[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。

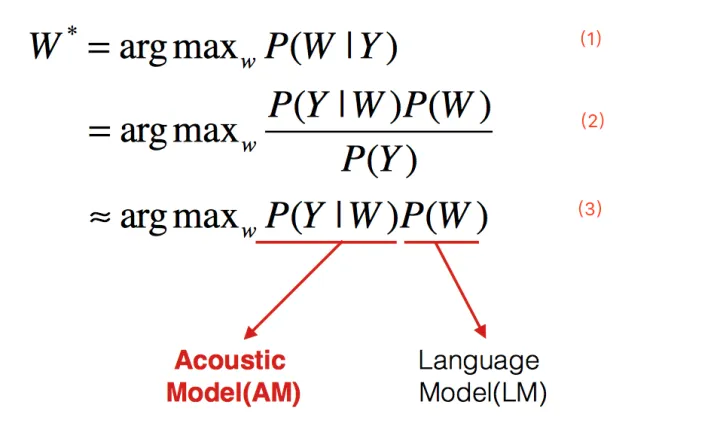
第二节主要介绍了智慧教室项目的整体分层架构，以及本文的自动字幕模块在其中处于什么位置。

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。

[第三章 语音识别](#__RefHeading___Toc29459285)

3.1语音识别技术简介

语音识别作为本系统的核心模型算法，是实现自动字幕生成功能的基础。主要就是对输入的语音信号进行特征提取得到一个特征向量序列，再通过一些识别算法模型处理，得到对应的语音识别文本序列。其数学模型为：



其中W表示结果文本序列，Y表示输入语音信号。公式1表示语音识别的优化目标是在给定语音输入的条件下，找到一个文本序列，使得条件概率结果最大。根据贝叶斯公式，可以得到公式2，其中分母表示得到这条语音信号的自然概率，它的大小和优化目标没有参数关系，可以在求解时不予考虑，从而得到公式3。公式3的第一部分表示给定一个文本序列，出现当前语音信号的概率，它就是语音识别的声学模型，也是语音识别算法的基本实现模型；第二部分则表示此文本序列出现的自然概率，就是语音识别中的语言模型，加入语言模型这个先验信息可以很大程度提高语音识别的准确率。

图3.1为语音识别基本流程图

3.1.1特征提取

本文主要讲述如何根据音频信号提取 MFCC 和 FBank 特征，这也是目前在语音识别任务中使用最广泛的两种特征。

人类的语音信号的频率大部分在 10000Hz 以下，根据奈奎斯特采样定理，20000Hz 的采样率就足够了。我们通常使用麦克风进行音频录制的采样率为 16000Hz，一个采样点使用 16bit 来存储。

图3.2为特征提取流程图



输入的语音信号，首先经过预加重处理，在音频录制过程中，高频信号更容易衰减，而像元音等一些因素的发音包含了较多的高频信号的成分，高频信号的丢失，可能会导致音素的共振峰并不明显，使得声学模型对这些音素的建模能力不强。预加重是个一阶高通滤波器，可以提高信号高频部分的能量。

第二步是分帧加窗操作，语音信号是一个非稳态的、时变的信号。但在 短时间 范围内可以认为语音信号是稳态的、时不变的。这个短时间一般取 10-30ms，因此在进行语音信号处理时，为减少语音信号整体的非稳态、时变的影响，从而对语音信号进行分段处理，其中每一段称为一帧，帧长一般取 25ms。为了使帧与帧之间平滑过渡，保持其连续性，分帧一般采用交叠分段的方法，保证相邻两帧相互重叠一部分。相邻两帧的起始位置的时间差称为帧移，我们一般在使用中帧移取值为 10ms。因为后面会对信号做 FFT，而 FFT 变换的要求为：信号要么从 -∞ 到 +∞ ，要么为周期信号。现实世界中，不可能采集时间从 -∞ 到 +∞ 的信号，只能是有限时间长度的信号。由于分帧后的信号是非周期的，进行 FFT 变换之后会有频率泄露的问题发生，为了将这个泄漏误差减少到最小程度，我们需要使用加权函数，也叫窗函数。加窗主要是为了使时域信号更好地满足 FFT 处理的周期性要求，减少泄漏。

接下来的操作是 离散傅里叶变换（缩写为 DFT），将每个窗口内的数据从时域信号转为频域信号，从而更方便提取信号的标志性特征。DFT 的时间复杂度为O(n2) ，为了满足系统的实时性要求，在实际中使用的是快速傅里叶变换（缩写为 FFT），其时间复杂度是O(nlog(n)) 。

