Hans汉斯

撤稿声明

撤稿文章名:	基于深度学习的视频语音提取文本系统设计与实现
作者:	谢煜颖
* 通讯作者:	735308107@qq.com
期刊名:	软件工程与应用(SEA)
年份:	2021
卷数:	10
期数:	4
页码 (从X页到X页):	528-541
DOI (to PDF):	https://doi.org/10.12677/SEA.2021.104057
文章ID:	2690558
文章页面:	http://www.hanspub.org/journal/PaperInformation.aspx?paperID=44437
撤稿日期:	2021.10.21

撤稿原因 (可多选):

☑ 所有作者

☐ 部分作者:

☐ 编辑收到通知来自于 ○ 出版商 ○ 科研机构: ○ 读者: ☐ 其他:

撤稿生效日期: 2021.10.21

撤稿类型 (可多选):

☐ 结果不实
- ○ 实验错误 ◎数据不一致 ◎ 分析错误 ○ 内容有失偏颇
- ○ 其他:

☐ 结果不可再得

☐ 未揭示可能会影响理解与结论的主要利益冲突

☐ 不符合道德

☐ 欺诈
- ○ 编造数据 ○ 虚假出版 ○ 其他:

☐ 抄袭 ☐ 自我抄袭 ☐ 重复抄袭 ☐ 重复发表*

☑ 侵权 ☐ 其他法律相关:

☐ 编辑错误
- ○操作错误 ○ 无效评审 ○ 决策错误 ○ 其他:

☐ 其他原因:

出版结果 (只可单选)

☐ 仍然有效

☑ 完全无效

作者行为失误(只可单选)

☑ 诚信问题

☐ 学术不端

☐ 无 (不适用此条，如编辑错误)

* 重复发表: "出版或试图出版同一篇文章于不同期刊."

Hans汉斯

历史
作者回应:
□ 是，日期:
☑ 否

信息改正:
□ 是，日期:
☑ 否

说明:

“ 基于深度学习的视频语音提取文本系统设计与实现”一文刊登在 2021 年 10 月出版的《软件工程与应用》2021 年第 10 卷第 4 期第 528-541 页上。因该论文属团队全体成果，并非作者个人成果，文章署名不实，侵占团队权益，作者主动申请撤稿。编委会现决定撤除此稿件，保留原出版出处：

文章引用: 谢煜颖. 基于深度学习的视频语音提取文本系统设计与实现[J]. 软件工程与应用, 2021, 10(4): 528-541. https://doi.org/10.12677/SEA.2021.104057

所有作者签名：谢煜颖

Software Engineering and Applications 软件工程与应用, 2021, 10(4), 528-541
Published Online August 2021 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2021.104057

Hans汉斯

基于深度学习的视频语音提取文本系统设计与实现

谢煜颖

浙江理工大学信息学院，浙江 杭州
Email: 735308107@qq.com

收稿日期：2021年6月29日；录用日期：2021年8月2日；发布日期：2021年8月9日

摘　要

21世纪是信息化的时代，多媒体技术在网络教学中的应用越来越普及。在新冠疫情防控形势严峻的时期，网络教学凭借得天独厚的优势起到了重大作用。但当前市场上的在线视频编辑平台功能单一、效率低下、用户体验繁琐，本文利用基于循环神经网络和卷积神经网络实现的语音识别对视频进行文本提取，并且使用注意力机制算法实现的语音合成对视频文本的修改，使用Ffmpeg对视频进行处理，同时使用多线程和异步队列提升系统性能。本文主要针对如何实现语音识别和语音合成，以及如何提升语音合成效果做了主要的研究，最终实现识别普通话的准确率为90.52%，以及声音合成近乎为ground truth的合成引擎，将其应用于产品实现了能对视频精准编辑，体验良好的视频语音提取文本系统。

关键词

语音识别，语音合成，视频处理，深度学习

Design and Implementation of Video Speech Extraction Text System Based on Deep Learning

Yuying Xie

School of Information, Zhejiang Sci-Tech University, Hangzhou Zhejiang
Email: 735308107@qq.com

Received: Jun. 29th, 2021; accepted: Aug. 2nd, 2021; published: Aug. 9th, 2021

Abstract

The 21st century is an information age, and the application of multimedia technology in network

