**大连理工大学大学生创新创业训练计划**

**项目开题报告**

|  |  |
| --- | --- |
| **项 目 编 号：** | 2019101410901010686 |
| **项 目 名 称：** | 基于大数据技术的信息传输优化与语音识别 |
| **项 目 级 别：** | 校级 |
| **项目负责人：** | 张爽 |
| **项 目 类 型：** | ☑创新训练 创业训练 创业实践 |
| **指 导 教 师：** | 杨南海 |
| **所在学部学院：** | 软件学院 |

**大 连 理 工 大 学**

**2019 年 3 月**

目录

一．项目来源及研究的背景和意义3

1.1项目来源 3

1.2研究背景及意义3

1.2.1视频字幕实时生成 3

1.2.2网络传输协议优化 3

二．项目研究内容4

三．已完成的前期研究工作和成果4

3.1视频实时字幕生成原理及模型准备4

3.2网络传输协议优化改进 7

3.2.1流式输出协议研究7

3.2.2卡尔曼滤波器在网络状况分析中的应用8

3.2.3基于RTCP反馈的网络拥塞控制机制9

四．项目研究方案、进度安排及预期目标9

4.1项目研究方案 9

4.2项目进度安排9

4.3项目预期目标9

五．预计完成过程中可能遇到的困难及解决措施 10

六．参考文献 10

七．经费使用预算12

**一．项目来源及研究的背景和意义**

**1.1项目来源**

本项目来自学生自选，学生的积累和兴趣。

**1.2研究背景及意义**

**1.2.1视频字幕实时生成**

视频翻译在我们的日常生活中具有重大意义。一方面它可以更为清晰直接地向本土人民传递所要传递的信息，另一方面它又可以促进跨国之间的语言文化交流。但是当前市场上并不是所有视频都带有字幕翻译，这就给一些听众极其不好的感受，例如听力有障碍的人群，语言不通的人群等等。另外，人工为视频添加字幕也是一笔很大的劳动力开销[3]。

谷歌在2009年为Youtube上线了自动字幕系统（Automatic Captioning System），借助人工智能技术，在视频里实时生成字幕。截至目前，YouTube宣布已经有10亿多条视频采用了该系统生成的字幕，而全球日活跃用户也达1500万。反观国内主流视频网站，至今尚未推出过类似的自动字幕系统，翻译外国语言影视资源的任务一直是由民间字幕组成员来承担，在翻译准确度，版权归属等问题上一直存在不小的争议。2011年《电子学报》的一篇论文称，以《新闻联播》为语料进行训练，实现了音频提取、分类与切分[2]、说话人识别、连续语音识别、视频文件播放、文本字幕自动生成等功能，从而实现“全自动中文新闻字幕生成”，但此论文中并未提及是否可以做到“实时”翻译，而且新闻联播的视频背景噪声较小，对于一般的声音嘈杂的视频来说没有普遍应用性。

**1.2.2 网络传输协议优化**

若做到视频实时字幕的生成，网络传输速度必是我们要考虑的重要因素之一。若网络状况不好，字幕数据包时常丢失或者字幕生成不及时同样会给听众带来极差体验。当前在网络通信中的视频数据在传输层普遍使用的是UDP协议，其在数据传输过程中延迟小、数据传输效率高，可以保证网络中视频流资源最大化交付。但由于不对传送数据包进行可靠性保证，故很容易导致客户端资源丢包，造成视频在播放时出现延迟抖动等现象，降低传输视频的质量。相比之下，由于TCP协议数据超时重传机制的存在，导致信道传播大量数据（视频流）引起丢包时，发送端会重传该数据包，延误了其他数据包的传输造成网络延迟，进而给用户带来糟糕的收视体验。

