出来，并返回给智能电视系统层；最后智能电视系统层对返回的信息映射成控制命令或者相关信息输入来对智能电视进行操控。

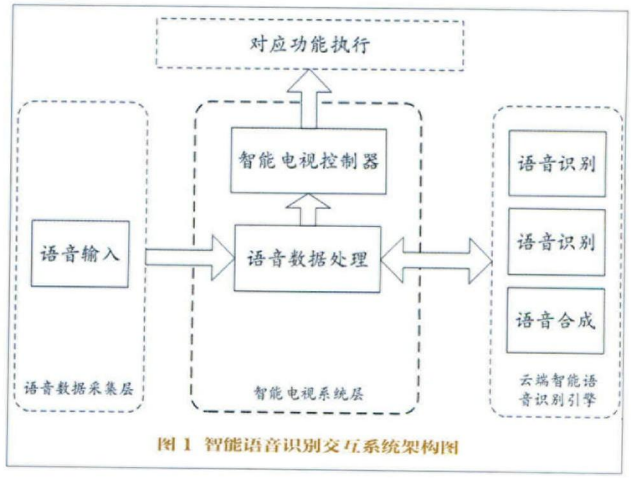

图1 智能语音识别交互系统架构图

* 本论文所属项目：电子信息发展基金资助项目——新型人机交互智能电视机研发与产业化

责任编辑：李玉薇　邮箱：liyuwei@abrs.gov.cn

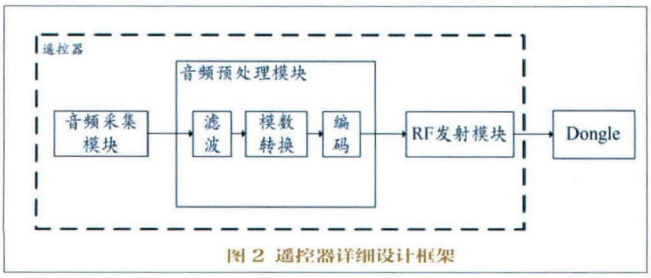

图 2 遥控器详细设计框架

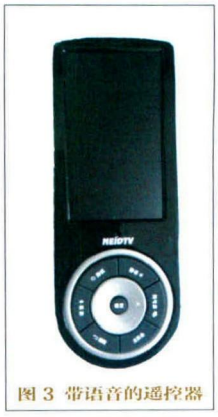
图 3 带语音的遥控器

2.1 语音数据采集

语音数据的采集是通过在遥控器上添加一个语音模块来完成的，其在进行语音输入时，将麦克风输入的模拟语音信号进行预处理，主要包括预滤波、采样、量化、模数转换、编码、断点侦测等；然后在遥控器添加一个 2.4GHz 无线传输模块用来传输预处理后的语音信号；最后通过电视端的 2.4GHz 无线 Dongle 模块来接收语音信号并进行相应处理。遥控器端的设计框架图如图 2 所示。根据遥控器端的设计框架图设计出遥控器的最终版本如图 3 所示。

2.2 语音数据接收 / 发送

智能电视语音数据的收发工作在语音数据采集完成后进行。当电视上的 Dongle 接收到来自遥控器端的 2.4GHz 模块中预处理后的音频模拟信号后，对其解码，转换格式，最后通过 Internet 将数据传输给云端的语音识别引擎进行解析。

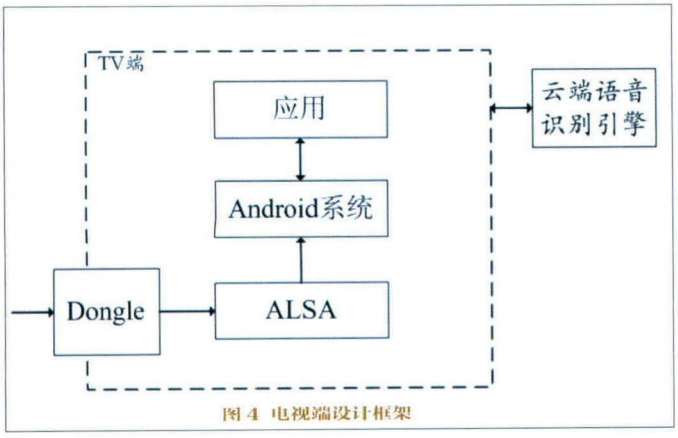

图 4 电视端设计框架

云端的语音识别引擎对数据进行识别解析后，将识别后的数据同样以 Internet 方式返回给电视端，电视端再进行相应的功能处理。使用云端智能识别引擎可以有效降低智能电视的硬件成本，提高识别精度，从而给用户一个良好的体验。电视端的设计框架如图 4 所示。