文章引用: 谢煜颖. 基于深度学习的视频语音提取文本系统设计与实现[J]. 软件工程与应用, 2021, 10(4): 528-541.
DOI: 10.12677/sea.2021.104057

teaching is becoming more and more popular. In the severe period of COVID-19 prevention and control, online teaching has played an important role by virtue of its unique advantages. However, the current online video editing platform in the market has the disadvantages of single function, low efficiency and cumbersome user experience. This paper uses speech recognition based on cyclic neural network and convolutional neural network to extract the video text, uses speech synthesis realized by attention mechanism algorithm to modify the video text, and uses FFmpeg to process the video, it also uses multithreading and asynchronous queues to improve system performance. This paper mainly focuses on how to realize speech recognition and speech synthesis, and how to improve the effect of speech synthesis. Finally, the accuracy rate of Mandarin recognition is 90.52%, and the sound synthesis engine is close to the ground truth, which can be applied to the product to realize accurate video editing and experience a good video-speech extraction text system.

Keywords

Speech Recognition, Speech Synthesis, Video Processing, Deep Learning

1. 引言

相对于线下教育，远程线上教育有着得天独厚的优势，正有越来越多的人形成线上学习的习惯。当今社会教育资源分布不均的问题依然存在，并难以解决，而网络教学给了所有人接受到高质量教学的平等机会，不仅是学生群体，很多走上岗位的人员也会通过网络学习提高自己。教学视频处理是实施网络在线教育的第一步，因此，随着网络教育的市场不断扩大，视频处理平台也将随之蓬勃发展，市场规模会越来越大，因此开发一个系统的教学视频语音提取文本系统非常重要。

网络教育市场不断扩大，行业内教学平台百花齐放，但普遍缺乏核心竞争力，传统的视频处理平台功能单一、效率低下、用户操作繁琐。在激烈的竞争中，开发视频语音提取和语音合成功能，实现极速音文转换，提高教学视频质量无疑会占据有利地位。功能强大而又操作简便的视频语音提取，既能凭借较高的视频质量吸引学生，又能凭借良好的用户体验吸引教师。该系统的开发必定能帮助网络教学平台迅速发展[1]。

视频语音提取文本系统的核心业务是实现教师对教学视频语音的精准编辑。系统分为网页端和后台管理系统。网页端为讲师提供视频仓库模块、视频切片模块、视频编辑模块、语音合成模块、审计模块和个人中心模块等功能。后台管理系统为系统管理员提供系统管理、用户管理、角色管理、视频仓库管理、视频编辑管理、语音合成管理和审计管理等功能。系统采用 B/S 架构，采用服务网格的设计模式。前端使用 uni-app 实现，并用 Echarts 实现数据的可视化报表审计。后端整体采用 java 编写，使用 Spring Boot 框架，并用 MyBatis-Plus 实现数据持久化，使用 Redis 缓存数据并使用 RabbitMQ 异步处理任务，提升系统性能，使用 k8s 和 Istio 部署语音识别和语音合成模块，与后端项目一起组成完整的视频语音提取文本系统[2]。

针对主要语音识别和语音合成这两个主要的功能分别采用 DeepSpeech2 算法和 FastSpeech2 算法实现，在数据集上分别收集了 3000 小时的视频和 85 小时的音频进行训练，最终实现了对普通话识别率高达 90.52%，声音合成 MOS 值高达 3.83。

2. 相关技术简介

2.1. FFmpeg 实现视频的处理

服务器采用 FFmpeg 对用户上传的视频进行格式修改，频率修改等操作。FFmpeg 是领先的多媒体框架，能够解码、编码、转码、混合、解密、流媒体、过滤和播放人类和机器创造的几乎所有东西。它支持最晦涩的古老格式，直到最尖端的格式。无论它们是由某个标准委员会、社区还是公司设计的。它还具有高度的便携性。

FFmpeg 可以在 Linux、Mac OS X、Microsoft Windows、BSDs、Solaris 等各种构建环境、机器架构和配置下编译、运行，并通过测试基础设施 FATE。

它包含了 libavcodec、libavutil、libavformat、libavfilter、libavdevice、libswscale 和 libswresample，可以被应用程序使用。还有 FFmpeg、FFplay 和 FFprobe，可以被终端用户用于转码和播放。