**二．项目研究内容**

2.1实时生成视频字幕，解决长期以来依靠字幕组人工翻译视频的相关问题，也给用户尤其是“特殊用户”带来良好的收视体验。

2.2为实现视频字幕的实时性，对比TCP和UDP两种协议，采用两种协议折中的方式—RTCP协议，通过实时分析网络传输的数据，预测数据包是否可能会丢失，从而决定是否对数据包重传，优化传输性能，缓解在视频收看高峰期的网络拥塞问题，尽最大可能确保视频及字幕的实时跟进。

**三．已完成的前期研究工作和成果**

**3.1视频实时字幕生成原理及模型准备**

（1）嘈杂视频中的人声分离[6]

利用卷积神经网络[5]对视频进行处理。输入嘈杂背景干扰下的包含单人或多人语音信息的视频，输出是将输入视频的音轨分解成纯净的音轨，并对应上相应的说话者。

（2）压缩音频文件转非压缩波形文件

常见的mp3等格式均为压缩格式，必须转成非压缩的纯波形文件来处理，如wav文件。Wav文件中存储的除文件头外，就是声音波形的一个个点。下图为波形示例。



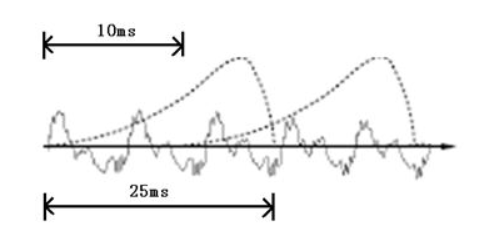
（3）VAD操作（静音切除）

在开始语音识别之前, 有时需要把首尾端的静音切除, 降低对后续步骤造成的干扰。这个静音切除的操作一般称为 VAD ,需要用到信号处理技术。

（4）对声音分帧

要对声音进行分析,需要对声音分帧,也就是把声音切开成一小段一小段, 每小段

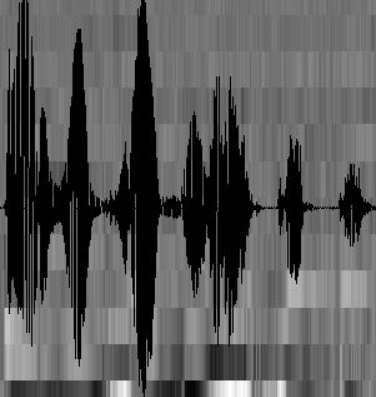
称为一帧。分帧操作使用移动窗函数来实现。帧与帧之间一般是有交叠的,就像下图这样:

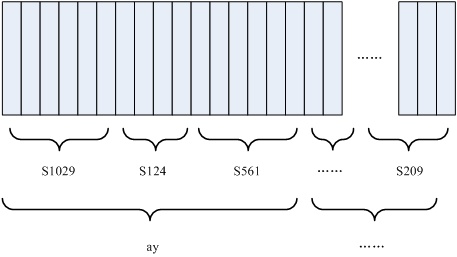


图中,每帧的长度为 25毫秒,每两帧之间有 25-10=15毫秒的交叠。称之为以帧长25ms、帧移10ms分帧。分帧后,语音就变成了很多小段。但波形在时域上几乎没有描述能力, 因此必须将波形作变换。

（5）对波形做变换

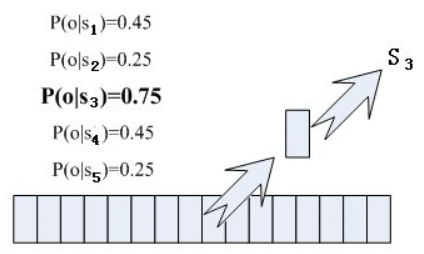
常见的一种变换方法是提取MFCC特征,根据人耳的生理特性,把每一帧波形变成一个多维向量,这个向量包含了这帧语音的内容信息。这个过程叫做声学特征提取。至此,声音就成了一个M行(假设声学特征是M维)N列的一个矩阵,称之为观察序列,这里N为总帧数。观察序列如下图所示,图中,每一帧都用一个M维的向量表示,色块的颜色深浅表示向量值的大小。