之后是重要的梅尔滤波器处理，从 FFT 出来的结果是每个频带上面的幅值，然而人类对不同频率语音有不同的感知能力，频率越高，感知能力就越差。[梅尔刻度](https://link.zhihu.com/?target=https%3A//zh.wikipedia.org/wiki/%25E6%25A2%2585%25E5%25B0%2594%25E5%2588%25BB%25E5%25BA%25A6)（Mel）是一种非线性刻度单位，表示人耳对音高变化的感官，基于频率定义。在 Mel 频域内，人的感知能力为线性关系,能更好的符合物理世界基本原理。将梅尔域上每个三角滤波器的起始、中间和截止频率转换线性频率域，并对 DFT 之后的谱特征进行滤波，得到 40 个滤波器组能量，为了使特征对输入信号的扰动不敏感，再进行 log 操作，就得到了 40 维的 Fbank（Filter Bank）特征。

最后使用离散余弦变换（IDFT）对滤波器组系数去相关处理，并产生滤波器组的压缩表示。通常，保留所得到的2-13阶倒频谱系数，就是MFCC特征结果，其余部分被丢弃，丢弃其他系数的原因是它们代表了滤波器组系数的快速变化，并且这些精细的细节对语音识别没有贡献。

MFCC是在FBank的基础上进行的，所以MFCC的计算量更大，且特征相关性小，抵抗噪声的鲁棒性低，一般用于GMM的训练。而fBank特征保留了更多原始特征，其相邻的特征高度相关（相邻滤波器组有重叠），适用于DNN建模，目前基于深度学习模型的语音识别算法一般多采用fBank特征。

3.1.2语音识别声学模型

声学模型可以理解为是对声音信号产生过程的建模，它能够把语音输入转换成声学表示的输出，更准确的说是给出语音属于某个声学符号的概率。

1.HMM

HMM模型，也叫做隐马尔科夫模型，是一种经典的机器学习序列模型，描述为一个隐藏的[**马尔科夫链**](https://zhuanlan.zhihu.com/p/26453269)随机生成不可观测的隐**状态序列**，每个状态生成一个观测，而由此产生一个**观测序列。**

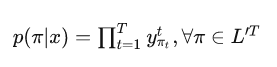
使用HMM对声学模型进行建模时，首先会将语音信号根据发音字典编码成三音素序列，然后穷举当前三音素序列所有状态转移序列，转移概率使用状态转移矩阵进行表示，同时使用发射矩阵对音素到文本token的发射概率进行表示，通常使用GMM或DNN对发射概率进行建模，最终使用维特比动态规划算法计算出概率最大的那条状态序列，从而解码出识别文本结果。

在基于DNN-HMM架构的语音识别声学模型中，训练DNN通常需要帧对齐标签。而在GMM中，这个对齐操作是通过EM算法不断迭代完成的，再加上其自身马尔科夫假设这一限制，对真实发声这一物理过程的建模表达能力有限，现在逐渐转向基于深度学习的端到端的语音识别系统的研究。

2.CTC

CTC的全称是Connectionist Temporal Classification，就是连接时序分类。它要达到的目标就是直接将语音和相应的文字对应起来，实现编码器输出特征向量和输出标签的对齐，同时再接入一个全连接神经网络预测字符token的概率分布，实现时序问题的分类。

其数学模型为



其中pi表示token字符序列，x表示特征编码器的输出。因为很多帧对应同一个token，同时很多帧没有任何输出，所以引入一个空白输入blank标签，相当于定义了一个多对一的函数，并在输出结果中，把重复的token合并，并把空白标签去除，得到最终的识别token序列。

图3.3为CTC识别效果示意图

在实际的语音识别系统的使用过程中，CTC一般用来作为模型训练过程中编码输出和对应标签之间损失函数的计算。首先穷举出所有编码输出帧对应的预测结果token路径，并在这些路径中找到所有能通过CTC算法映射成标签序列的路径，然后逐一计算生成这些路径的概率，最终训练模型使这些路径的概率之和最大化。从这个算法过程可以看出，需要计算的路径数据会很大，时间复杂度最坏能达到O(n^t)，这是系统所不能接受的。所以我们一般使用动态规划算法降低时间复杂度，快速计算损失，其主要思虑就是若多种路径使用相同步长映射为相同输出，那么就把这些路径概率合并计算。

图3.4为CTC损失计算示意图



3.LAS

LAS为基于深度学习Transformer思路的一种端到端的语音识别声学模型，其中L为Listen代表模型的Encoder模块，S为Spell代表模型的Decoder，A代表Attention，表示模型使用了[注意力机制](https://so.csdn.net/so/search?q=注意力机制&spm=1001.2101.3001.7020)。

Listen模块即模型的encoder编码器模块，这个模块以由特征提取模块提取的特征向量序列(如MFCC)作为输入，输出一串相同长度的编码向量序列。因为输入序列中单位时间内就有非常多个向量，且根据发声的连续性原理，相邻的向量信息重合度高，导致训练效率低，所以 编码器通常会先对输入序列作降采样操作，然后通过基于堆叠CNN，RNN，MLP等网络block进行编码操作，同时加入一些和残差连接等操作提高网络训练效率。编码器的主要作用就是把不同的发声者的发出的相同词语句子的差异和语音数据中的噪声等移除，同时提取语音中与具体内容相关的信息。

