[第一章 绪论 1](#__RefHeading___Toc29459282)

[1.1 研究背景 1](#__RefHeading___Toc29459283)

1.在线教育的发展趋势

随着社会信息化技术的大力发展和普及，以及人们对子女教育的重视程度越来越高，并且叠加近几年疫情防控政策等现实因素的影响，教育形式多样性和高效性的需求越来越凸显。

而在线教育作为一种新型的教育模式，在近几年来发展迅速，且相对于传统线下教育优势明显，首先就是时间和地点的灵活性，学生可以自由安排学习时间和地点，其次可以适应不同学习风格的需求，在线教育提供了多种学习方式，如视频课程、自学课程、讨论组、实验室活动等，可以满足不同学习风格的学生的需求，最重要的一点就是可以提高教师的教学质量，线教育可以使用最新的教学资源和技术，例如，可以使用虚拟实验室、模拟软件、在线测验等工具来提高学生的学习效果。

基于在线教育的种种优点，中国的在线教育渗透率也在不断扩大。根据2021年1月20日教育部等五部门联合发布的《关于大力加强中小学线上教育教学资源建设与应用的意见》，可以看出国家也在大力引导教育信息化的发展趋势，也为在线教育的规范健康发展提供了保障。

而本文的智慧教室项目，即是一种基于各种人工智能技术赋能的在线教育平台，旨在用技术的手段提高教师远程授课的教学质量，比如应用计算机视觉技术可以实时监测学生上课状态并进行适当提醒，基于自然语言处理的技术可以提取教师课程内容的关键词辅助教学，也可以自动检查随堂测验的结果并进行统计标注等，这些人工智能技术的开发应用可以极大的方便教师的上课效率，提高学生的听课质量。

2.远程直播教学存在的问题

虽然智慧教室等在线教育平台有很多优点，但同时也存在很多现实的问题，比如缺乏直接的师生互动，容易被听课的电子设备上的其他信息推送干扰，以及容易受到当前的网络环境以及现实环境噪声影响等，而这些问题有的可以通过与家长学生沟通适当约束处理，有些则可以通过技术手段解决。本文就是旨在通过使用语音识别的技术实时的生成教师授课内容的字幕，可以很大程度的缓解因学生听课环境嘈杂或者跟不上授课节奏而降低听课效率的影响。而其他的一些问题还需要在线教育的从业者们继续去努力克服解决。

[1.2 国内外研究现状 1](#__RefHeading___Toc29459283)

[1.2.1 字幕技术研究现状 1](#__RefHeading___Toc29459284)

字幕技术是一种将文本内容显示在视频上的技术，通常用于将另一种语言的音频翻译成文本，或者为视频提供补充信息。字幕技术的研究可以追溯到20世纪60年代，随着计算机和互联网的发展，字幕技术也得到了极大的提升。近年来，字幕技术发展迅速，许多研究机构和企业都在投入巨资进行字幕技术的研究和开发。

当前字幕生成方案大致分为两大类，一类是采用人工识别并标注最后校验的方案，一般是在准确率要求较高的离线场景下采用，比如一些电视剧电影的字幕制作就是采用此种方案，但是此方案效率较低，需要耗费很高的人力成本和时间成本。另一类就是目前很受欢迎的自动字幕生成方案，随着当前语音识别、机器翻译和文本生成等方面的技术的进步与成熟，使自动的识别声音生成对应的文本字幕成文可能，此方案效率较高，但也分为离线和在线两种模式，且支持在线识别场景，能够满足一些比如视频会议，直播课程等实时性要求较高的场景，所以也越来越成为目前主流的研究方向。

关于在线教育领域，比如国内的网易公开课，慕课等课程平台目前也都开始支持录播课程讲述内容的自动字幕生成，极大的减少了教师的课程制作成本，可以把更多的精力用在课程内容本身，提高教学质量。而对于直播课程，受限于计算资源，算法性能等因素的影响，目前还没有成熟的解决方案，还需要研究者们探索研究，这也是本文的主要研究内容。