2.2. Spring Boot 框架

后端使用 Spring Boot 开发。Spring Boot 是由 Pivotal 团队在 Spring 框架基础上提供的全新框架，Spring Boot 的设计目的是用来简化新 Spring 应用的初始搭建以及开发过程。通过约定大于配置是它的理念，开发人员可以极快的编写 web 应用。

2.3. Istio 服务网格

Istio 是一个完全开源的服务网格，作为透明的一层接入到现有的分布式应用程序里。它也是一个平台，拥有可以集成任何日志、遥测和策略系统的 API 接口。Istio 多样化的特性使您能够成功且高效地运行分布式微服务架构，并提供保护、连接和监控微服务的统一方法。

3. 语音功能实现及算法分析

3.1. 语音识别实现

3.1.1. 整体方案

在语音识别上，调研了从动态时间规整模型，到混合高斯－隐马尔可夫模型，再到采用深度学习的端到端语音识别模型，最终采用百度开源的 PaddlePaddle 框架，在 DeepSpeech2 的基础上开发而来。采用端到端的语音识别技术，使用 CTC 损失函数，采用多层 CNN、RNN 深度神经网络模型训练而成，相比传统的 DNN/HMM 实现的语音识别有较大的准确率和高效率的提升[3]。语音识别的模型图如图 1 所示：

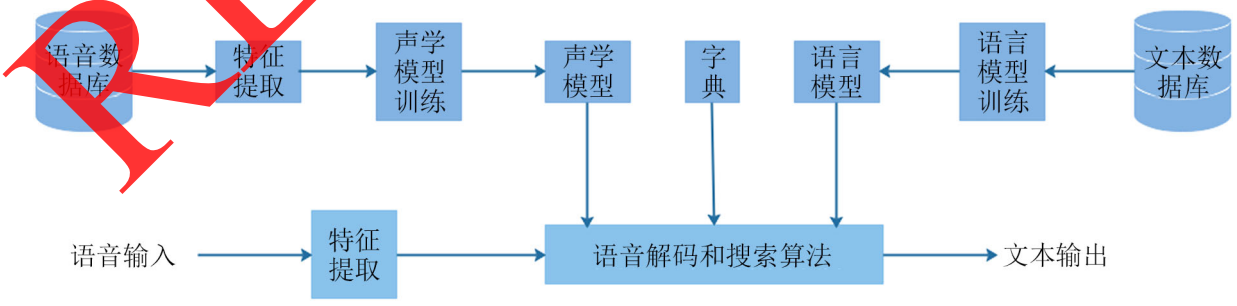

Figure 1. Speech recognition model diagram
图 1. 语音识别模型图

具体描述如下：

1) 特征提取：将语音数据库中的音频信息提取为计算机能识别的向量数据，是进行后续神经网络训练的基础。特性提取时，我们有常用的特征参数作为提取模板主要有两种，分别是线性预测系数(LPCC)和梅尔倒谱系数(MFCC)。LPCC 的基本思想是，当前时刻的信号可以用若干个历史时刻的信号的线性组合来估计，通过使实际语音的采样值和线性预测采样值之间达到均方差最小，即可得到一组线性预测系数。MFCC 则是受人的听觉系统研究成果推动而导出的声学特征，利用同态处理方法，对语音信号求离散傅立叶变换后取对数，再求反变换就可得到倒谱系数。相对于 MFCC 来说，LPCC 的实现更为简单，处理速度等快。同时特征处理的数据还将进行指数归一化和预处理等操作。

2) 声学模型：一个好的算法模型能逼近理论上限。项目组针对语音识别技术进行调研，采用学术界流行的识别框架和新颖的技术实现，并结合业务实际整合修改训练集以实现针对教学视频的语音识别系统。具体到实际项目组需要解决的三个问题：

a) 如何识别中英文混合的音频序列输入。

b) 在识别准确的前提下，如何做到快速识别，降低用户等待时间。

c) 针对特定的教学视频，涉及到大量的专业名词，以及发音人的地方口音，大大增加了识别的难度，如何获取大量的数据集以及修改模型以适应中文声学模型的自训练成为了不小的挑战。