Attention模块则将encoder的输出向量与关键字向量 z 0​进行attention注意力提取操作，计算当前输出所各个输入的权重参数，从而训练模型自动确定当前输出与哪些输入有关，提高模型的表达能力。

Spell模块即模型的decoder解码器模块，此模块使用前一刻和attention模块进行注意力的计算，然后通过一个全连接网络得到一个长度为输出词表大小的向量，代表当前时间步预测输出的分布，重复此操作，即可得到完整的预测序列。

由于神经网络强大的建模能力，相对于CTC，端到端模型的输出标签不需要进行细分对齐操作，这使模型优化训练变得简单灵活，但是从解码器的计算过程可以看出，必须编码器对所有输入进行一次性编码，解码器才能按时间步进行解码输出，其训练的时间成本较高，且只能应用于离线场景，在需要实时识别的场景使用中一般需要逻辑优化。

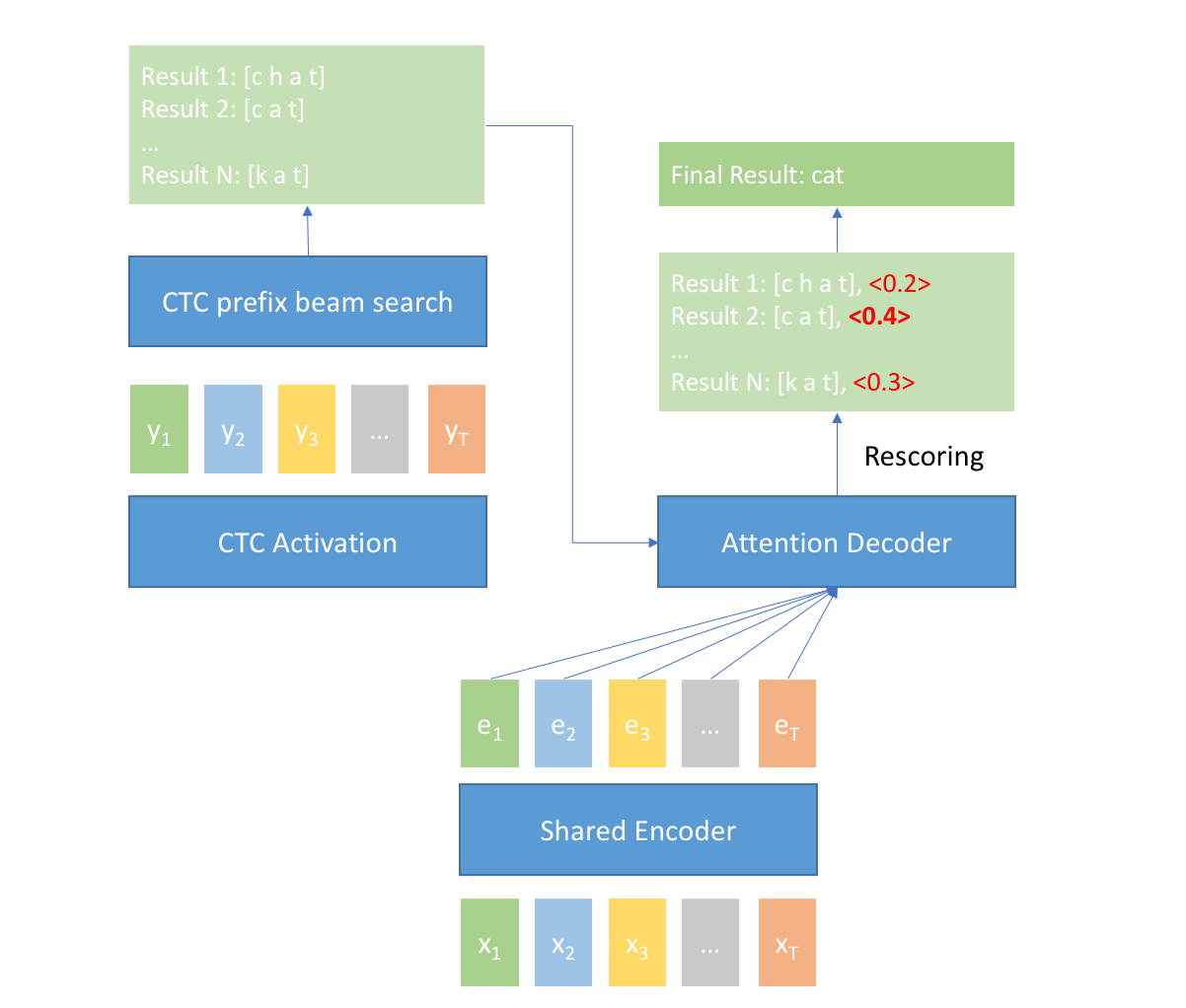
3.1.3语音识别语言模型

语言模型就是从大量的文本预料库中，统计出语言文本的一般性语法语义规律，并将这种规律使用某个模型进行表示。对于一种语言的一条文本序列，可以通过训练好的语言模型进行先验概率计算，得到的概率分值即代表了其语言合理性，概率分值越高，表示这条文本序列越符合此种语言的语法语义规律。