[1.2.2 语音识别技术研究现状 1](#__RefHeading___Toc29459284)

语音识别是一种将语音转换为文本的技术，最早可以追溯到上世纪50年代。当时，贝尔实验室的研究人员就开始尝试使用电子计算机来处理语音信息，主要是通过程序规则识别简单的词语并将其组合为文本，这也为接下来的20年基于规则的孤立词识别系统的研究奠定了开端。

到了20世纪80年代，语音识别系统的研究开始从孤立词的识别转向大量连续语音识别的研究。统计语言模型大行其道，将时序建模和观察概率建模相结合的GMM-HMM成为研究的主流，但是由于当时的技术研究还处于初期，可以用来训练的标注数据也不是很充足，所以识别的准确率还很低，无法达到商用的水平。

20世纪90年代到2010年间，随着对GMM-HMM的研究发展逐渐走向成熟，商用的语音识别系统开始出现，例如DRAGON系统，IBM的Via-vioce系统，微软的Whisper系统，英国剑桥大学的HTK系统等，尤其是开源的HTK系统，提供了一整套语音识别的软件工具，为语音识别的生态发展起到了很大的贡献。

进入2010年之后，由于深度学习，循环神经网络的重新兴起，基于深度学习的语音识别系统的技术研究实现了飞跃发展。从基于DNN-HMM开源语音识别工具kaldi到近些年出现的端到端的解决方案，比如DeepSpeech2，Espnet，Wenet等，语音识别模型的识别准确率不断提高，各个模型在某些特定数据集上都取得了很好的甚至超过了人类水平的测试结果，但是这些数据集是由专业人士在干净的环境下录制的，和实际的工业使用场景还是有很大差异，要想将其应用部署在实际复杂多变的智慧教室场景下，并取得可以实用的效果，还是要面临很大挑战。

[1.3 研究主要内容及意义 1](#__RefHeading___Toc29459283)

本文主要以智慧教室项目为依托，通过语音识别技术自动的为教师直播课程生成文本字幕，提高授课效率，并通过开发相应的服务接口，灵活的集成到智慧教室产品中，提高其功能完备性，也为后续基于语音识别技术的其他功能扩展提供服务基础。

本文的研究过程主要包括以下几个方面：

1.语音分割

通常对于输入的语音信号，会有很多无用的空白信号，有意义的人声信号片段只占一部分，语音分割是指使用信号特征提取和端点检测等技术将语音信号分割成若干个连续的话语片段，并且能够去除空白信号，减轻后端识别的压力。本文主要是基于webrtcvad模块进行静音帧检测，并根据检测的结果设计相应的算法逻辑对有效帧进行分割提取。

2.语音识别

本文的语音识别模块主要是基于开源的Wenet语音识别系统框架，在其提供的基于开源的wenetspeech预训练模型的基础上，在教育领域语音识别数据集上进行微调训练，同时加入通用语言模型和热词模型对候选结果进行加权打分，最终筛选出最优结果。并且实现了字级别时间戳对齐和语音信号端点检测，从而使其支持连续长语音信号的实时识别，为后续的实时字幕生成提供技术基础。

3.模型部署

为了提高模型的高可用性，提供稳定的语音识别服务，并应用在智慧教室产品中，本文使用效率更高的c++语言重新实现了模型推理和解码算法，减少语音识别的时间开销，同时基于Grpc将语音识别服务包装成远程调用接口，增强用户调用服务的灵活性，最终使用Docker容器技术将代码模型和环境打包成image文件，根据需求在合适机器上移植部署。

4.文本处理和字幕生成

对于语音识别模块识别的原始文本，通常包含很多错别字词以及一些不连贯的情况，本文使用无标注的通用语料训练了一个中文纠错模型对其进行纠错处理，进一步提高字幕生成的准确率。另外识别的结果文本是不包含标点符号的，所以本文通过使用语音分割技术和标点添加模型相结合，为得到的文本分句并添加合适的标点符号，提高最终生成字幕的可读性和连贯性。最终将处理好的文本字幕实时发送给客户端渲染显示。