结合以上三个问题并对比了经典的几个语音识别模型，动态时间规整(DTW)、混合高斯模型－隐马尔科夫模型(GMM-HMM)、上下文相关的深度神经网络－隐马尔可夫模型(CD-DNN-HMM)以及基于端到端的 CNN-RNN-CTC 模型。最后采用了模型体积小，识别率高，速度快的 CNN-RNN-CTC 模型，将声音频谱序列作为输入，即可得到相应的中文字符[2] [3]。

3) 语言模型：考虑到只使用 CTC 的端到端语音识别技术在准确上还有提高的空间，为了进一步符合识别的上下文语境，引入了语音模型，来提高识别的准确率。这里我们使用了百度提供的已经训练好的语音模型。

解码和搜索算法：将经过 RNN 计算后的数据使用 softmax 激活函数进行处理，得到输出序列的概率，结合语言模型，需要找到一条所有概率之积最大的最优路径，有实现集束搜索(ctc_beam_search)、贪婪策略(ctc_greedy)这两种方法，集束搜索占用的内存空间较大，速度稍慢但准确率高，贪婪搜索速度快、占用空间小但准确率欠佳。在训练时使用贪心搜索来测试字错误率，实际应用则使用集束搜索提升准确性。

结合该项目改进的语音识别系统具体模型流程如图 2 所示。

3.1.2. 算法原理

1、CTC 损失函数

传统的语音识别需要对输入的每一帧音频特征所对应的音素进行标注，即给定输入序列 $X[X1,X2,\cdots,Xt]$ 以及对应的标签数据 $Y[Y1,Y2,\cdots,Yu]$(例如语音识别中的音频文件和文本文件)，我们的工作是找到 X 到 Y 的一个映射。语音对齐的过程本身就需要进行反复多次的迭代，来确保对齐更准确，这本身就是一个比较耗时的工作。

采用 CTC 作为损失函数的声学模型训练，是一种完全端到端的声学模型训练，不需要预先对数据做对齐，只需要一个输入序列和一个输出序列即可以训练。

为了解决对齐的问题，CTC 引入了空白字符 ε，语音识别中的停顿均表示为 ε。例如“你好”可以表示为“你你 ε 好好 $\varepsilon\varepsilon$”，所以，CTC 的对齐涉及去除重复字母和去除 ε 两部分。也就是说，对应标签 Y，其关于输入 X 的后验概率可以表示为所有映射为 Y 的路径之和，我们的目标就是最大化 Y 关于 $X = Y$ 的后验概率$p(Y|X)$。假设每个时间片的输出是相互独立的，则路径的后验概率是每个时间片概率的累积，公式及其详细含义如下。

Figure 2. End to end speech recognition model flow
图 2. 端到端语音识别模型流程

$$p(Y|X) = \sum_{A \in A_{X,Y}} \prod_{t=1}^{T} Pt(at|X)$$

上面的 CTC 算法存在性能问题，对于一个时间片长度为 T 的 N 分类任务，所有可能的路径数为 T^N，在很多情况下，这几乎是一个宇宙级别的数字，用于计算 Loss 几乎是不现实的。在 CTC 中采用了动态规划的思想来对查找路径进行剪枝，算法的核心思想是如果路径 π_1 和路径 π_2 在时间片 t 之前的输出均相等，我们就可以提前合并他们，如图 3 所示。

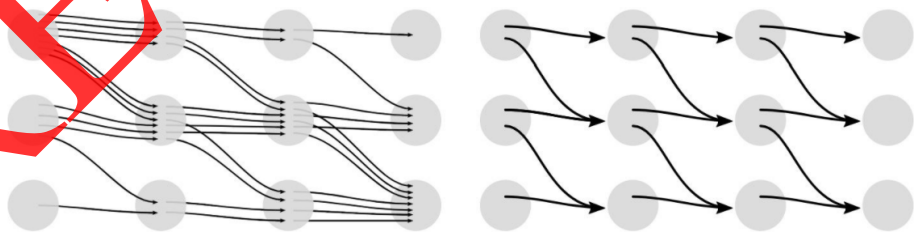

Figure 3. Merging diagram
图 3. 合并图解

由于$p(Y|X)$的计算只涉及加法和乘法，因此其一定是可导函数，进而我们可以使用 SGD 优化模型。对于数据集 D，模型的优化目标是最小化负对数似然如下：

$$\sum_{(X,Y) \in D} -\log p(Y|X)$$

当我们训练好一个 RNN 模型时，给定一个输入序列 X，我们需要找到最可能的输出，也就是求解：

$$Y^* = \arg\max_Y p(Y|X)$$

对 CTC 输出结果解码有两种方案，一种是 Greedy Search，第二种是 beam search。

2、CNN

卷积神经网络(CNN)是受到人类视觉神经系统的启发提出的。CNN 有 2 大特点，能够有效的将大数据量的向量降维成小数据量，能够有效的保留特征信息，符合音频处理的原则。CNN 主要有三部分组成，如图 4 所示：