将语言模型应用到语音识别算法中，可以引入此种语言的先验知识，很大程度上对直接的声学模型的输出进行错误的语法语义纠正，从而提高整体的语音识别准确率。

常见的语言模型有基于文本统计的n-gram语言模型和基于深度学习预训练的bert模型，两种模型在模型大小，性能和效果上各有特点，需要根据具体的使用场景需求进行选择。

[3.2 Wenet语音识别框架](#__RefHeading___Toc29459286)



Wenet是出门问问语音团队和西工大语音实验室联合开发并开源的一款面实际业务场景应用的语音识别框架工具包，主要实现了一种叫做U2的两阶段语音识别解码算法，其最重要的特点就是在同一个端到端模型下实现了流式识别和非流式识别两种识别模式，极大扩展了模型服务的应用场景，缩小了语音识别算法的研发和实际工业场景应用之间的鸿沟，这也是本文选择使用Wenet作为自动字幕模块的基本语音识别框架的最重要的原因。

近几年来，随着深度学习技术的成熟，基于端到端的语音识别系统在语音识别任务标准词错率（WER）上已经完全超过了 传统的混合语音识别框架，但是在实际的使用场景中，部署一个端到端系统还有度多现实的问题需要解决。

而Wenet语音识别框架的两大优势就是能提供流式识别和模型轻量级易于部署。首先Wenet使用基于block的conformer网络结构作为编码器，支持流式输入，decoder模块根据系统检测到的序列端点进行适时的重打分操作，这样即满足了流式识别的实时性要求，又保证了系统有较高的识别准确率。另外，相对于其他优秀的语音识别开源框架，比如Espnet，Deepspeech等，Wenet的核心设计目标就是高性能易于部署，其没有对很多序列化识别任务进行统一抽象，而是完全聚焦于语音识别任务，同时不依赖于像Kaldi等大体量的传统语音识别工具包，完全基于Pytorch及其生态系统开发，代码简洁，安装方便，极大提高了开发者的应用部署效率。

[3.1.1 模型网络结构](#__RefHeading___Toc29459284)

Wenet语音识别框架实现了U2算法，如下图3.5所示，其模型结构包含三个网络部分，分别是共享编码器（Shared Encoder），CTC解码器（CTC Decoder）和注意力解码器（Attention Decoder），其中共享编码器有多个Transformer或者Conformer网络层组成，其主要作用就是对特征提取的原始特征向量进行信息编码，通过自注意力机制提取高层抽象特征，为了保持良好的流式识别时延，会设定超参数在注意力计算时使用有限右侧信息，CTC解码器包含一个线性全连接层，提供编码输出到CTC激活映射的转换，对于注意力解码器，包含多个Transfromer解码器层。