1.4拟解决的关键问题

模型推理时间开销大，不支持在线识别的问题：基于LibTorch库使用C++语言重写模型推理代码，并将pytorch框架训练生成的模型导出成LibTorch可加载的格式，以支持模型加载和在线推理。

不支持实时长语音识别问题：通过语音识别模块内部集成语音端点检测模块，能够识别用户完整语句的停止位置，并据此为依据输出完整连续的用户语音识别结果。

输出的语音识别结果存在错误字词和不连贯的问题：因为本文的字幕生成的应用场景是在单一的在线教育领域，会有很多领域常用词存在，所以可以使用热词增强技术，构建一个热词语境图，在解码搜索阶段，加入热词前缀序列得分增强。同时支持语言模型，使用先验的语言知识对结果进行打分纠错。使用这两种方法可以很好的减少常用词识别错误率。

[1.5 论文章节安排 1](#__RefHeading___Toc29459283)

第一章绪论，主要介绍了本文研究的相关时代背景和技术背景，字幕技术和语音识别技术当前的发展现状，并结合智慧教室中自动字幕的现实需求，简单介绍了本文研究的各个方面和主要解决的问题，梳理出本文的主要研究过程。

第二章系统设计概要和相关技术简介，先是系统分析了本文研究内容的各项需求，之后展示了本文的系统结构设计和数据流设计，最终系统的介绍了各个功能模块用到的相关技术和实现逻辑。

第三章语音识别，主要介绍了本文用到的语音识别框架及其实现算法，开源数据集整理及其训练过程，另外重点讲述了为了提高模型识别准确率而加入的语言模型和一些相关辅助功能实现，是本文的重点算法章节。

第四章模型服务与部署，主要介绍两种模型推理方案的对比和选择，训练好的模型到处和量化方式，以及语音识别模块对外的服务接口设计，最终介绍了Docker容器环境部署流程。

第五章自动字幕生成，是本文的功能统领章节，按照数据流的顺序，逐节阐述了从原始的语音信号输入到最终生成字幕的各个算法实现环节。

第六章系统测试与分析，主要是配置本地环境，编写测试样例，对本文实现的自动字幕模块进行功能和性能测试，并对测试结果进行系统分析和总结。

第七章研究总结与启示，主要分析了本文研究过程中存在的一些问题和一些解决办法，以及基于此研究未来的一些改进完善的方向。

[第二章 系统设计概要和相关技术简介 2](#__RefHeading___Toc29459285)

[2.1 需求分析 2](#__RefHeading___Toc29459286)

智慧教室是我们研发的一个在线教育平台，其主要的特点就是大量使用人工智能的技术开发出各种自动化辅助教学工具，提高线上授课效率，增强教师和师学生的沉浸式授课体验，既能实现线上教室的功能性，又能体现快捷高效的智慧性，这也是我们智慧教室项目研发的初衷。本文的自动字幕技术就是在这种初衷的指引下，解决课程中缺乏字幕影响授课效率的痛点，提高智慧教室的功能完备性。

而作为一种线上教育基于平台，基本功能就是提供教师授课和学生听讲的一种连接平台。而这种连接包括离线课程录播和在线远程直播两种方式，相应的本文的自动字幕模块也分为离线课程自动字幕生成和直播课程实时字幕显示两大基本需求。而具体的细分需求会在下述章节中详细阐述。

[2.1.1 功能性需求分析 1](#__RefHeading___Toc29459284)