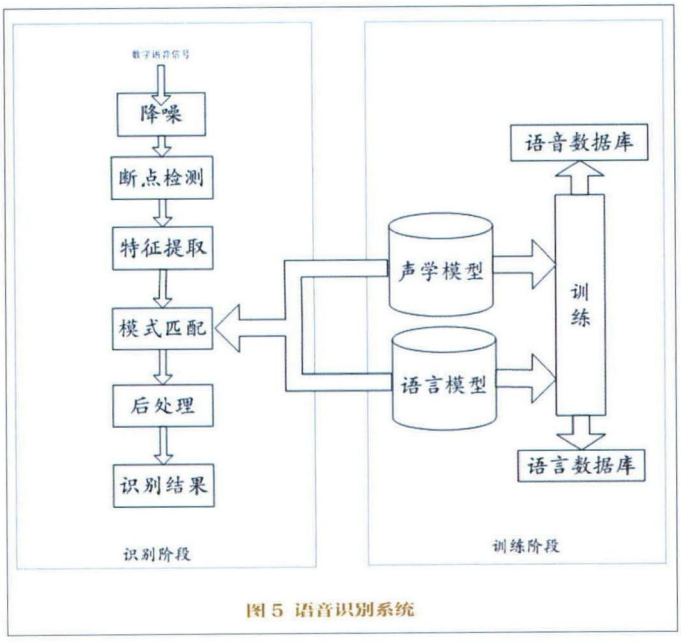

图 5 语音识别系统

3 语音识别

3.1 语音技术

语音技术一般主要包括语音合成和语音识别技术。其中语音合成也就是说将文本内容转换成语音信息并朗读出来。这一过程涉及多个学科，它的关键在于如何将文本信息转换成可以听的声音信息，在智能电视人机交互中，这一技术可以实现让智能电视“开口”讲话的功能。而语音识别又称自动语音识别（ASR），它主要是通过模数转换将机器“听”到的自然语言转换成相应的文本或者命令，从而进行更进一步的操作。语音识别系统本质上是一种多维模式识别系统，它的识别过程与人对语音识别的处理过程基本上是一致的。语音识别一般分为两个过程，一个是系统的训练过程，一个是系统的识别过程，它们有各自不同的任务要求，如图 5 所示。

在语音识别系统的训练阶段建立了对应的模式库，库中每一个词都对应一个参考模式，它由这个词重复发音多遍，再经特征提取和某种训练得到。而识别阶段则是整个系统的核心，它的作用是按照一定的准则求取待测语音特征参数、语音信息与模式库中的相应模型的匹配度，匹配度最高的就是识别结果。这种训练和匹配的方法有多种，目前较为常见的方法主要有动态时间规整（DTW）、隐马尔可夫链（HMM）、人工神经网络（ANN）等。

从语音识别的过程来分析，语音识别的效果受处理速度和存储容量的限制，尤其是在智能电视硬件资源不够丰富的平台下，语音识别通常要求

责任编辑：李玉薇 邮箱：liyuwei@abrs.gov.cn

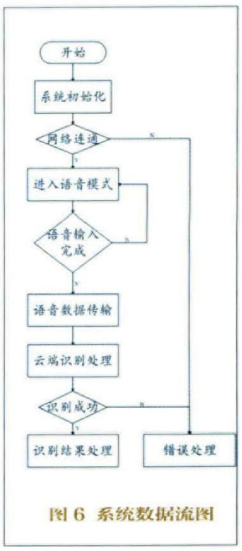

图 6 系统数据流图

有一定的句式，这样就容易造成识别性能弱，交互固定等不好的用户体验。

3.2 云端语音识别引擎

云计算是一种基于互联网的计算方式，通过这种方式，共享的软硬件资源和信息可以按照需求提供给计算机和其他设备。这对用户而言，降低了用户端的负载和成本，赋予了用户前所未有的计算能力。