（6）将帧识别为状态

单词的发音由音素构成。项目中涉及的英语音素集采用卡内基梅隆大学的一套39个音素构成的集合。汉语采用声母和韵母作为音素集。状态是比音素更为细致的语音单元。通常把一个音素划分为3个状态。如下图所示：

图中, 每个小竖条代表一帧,若干帧语音对应一个状态, 每三个状态组合成一个音素,若干个音素组合成一个单词。知道每帧语音对应的状态,便可得到语音识别的结果。

要确定每帧音素对应的状态，就要看帧对应哪个状态的概率最大，那么帧就属于哪个状态。比如下图，该帧在状态S3的概率最大，所以属于S3

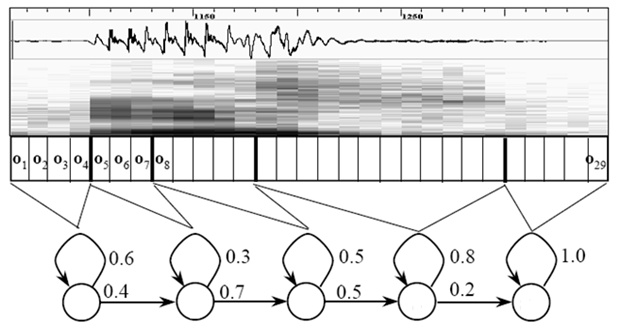


用到的概率，从“声学模型”中提取。声学模型中存了许多参数，通过参数得知帧和状态对应的概率。可以通过使用大量的语音数据进行分析训练，获取模型参数。

（7）状态组合为音素

使用HMM(隐马尔可夫模型)，解决相邻帧状态差异问题[4]。HMM可以构建一个状态网络，并从状态网络中寻找与声音最匹配的路径。这样，结果被限制在预先设定的网络中。这个网络需要足够大，要包含任意文本的路径。但不能过大，会造成识别准确率降低。所以要选择大小适当的网络。

（8）音素组合成单词

 在状态网络中搜索最佳路径，语音对应这条路径的概率最大，称为“解码”。采用Viterbi(动态规划剪枝)算法，寻找全局最优路径。

图中的累计概率由三部分组成：

1.观察概率：每帧和每个状态对应的概率

2.转移概率：每个状态到自身或者转移到下个状态的概率

3.语言概率：根据语言统计规律得到的概率

前两种概率从声学模型中获取，语言概率从语言模型中获取。语言模型通过大量文本训练得到，可利用某种语言本身的规律提升识别的准确率。

**3.2网络传输协议优化改进**

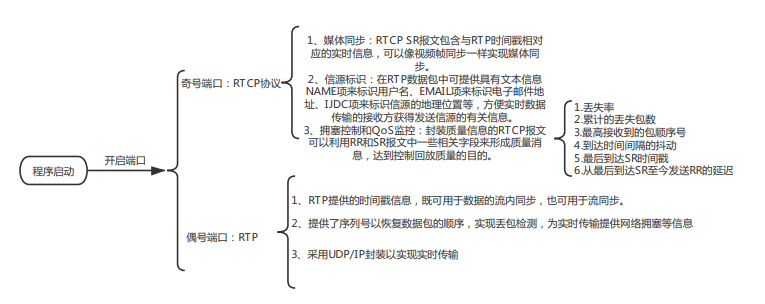
**3.2.1流式输出协议研究**

基于UDP协议不对传送数据包进行可靠性保证，很容易导致客户端资源丢包； TCP协议数据超时重传机制重传丢失数据包，延误了其他数据包的传输造成网络延迟等问题，我们应用了综合两者协议优点的协议—RTP[1]，RTP由IETF的A，V工作组于1996年公布作为RFCl889正式文档，是专门为交互式话音、视频、仿真等实时媒体数据的端到端传输服务而设计的传输协议。RTP包含两个紧密联系的部分：RTP用于携带具有实时特性的数据；RTCP用于监视服务质量并传送正22在进行的会话中参与者的信息。RTP以报文形式携带着媒体数据，而RTCP则利用RTP报文的相关传输质量信息构成RTCP报文，以监视服务质量并传送币在进行的会话中参与者的信息。