1.流式输入和非流式输入的统一结构

Wenet针对工业届应用的具体场景需求，兼顾流式响应速度和识别效果，采用了一套两阶段解码的方案，用一个统一的模型结构同时有效的支持流式输入和非流式输入场景。

CTC解码器进行第一遍流式解码，根据解码分值排序取前n个序列作为候选结果，其中分值最高的结果序列可作为流式识别结果实时返回。

多个候选结果再通过注意力解码器做一遍teacher forcing rescoring操作，根据得分重新排序。得到更好的识别结果。

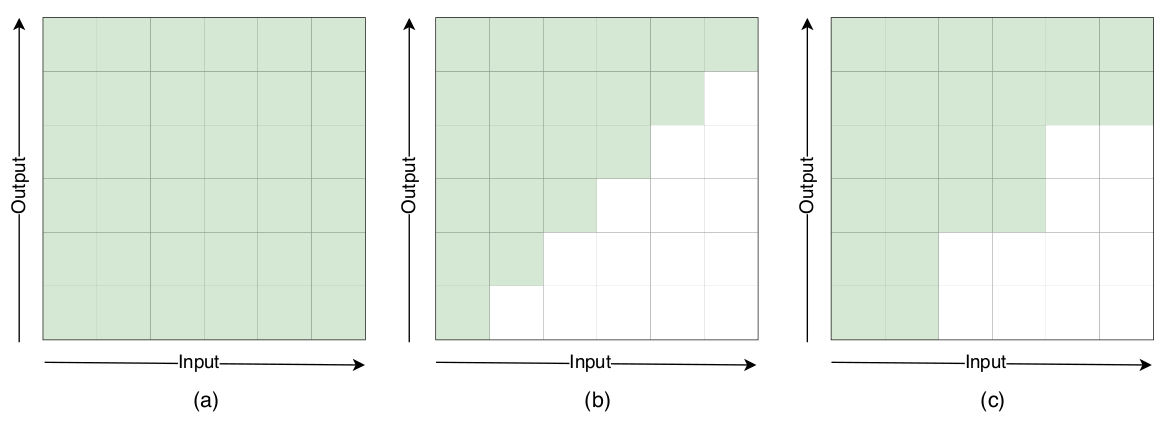
2.对编码器Conformer结构进行流式改进

标准的Conformer网络结构，self-attention层需要对整个输入序列进行attention得分计算，不能支持流式输入，并且卷积层的计算也需要依赖右侧固定长度的序列信息，依赖的长度也会随着模型层数的增加而增加。所以，需要对这些具体的网络结构进行改进，

使其仅仅依赖有限长度的右侧上下文信息，并且支持流式推理。

对于self-attention层，Wenet选择使用了chunk方案进行改进，主要将序列分成多个chunk，每帧在chunk内部做attention ，这种方案可以保证对右侧依赖的长度与网络层数无关。在基于chunk的方案中，chunk的大小会影响着流式最大延时和模型识别率性能，大的chunk延时大但性能更好，小的chunk延时小但性能会变差。Wenet使用了一种新的动态chunk训练算法，可以使得模型的chunk大小动态可变，在运行时，可根据当前场景延时要求和识别率需求手动调整chunk大小，而无需重新训练模型。 而对于convolution层，Wenet使用一个左侧卷积，以支持流式推断。

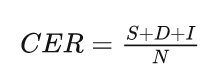
下图3.6为conformer网络结构改进图



[3.1.2 预训练模型](#__RefHeading___Toc29459284)

Wenet作为一个完善的基于端到端模型的语音识别框架，其提供了很多开源语音识别数据集的预训练模型，没有充足计算资源的研究开发人员提供了一个通用的语音识别能力，也为开发者根据业务需求使用领域数据集进行微调训练提供基础，这也是本文选择Wenet的又一重要原因。

表3.7为Wenet基于不同数据集预训练模型的CER（字错率），CER是语音识别任务的常用评价指标，计算公式为



其中S（substitution） 表示替换的字符数目，D（deletion）表示删除的字符数目，I（insertion）表示插入的字符数目，N 表示参考序列中字符总数，可以看出，CER越小，代表语音识别效果越好。

测试集为智慧教室中的课程视频，标签字幕文本已由人工标注。

| **Datasets** | **Languages** | **Testset CER** |
| --- | --- | --- |
| [aishell](../../asr_wenet/wenet/wenet/examples/aishell/s0/README.md) | CN | 10.57 |
| [aishell2](../../asr_wenet/wenet/wenet/examples/aishell2/s0/README.md) | CN | 9.83 |
| [multi\_cn](../../asr_wenet/wenet/wenet/examples/multi_cn/s0/README.md) | CN | 7.54 |
| [wenetspeech](../../asr_wenet/wenet/wenet/examples/wenetspeech/s0/README.md) | CN | 7.32 |

根据在业务数据上的测试对比，基于wenetspeech数据集的预训练模型的字错率最低，所以本文选择此模型作为本系统的初始训练模型。

[3.1.3 解码算法](#__RefHeading___Toc29459284)

Wenet在训练时同时使用了帧级别deocder的CTC loss和label级别decoder的loss。因此在识别推理时可以使用多种不同的解码方式进行灵活组合解码。

下图3.7为Wenet解码流程示意图。

目前Wenet支持四种解码算法

1.CTC greedy beam search， 帧级别解码，解码过程不合并前缀，每一个时间步都选择概率最大的作为结果，最终进行ctc序列处理。 此解码算法本质上使用贪心思路，得到的结果一般不是最优解，但是优点是解码搜索速度快，效率高。

2.CTC prefix beam search ， 帧级别的解码，每一个时间步合并相同的ctc序列前缀求和计算，选择概率和最高的n个前缀序列作为当前候选集合， 并重复此计算过程，最终得到n个最优解码结果。其本质是一种动态规划算法，保证得到最优解的前提下提高计算效率，也是语音识别领域中最常用的解码方式。下图3.7为CTC prefix beam search 解码流程示意图。

3.Attention decoder beam search，label级别解码， 基于cross-attention的注意力解码器单独解码，不支持流式识别。

4.CTC + attention rescoring， 将CTC decoder的n-best结果，通过attention decoder进行重打分 。

CTC + attention rescoring方案可以利用CTC decoder的输出作为流式结果，在损失很小准确率的情况下支持场景的流式需求，并最终利用rescoring改进结果，得到一个识别率更高的最终结果。这也是本系统在模型推理实现和应用时采用的解码方案。