云计算服务应该具备以下几个特征：随需自助服务，随时随地用任何网络设备访问，多人共享资源池，快速重新部署灵活，可被监控与测量的服务，基于虚拟化技术快速获得资源，减少用户端的处理负担，降低用户对专业知识的依赖等。

随着互联网的快速发展，从多个渠道获取的大量文本或语音方面的资料，为语音识别中的语言模型和声学模型的训练提供了丰富的资源，使得构建通用的大规模语音模型和声学模型成为可能。

基于云计算的这些特征，我们的智能语音人机交互方案中的语音识别使用云端的语音识别引擎进行识别解析处理，这样可以通过各种复杂的算法来提高语音识别的精度，将复杂的、开放式的语音输入交给云端进行处理，给用户带来了新的体验。

目前在国外的语音识别引擎以 Nuance 和谷歌为主，而国内方面，科大讯飞、云知声、小 i 机器人、捷通华声等都有最新的语音识别技术。鉴于云端语音识别引擎的接入便捷性和易扩展性，本文的智能语音人机交互方案中的云端使用科大讯飞的语音识别引擎。

3.3 智能语音人机交互

整个人机交互的方案采用用户主动式语音输入。用户按下遥控器上的语音按键之后进入语音模式，语音模块采集用户的声音经过预处理交由 RF 模块进行传输，同时在进入语音模式之前电视端需要检查网络是否连通，在确定网络畅通之后电视端系统将遥控器端采集完成的语音数据通过 HTTP 的 POST 方法发送至云端进行处理，然后对云端返回的数据进行解析并显示给用户或者进行相关命令控制操作。该方案的数据处理流程图如图 6 所示。

4 实验测试

语音识别过程中存在多种问题，语音模式不仅对不同的说话人不同，对同一说话人也是不同的，例如一个说话人在随意说话和认真说话时语音信息是不同的，一个人的说话随着时间的变化而变化；说话者在讲话时，不同的词可能听起来是相似的；环境噪声和干扰对识别也有着严重的影响。

综合考虑这些因素，对本系统进行测试时，选取一男一女分别对系统进行测试。为了降低一个人说话在不同时刻不同效果的影响，首先建立一个含有 30 个语音命令的模板库，每个命令词汇长度不定，然后对每一个命令词汇进行录音并分别存储，最后在相同的环境下（安静和有噪声）进行语音的识别，统计并计算男、女的平均识别准确率。此外，针对不同语速对识别也会造成影响的事实，统计了在相同环境下（安静）的不同语速识别的准确率，识别结果如表 1 所示。

表 1 系统性能测试结果

测试人	安静	有噪声	慢速	中速	快速
男	90.2%	76.3%	81.4%	90.2%	78.7%
女	90.7%	75.9%	82.1%	90.7%	79.3%

从这两个测试来看，系统的识别率达到了预期的目标，基本可以满足智能电视人机交互对性能的需求。

5 结语

本文介绍了一种新型的智能电视人机交互方案，它采用 2.4GHz 无线传输模块来传输语音数据及云端智能语音识别引擎来进行语音的识别操作，并在 Android 平台上进行了实现，充分利用了网络资源，降低了用户终端设备的成本。通过测试表明该系统符合实时性的要求，系统稳定性能好，更有利于智能电视的人机交互实施。RTI

参考文献

[1] 谢世海，刘苏. 浅谈智能电视人机交互方式 [J]. 科技信息，2013（3）.
[2] 吴进强，苏凯雄. 基于智能电视的语音识别系统的设计与实现 [J]. 电视技术，2013,37（10）：27-30.
[3] 吴佳兴，李爱国. 基于云计算的智能家居系统 [J]. 计算机应用与软件，2013,30（7）：240-243.