实时数据传输对于传输实时性的要求远高于传输可靠性，因此RTP数据通常  
采用UDP／IP封装，利用UDP提供的复用和校验功能，共同完成网络传输层功能。



**UDP/IP封装的RTP数据**

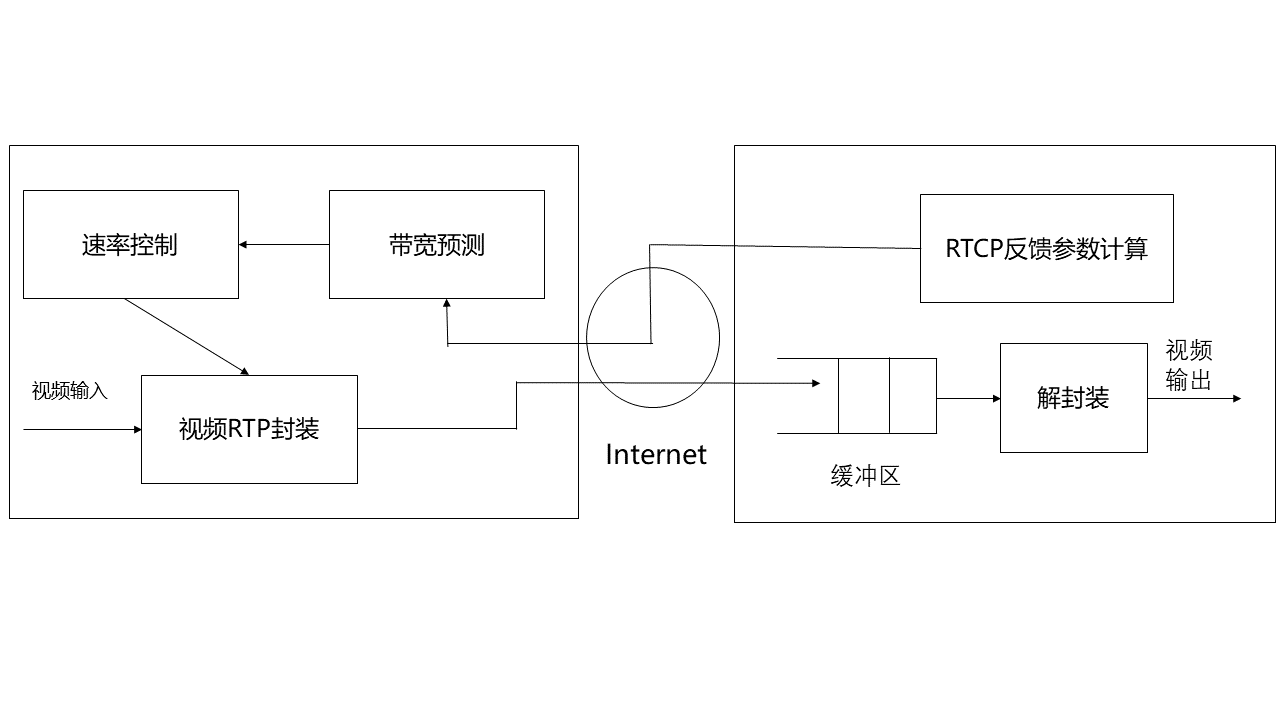
协议应用流程：

**3.2.2卡尔曼滤波器在网络状况分析中的应用**

卡尔曼滤波在数学上是一种统计估算方法，通过处理一系列带有误差的实际测量数据而得到物理参数的最佳估算。卡尔曼过滤器通过建立一系列数学方程，  
运用递归的方式对离散随机事件状态进行预测。通过统计历史包丢失率，我们运用卡尔曼滤波算法通过迭代过程对RTCP间隔内的包丢失率进行预测。根据丢包率的反馈，发送端探测到网络的状况和可用带宽。在接收端采用设置缓冲区的方式来进行回放视频的平滑处理。