[3.2 领域数据集微调训练](#__RefHeading___Toc29459286)

[3.3 语言模型](#__RefHeading___Toc29459286)

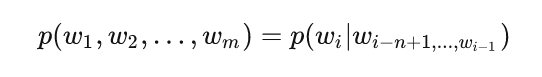
根据上述章节3.1.3可知，语言模型对语音识别系统至关重要，加入语言模型可以很大程度提高识别准确率，但同时也应注意到，语言模型的引入也会提高系统计算量，增大系统识别延迟，所以本文在引入语言模型的同时，使用一些逻辑算法策略，在合适的时机进行语言模型打分计算，兼顾系统识别准确率和效率。

图3.8为系统解码流程图，其中解码算法有两个分支，一条是使用标准的CTC prefix beam search对CTC激活输出进行解码，另一条是加入语言模型的解码搜索算法，在 prefix beam search的基础上，使用语言模型进行筛选过滤候选路径，重复更新搜索，最终输出N-best结果。

[3.3.1 基于Kenlm的ngram语言模型 1](#__RefHeading___Toc29459284)

n-gram是一种基于统计的传统语言模型，其主要思想就是将语言文本序列按照大小为n的滑动窗口进行操作，每次截取一个长度n的文本片段序列，文本序列中的每一个token叫做gram,在语料库中，统计所有gram出现的频率以及所有截取的ngram出现的频率，从而能够计算出截取片段在当前gram为前缀的条件下出现的概率，对于一个完整的句子，可以把所有的条件概率相乘得到句子出现的概率，并将其作为评价句子合理性的评分指标。

对于一个m个词组成的文本序列，我们希望计算其出现概率，即句子的合理性概率，根据条件概率链式法则，同时加入马尔科夫假设，得到如下计算公式



其中wi为序列中的第i个词。

每一个词w都具有v（词典大小）种取值，这样构造的理论词对数量非常多，但根据语言规则，实际的语料中不会有这么多词对组合，根据最大似然估计准则，得到的概率大部分都为0，这就为连乘计算带来麻烦，为了解决这个问题，会引入各种数据平滑算法，包括加一平滑，回退平滑等。

n-gram由于计算简单，部署灵活，效率较高，但是由于其使用了马尔科夫假设，不能考虑长远的上下文语义，所以一般只能用来实现一条句子的语法通顺性的打分操作。

本文使用Kenlm来实现ngram打分功能，Kenlm是一个轻量级的语言模型工具包，并对计算过程进行了很多工程和算法优化，推理速度快，可以高效完成对大量候选序列的语法合理性评价筛选工作。

[3.3.2 基于bert预训练语言模型](#__RefHeading___Toc29459284)

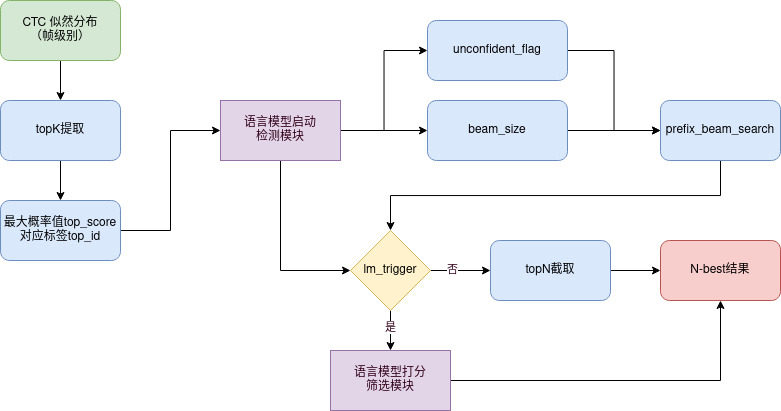
bert是一个预训练的语言表征模型，由于其内部的自注意力机制和在大量无监督语料上进行了预训练，能更好的关联上下文语义和句子的整体情感信息，所以具有出色语言表征能力。

bert采用遮蔽语言模型MLM进行自回归训练，就是随机mask掉句子中15%的内容，根据前后已知的词来预测mask的内容，相当于完形填空，可使模型学习词与词的关联。所以我们计算bert MLM层输出的似然分布概率，其代表了当前词在当前句子的上下文语境下出现的合理性，然后把所有的概率值相乘，其结果作为bert语言模型对句子的评分值，最终使用此评分结果对ngram语言模型的粗筛结果进行精筛操作，得到语法语义都更为合理的候选集。

3.3.3语言模型搜索处理逻辑

本文中语言模型的引入采用候选序列重打分机制，包括ngram粗筛和bert精筛两个步骤，最后对筛选结果中的序列通过语言模型得分和CTC似然得分的加权分值进行排序，获得最终的n-best识别结果。