本文的自动字幕模块的总体功能需求字幕请求和字幕返回功能，音频信号处理功能，语音识别功能，文本加工功能。

图2.1为自动字幕模块整体需求用例图。



1.流媒体传输功能需求分析

此功能属于字幕生成模块和智慧教室系统的通信接口，主要需求是实现字幕生成服务的请求调用和结果返回，并保证数据传输的稳定性和可靠性。

根据智慧教室业务需求逻辑不同，主要分为音频文件请求和实时音频数据流请求。

系统会定时监测智慧教室数据存储服务器中的课程变化情况，若检测到有新增的教师课程视频时，会自动的发起请求调用，对新增课程进行音视频分离，并将提取的音频文件发送给自动字幕模块，同时，教师客户端侧也可以主动发起字幕生成请求，自动字幕模块通过内部的软件算法生成对应的字幕文件，之后将结果发送回智慧教师后端以及教师客户端。

若智慧教室启动一次直播课程，系统后端会将直播音频信号实时转发给自动字幕模块，自动字幕模块生成字幕文本并进行时间戳对齐后按照帧格式发送给系统以及客户端进行实时的渲染显示。

2.字幕生成功能需求分析

（1）音频信号处理需求

由于音频录制设备的差异或者设置模式的原因，音频信号的通道数和采样率一般不同，而后端语音识别模块要求的输入信号一般为固定参数，所以需要根据具体设定进行信号的参数对齐转换。另外，原始的语音信号中通常会包含很多教室上课的环境噪声以及一些其他人的干扰信号，同时在远程传输中也会加入一些噪声信号，为了提高模型识别的准确率，我们会使用语音信号处理算法进行降噪和人声提取处理，提高模型输入信号的信噪比。

（2）语音识别需求

此需求是本模块的核心功能需求，为了提高语音识别算法的泛化能力，要使用大量语音识别数据集训练高效的算法模型，并且加入辅助的逻辑规则，生成稳定的语音识别文本结果。由于本文支持离线和在线字幕生成两种生成模式，所以要求语音识别算法也能支持流式和非流式两种信号输入模式。同时为了支持连续的长语音输入，所以算法内部还需有语音分割和端点检测的功能，并且需要输出识别结果文本对应的时间戳，以便满足后续文本处理的需要。

（3）文本处理需求

在深度学习和大数据的加持下，语音识别算法识别的文本结果虽然正确率得到了很大提高，但还是会有很多错误字词和语义不连贯的情况，所有需要使用一些通用的语言纠错模型进行错误字词检查和纠正。另外，得到的文本是不包含标点符号的，这对于书面文本的阅读很不友好，所以还需要通过端点检测和对文本语境的理解在合适的位置添加合适的标点符号，提高生成字幕的可读性。

[2.1.2 非功能性需求分析 1](#__RefHeading___Toc29459284)

本系统研发后会最终应用在智慧教室产品中，所以除了实现基本功能之外，还需要考虑产品功能稳定性，部署成本和维护成本以及对未来新业务需求的扩展支持。

1.稳定性和高可用性

稳定性和鲁棒性作为业务软件开发的基本要求至关重要，首先需要支持各种异常处理，包括信号输入格式不一致，通信中断恢复等情况的处理。其次是进行冗余设计，若发现处理结果异常时，尝试更换数据处理链路。最重要的要有服务自检机制，如发现自身服务异常或停止时，尝试自动重启恢复，若无法恢复需向智慧教室后端发送服务停止消息和日志，方便排查维护。

2.响应实时性

对于离线字幕生成请求，一般对实时性要求不高，对于多个请求同时发送的情况，可以使用请求队列或者多进程多服务的方式满足需求。而对于实时字幕生成请求，则需要考虑算法处理延迟问题，若延迟较大会极大影响用户体验甚至失去字幕生成功能存在的意义，所以需要使用模型量化和工程优化等技术较少延迟，提高用户需求相应实时性。

3.可扩展性

在模块设计之初，就要考虑各个功能模块的解耦设计，在满足性能的前提下，各个功能尽量包装成服务，并在考虑到各种调用场景的情况下灵活的设计好服务接口，方便主业务逻辑层调用。同时为日后其他基于语音识别技术的业务需求开发提供基础服务，较少日后的重复开发工作量。