**3.2.3基于RTCP反馈的网络拥塞控制机制**

利用视频接收端的RTCP反馈数据包来为发送端发送速率的变化提供调控的依据。据传输媒体特点，合理运用码率切换，采用设置接收端缓冲区的方式，并且动态调整缓冲区的大小。所有这些方面的组合构成了能够用于大数据量传输的拥塞控制机制，这其中涉及到的步骤分别为回环时间RTT的测定；输出时限的设定；包丢失率的计算；发送端速率调整。

**网络拥塞控制原理图**

**四．项目研究方案、进度安排及预期目标**

**4.1项目研究方案**

项目前期分两条主线开展：流媒体系统开发和视频字幕实时生成模型构建。

第一条主线流媒体系统开发包含云端服务器搭建及客户端开发。云端服务器拟利用阿里云平台，对数据进行处理分析。客户端能实现用户对视频资源的实时访问。

第二条主线视频字幕实时生成模型包含训练数据集获取，训练数据预处理，模型训练，模型优化。数据集获取将利用网络爬虫技术，从主流视频网站上获取视频和音频资源。首先将获取得到的嘈杂视频进行人声分离，得到纯净的音轨，并对应上相应的说话者。之后对分离的人声音频进行预处理，其基本步骤包含：压缩音频文件转为非压缩波形文件、VAD操作（静音切除）、对声音进行分帧、对波形做变换、利用隐马尔可夫模型将帧识别为状态、转态组合为音素、音素组合为单词等。同时，利用卷积神经网络对音频中的干扰声进行处理。项目用到的模型包括卷积神经网络模型、隐马尔可夫模型、声学模型和语言模型。其中，利用音频文件对前两者进行训练。语言模型通过搜集的大量文本数据进行训练，提高语音识别的准确率。基于训练结果，对模型进行进一步的优化，利用动态规划算法解决隐马尔可夫模型局部最优解问题以获得全局最优解。

最后将优化改进后的字幕生成模型应用到流媒体系统中，考虑到视频及字幕播放的实时性，再对流媒体系统进行网络传输优化。基于UDP协议不对传送数据包进行可靠性保证，很容易导致客户端资源丢包； TCP协议数据超时重传机制重传丢失数据包，延误了其他数据包的传输造成网络延迟等问题，我们应用了综合两者协议优点的协议—RTP协议。最后将经过优化流媒体平台进行全面推广。

**4.2项目进度安排**

* **2019.4**

学习视频字幕实时生成基本原理，搜集视频音频训练数据并预处理

完成流媒体系统云端服务器搭建

* **2019.6**

利用卷积神经网络算法实现嘈杂视频中的人声分离

完成流媒体系统客户端开发

* **2019.7**

优化VAD和信号处理技术。训练语音数据和文本数据，优化声学模型和语言模型，提高语音识别准确率

* **2019.9**

利用HMM模型建立状态网络，解决相邻帧状态差异问题。

* **2019.11**

利用Viterbi算法优化HMM模型，解决局部最优解问题

* **2020.1**

实现模型在流媒体系统上的应用

对系统进行网络传输优化，降低传输时延

* **2020.3**

进行平台上线，成果推广

**4.3项目预期目标**

* 模型：生成视频字幕实时生成可应用模型
* 应用：生成一整套流媒体系统，包含服务器端与客户端，并将视频字幕实时生成模型应用到客户端视频播放中，在服务器端与客户端进行网络传输优化RTCP协议应用。
* 竞赛：参加AI/大数据相关竞赛并获奖
* 发表国内外Ei/Sci检索会议核心期刊

**五．预计完成过程中可能遇到的困难和问题以及解决的措施**

1.存在问题：声学模型和语言模型识别准确率不高。

解决措施: 收集大量的语音和文本数据对模型进行训练。

2.存在问题：语音和文本数据的获取问题。

解决措施：通过网络爬虫获取大量数据

3.存在问题：电脑配置达不到训练数据的要求

解决措施：利用云平台在线训练模型

**六．参考文献**

[1] 大数据量实时流媒体输出拥塞控制机制研究. 计算机应用. 2016.5.