图3.9为语音识别系统语言模型处理流程图。

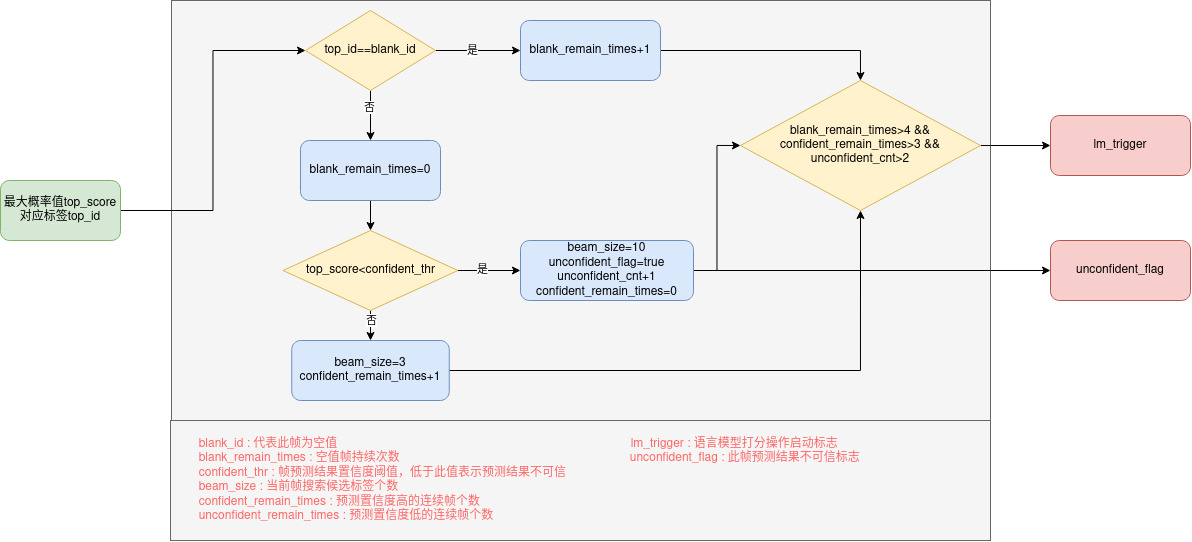


首先读取帧级别的CTC似然概率分布值，然后对每一帧根据概率值排序，获取前k（k为系统超参数）大的标签，然后经过语言模型启动检测模块进行逐帧检测，若未检测到语言模型启动标志，则进行普通的prefix beam search进行搜索更新，但同时将预测结果不确信的帧进行冗余搜索，相反，若启动标志为真，则调用语言模型打分筛选模块对冗余候选结果进行筛选，得到最终的N-best结果。

1.语言模型启动检测模块

主要功能就是对CTC的预测输出进行检测判断，从而确定启动语言模型进行纠正筛选的合理时机，在确保语言模型打分准确的前提下，同时满足系统的实时识别的需求。

图3.10为语言模型启动检测模块程序流程图



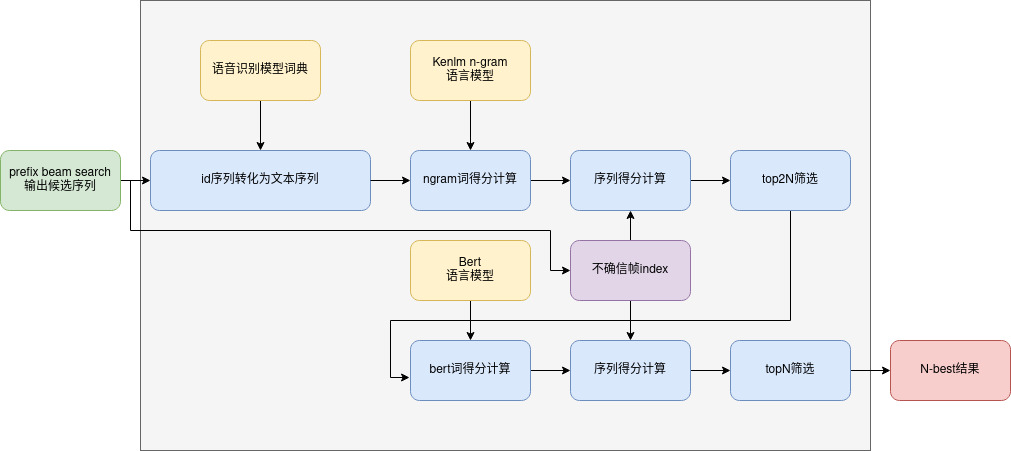
其基本思路就是设定一个CTC预测输出的置信度阈值，若某帧的最大概率低于此阈值，则认为此帧的CTC输出不可信，同时统计空白帧持续次数，可信帧持续次数和不可信帧一共出现的总次数，若当前序列中出现超过2个不可信帧同时当前帧为空白帧状态时，则将语言模型启动标志置为真。并且，对于每一帧，都会将当前帧的置信状态进行输出，若当前帧为不可信帧，则在prefix beam search 过程中将此帧的top10似然概率值的标签都加入到搜索序列中，防止正确结果漏出，为之后的语言模型筛选提供冗余候选集。