Figure 4. CNN diagram
图 4. CNN 图

首先是卷积层，结合本项目，就是采用线性预测系数(LPC)将输入的音频提取特征向量。

对于池化层，引入了 Batch Normalization 算法，对每一层 RNN 的输出结果进行批归一化处理，加快 RNN 训练的收敛速度。

全连接层将经过多次 RNN 处理后的数据进行输出。

3、GRU

GRU (Gate Recurrent Unit)是循环神经网络(Recurrent Neural Network, RNN)的一种。和 LSTM(Long-Short Term Memory)一样，也是为了解决长期记忆和反向传播中的梯度等问题而提出来的。GRU 和 LSTM 在很多情况下实际表现上相差无几，但是 GRU 更加方便计算。GRU 的输入输出结构与普通的 RNN 是一样的。有一个当前的输入 x^t，和上一个节点传递下来的隐状态(hidden state) h^{t-1}，这个隐状态包含了之前节点的相关信息。结合 x^t 和 h^{t-1}，GRU 会得到当前隐藏节点的输出 y^t 和传递给下一个节点的隐状态 h^t。GRU 的内部结构如图 5 所示。

其中图中的⊙是 Hadamard Product，也就是操作矩阵中对应的元素相乘，因此要求两个相乘矩阵是同型的，⊕则代表进行矩阵加法操作。

3.2. 语音合成实现

3.2.1. 整体方案

考虑到项目需要实现声音复刻，不单单是简单语音合成，所以模型的训练时间和所需的数据集不能过于复杂，因此我们对比了现有的语音合成算法，采用传统的自回归模型的 Tacotron2 模型，使用自注意力机制的 FastSpeech2 模型，最终比较了实现效果后，选择了训练时间更短、效果更好的 FastSpeech2 算法。最终得到下面的语音合成模型，如图 6 所示：