4.成本

自动字幕模块的附加成本包括部署成本和维护成本。通过程序算法优化，尽量做到模型部署轻量化，降低cpu和内存的资源消耗，并在部署时考虑使用负载均衡策略，优化运行成本。同时也要考虑模块的可维护性，开发中注意加入注释和说明文档整理，系统中加入分级日志机制，总结出异常恢复的常用办法，方便功能模块的部署维护和异常处理。

2[.2 系统](#__RefHeading___Toc29459286)架构设计

基于上述章节描述的各种功能和性能需求，本文基于软件工程系统架构的设计方法，对字幕生成模块进行系统设计，使其方便的集成到智慧教室系统中。

智慧教室作为一种在线课程的网络平台，使用较为常见的客户端服务器的架构模式，客户端包括学生客户端和教师客户端，通过使用WebRTC实时音视频传输协议，将客户端采集的音视频信号发送给服务后端，后端的流媒体服务器集群会对原始的音视频信号进行编码解码，之后调用各种AI服务进行后处理操作，最终将处理好的视频流推送给相应客户端，同时转存到文件存储服务器。

图2.2为智慧教室系统架构图

而本文主要工作是为智慧教室系统提供语音识别和字幕生成服务功能。其中客户端侧使用开源的ckplayer实现字幕的加载和播放工作，在智慧教室服务端，请求处理层主要是提供网络负载均衡的功能，使所有的流量得到合理分配，保证服务请求的及时相应。业务处理层为系统提供各种基础服务，包括最基本的完成音视频信号推流拉流工作的流媒体服务，

智慧教室项目使用的流媒体传输技术包括WebRTC和RTMP两种传输方案，而流媒体服务层则会调用包括姿态识别，字幕添加等AI服务等进行处理，本文的主要内容就是设计在此层。最下层是数据处理层，首先设计了两个中间件模块，通过使用Kafka消息队列和Redis业务数据缓存的方式缓解系统数据读写压力，底层数据库则完成系统业务数据的存储功能，其中mysql负责存储业务信息，用户信息，课程信息等元数据，而FastDFS文件存储集群则用来分布式存储体积较大的课程音视频数据。

[2.3](#__RefHeading___Toc29459286) 数据流设计

根据智慧教室的业务需求，分为录播回放和课程直播两种业务场景，相应的自动字幕生成服务也分为离线音频文件字幕生成和在线音频流字幕生成两条数据流模式。

图2.3为智慧教室自动字幕生成数据流图

对于离线录播课程的场景，教师把事先录制好的课程视频上传，智慧教室后端在重新解码编码等基本处理后将课程信息和数据存储在数据库和文件存储集群中，课程监测模块监测到新增课程标志后，在确认教师开启自动字模生成选项后，会向字幕生成模块发送离线字幕生成请求，并同时进行音视频分离操作，提取出wav格式的音频文件，然后调取字幕生成服务生成字幕文件，最终将字幕文件存储到FastDFS文件服务器中并更新数据库中的课程信息。之后学生若观看课程视频发起课程请求时，系统会将课程视频和字幕文件同时发送给客户端，由客户端播放器进行字幕的添加和渲染。

而对于教师直播授课的场景，教师的直播音视频流会推送到后端，首先经过Nginx负载均衡服务器进行分流操作，之后经过SRS流媒体服务器进行流媒体的处理，同时向字幕生成模块发起在线字幕生成请求，并将音频数据流实时转发给字幕生成服务，字幕生成服务会实时生成带有时间戳的文本字幕帧，最终再由流媒体服务器将字幕帧数据发给学生客户端进行实时显示。

[2.4 本章小结](#__RefHeading___Toc29459286)

本章主要对自动字幕模块进行了全面的需求分析和系统架构设计。

第一节使用软件工程的方法对自动字模的功能需求和非功能需求进行了系统分析，从而为之后的系统实现提供清晰的指引。

第二节主要介绍了智慧教室项目的整体分层架构，以及本文的自动字幕模块在其中处于什么位置。

第三节阐述了系统整体数据流的转移处理过程，详细展示了各个子功能模块的逻辑连接关系，向读者讲述了系统的整体工作流程。