2.语言模型打分筛选模块

此模块的主要功能就是为候选序列集合进行语言模型评分，从而筛选出语法语义都更合理的序列结果。由3.3.1和3.3.2章节的研究可知，ngram的效率高，但效果一般，而bert模型的效果较好，但是时间性能较差，表3.2展示了两种语言模型的时间性能对比结果，所以本文将两种模型联合使用，设计了两层筛选机制。

|  |  |  |
| --- | --- | --- |
|  | time(s) | score |
| ngram | 5.72E-04 | 0.079 |
| bert | 0.013 | 0.149 |
| 测试序列 | 那年冬天祖母死了父亲的也差事差也交卸了 | |

图3.11为语言模型打分筛选模块处理流程示意图

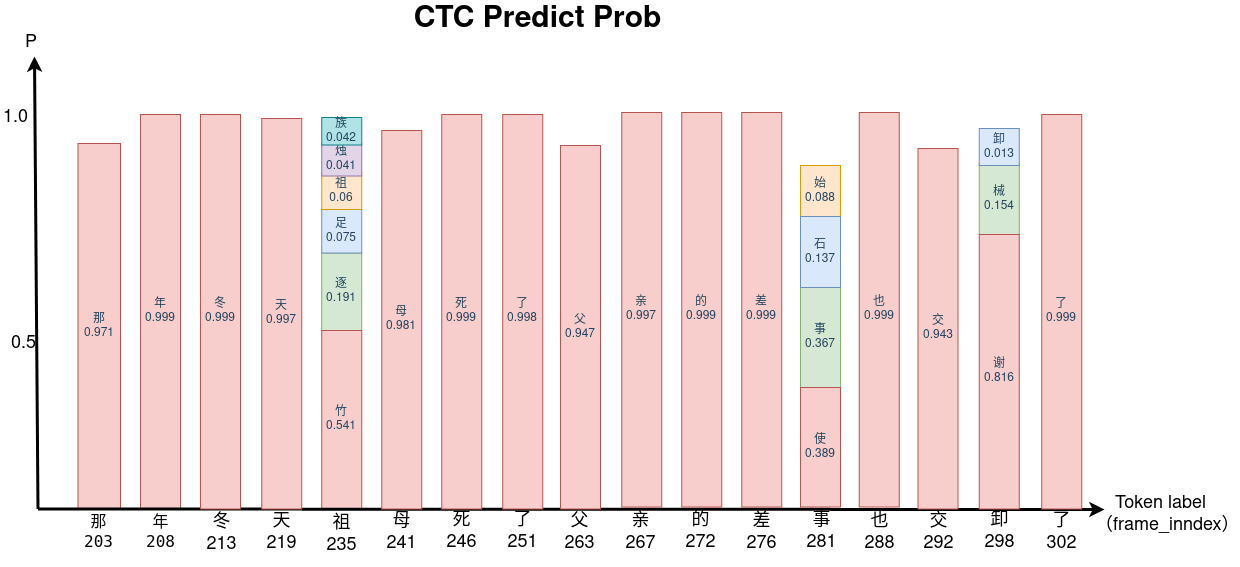


首先会对候选ID序列基于语音识别模型词典转化为文本序列，因为考虑语言模型设计的独立性，语言模型和语音识别模型一般不采用同一个标签词典，所以转化为统一的文本序列后便于语言模型处理。

之后对所有的候选序列经过ngram语言模型进行打分操作，为了不受其他其他确信帧的语言模型得分值的影响，根据输入的不确信帧的序号统计不确信帧的得分之和从而得到完整序列的得分，得到2n个候选结果。然后调用Bert语言模型重复此操作，从而筛选出n个最终候选集。

3.两层语言模型筛选实例

下图3.12为一条测试样例序列的CTC预测输出结果。其中红色柱状图中的token序列是CTC的预测结果，横坐标下的为真实的标签序列。

根据语言模型检测模块的识别检测，第235，281，198帧为不确信帧，且不确信帧数大于最小阈值，所以启动语言模型筛选，不考虑CTC规则中的重复合并情况，一共有6\*4\*3条候选序列组合，将所有候选序列通过语言模型进行打分操作，并根据分值进行排序截取，得到最终的N-best结果。

下图3.13是语言模型打分排序结果，可以看出，正确的序列组合分值最高，即是通过语言模型筛选得到正确预测结果序列。



[3.4 热词增强模型 2](#__RefHeading___Toc29459286)

[3.5 时间戳对齐 2](#__RefHeading___Toc29459286)

[3.6 端点检测分割 2](#__RefHeading___Toc29459286)

[3.7 本章小结 2](#__RefHeading___Toc29459286)