[2] [基于深度学习的语音识别方法研究](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=2&CurRec=11&recid=&FileName=DLXZ201902031&DbName=CJFDLAST2019&DbCode=CJFQ&yx=&pr=&URLID=). [智能计算机与应用](http://kns.cnki.net/kns/NaviBridge.aspx?bt=1&DBCode=CJFD&BaseID=DLXZ&UnitCode=&NaviLink=%e6%99%ba%e8%83%bd%e8%ae%a1%e7%ae%97%e6%9c%ba%e4%b8%8e%e5%ba%94%e7%94%a8). 2019.2

[3] [人工智能技术在语音交互领域的探索与应用](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=2&CurRec=15&recid=&FileName=DZBZ2019Z1012&DbName=CJFDTEMP&DbCode=CJFQ&yx=&pr=&URLID=). [信息技术与标准化](http://kns.cnki.net/kns/NaviBridge.aspx?bt=1&DBCode=CJFD&BaseID=DZBZ&UnitCode=&NaviLink=%e4%bf%a1%e6%81%af%e6%8a%80%e6%9c%af%e4%b8%8e%e6%a0%87%e5%87%86%e5%8c%96). 2019.2

[4] [基于HMM的声调语音模型研究](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=0&CurRec=56&recid=&FileName=WHHG201806021&DbName=CJFDLAST2019&DbCode=CJFQ&yx=&pr=&URLID=). [武汉工程大学学报](http://kns.cnki.net/kns/NaviBridge.aspx?bt=1&DBCode=CJFD&BaseID=WHHG&UnitCode=&NaviLink=%e6%ad%a6%e6%b1%89%e5%b7%a5%e7%a8%8b%e5%a4%a7%e5%ad%a6%e5%ad%a6%e6%8a%a5). 2018.12

[5] [基于改进卷积神经网络算法的语音识别](http://kns.cnki.net/kns/detail/detail.aspx?QueryID=0&CurRec=64&recid=&FileName=YYSN201806018&DbName=CJFDLAST2019&DbCode=CJFQ&yx=A&pr=&URLID=11.2121.O4.20181108.1203.014). 应用声学. 2018.11

[6] Inbar Mosseri,Oran Lang. Looking to Listen at the Cocktail Party:

A Speaker-Independent Audio-Visual Model for Speech Separation(2018).

**七．经费使用预算**

|  |  |  |
| --- | --- | --- |
| **项 目** | **预算金额(元)** | **描述** |
| 实验材料费 | 300 | 实验室用品、实验材料、实验试剂、实验耗材、电子元器件、专用工具和仪器等 |
| 实验设备购置 | 100 | 实验所需设备，购置单价超过1000元的通用设备或单价超过1500元的设备须报学校固定资产 |
| 电子设备购置 | 500 | 购买电脑，U盘、内存等电子配件 |
| 软件购置费 | 1000 | 构建信息网络方面的支出，单价8000元以上需报软件类固定资产 |
| 维修（护）费 | 100 | 实验用仪器设备维修费 |
| 办公用品费 | 100 | 日常办公用品支出，如纸张、文具等（每个项目不得超过300元） |
| 加工测试费 | 100 | 加工费、测试费、图文制作费等 |
| 书籍资料费 | 500 | 购买书籍、资料等支出 |
| 印刷费 | 100 | 打印、复印费、印刷费 |
| 差旅费 | 1000 | 车船票、住宿费、会务费等。参加会议须有会议通知、调研须有调研报告 |
| 论文、专利费 | 500 | 版面费、审稿费、专利费等。 |
| 其他费用 | 0 | 提出申请 |
| 合计 | 4300 |  |