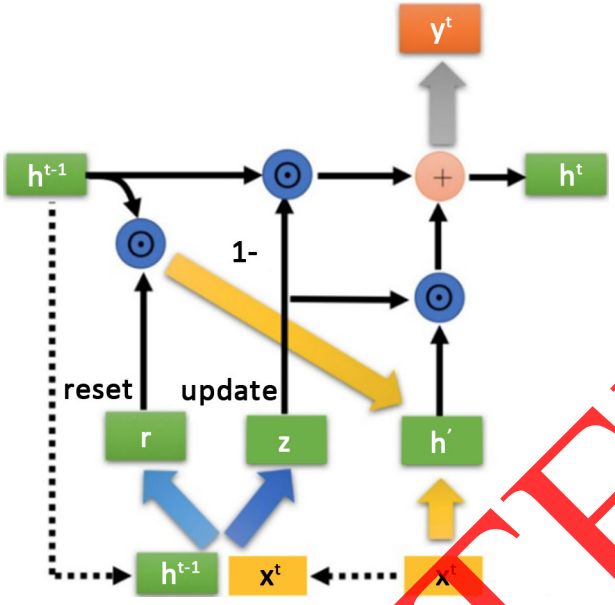

Figure 5. The internal structure of GRU
图 5. GRU 的内部结构

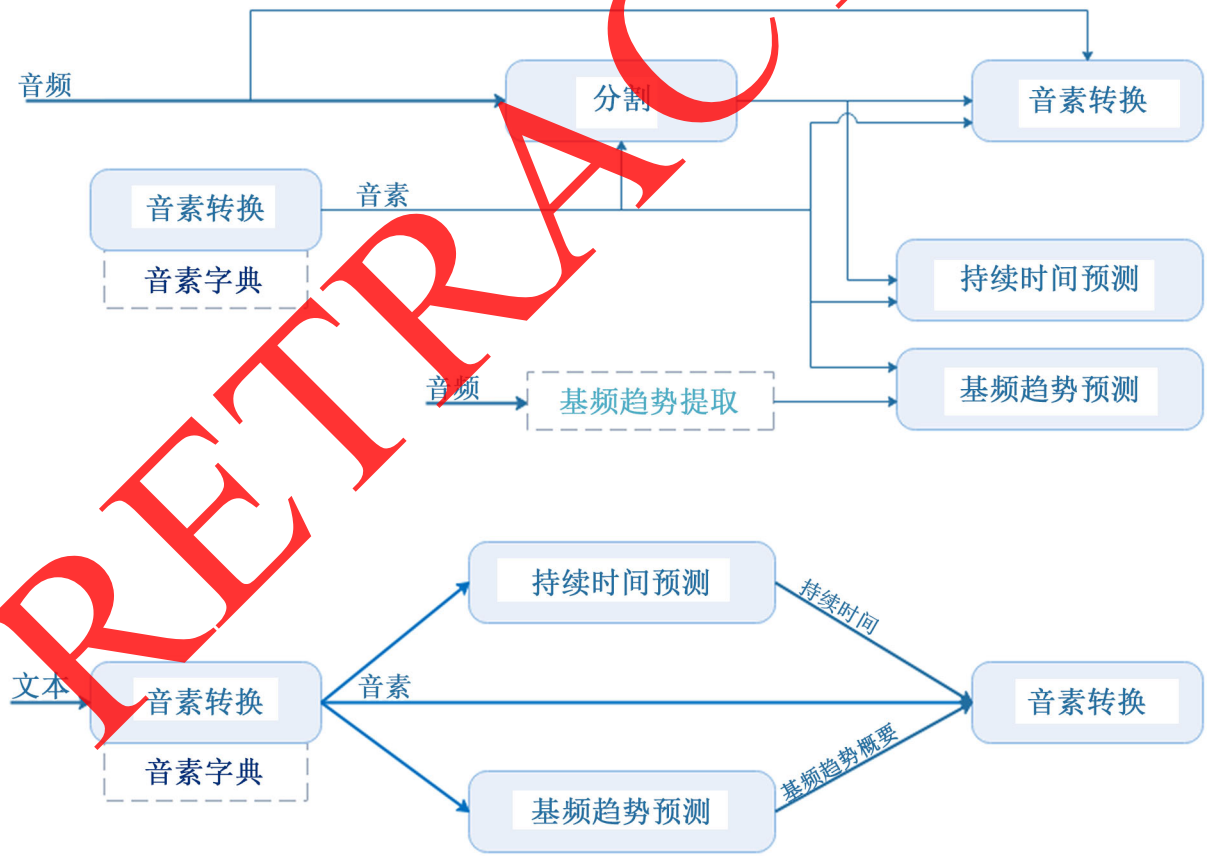

Figure 6. Speech synthesis model
图 6. 语音合成模型

主要涉及如下几个模型[4]：

Grapheme-to-Phoneme 模型

字素到音素模型是基于(Yao & Zweig, 2015) [5]开发的编码器－解码器架构。然而，我们使用了具有门控递归单元(GRU)非线性的多层双向编码器和同样深度的单向 GRU 解码器。每个解码器层的初始状态初始化为对应编码器转发层的最终隐藏状态。该体系结构采用教师强制训练，采用波束搜索进行解码。使用 3 个双向层，每个层 1024 个单位在编码器和 3 个单向层。

Segmentation 模型

分割模型，经过训练输出给定的话语和目标音素序列之间的对齐。用 CTC 训练生成音素序列的网络将为每个输出音素产生简短的峰值。粗略地将音素与音频对齐，但不足以检测到精确的深音：实时神经 TTS 音素边界。因此采用训练预测音素对的序列而不是单个音素。然后，该网络将倾向于在接近一对音素之间边界的时间步输出音素对。在解码方面，采用波束搜索进行解码。

音素持续时间和基频模型

使用单一的结构来共同预测音素持续时间和随时间变化的基频。该模型的输入是一个带有重音的音素序列，每个音素和重音都被编码为一个热载体。该架构包括两个完全连接的层，每个层有 256 个单元，然后是两个单向循环层，每个层有 128 个 GRU 单元，最后是一个完全连接的输出层。在初始全连通层和最后一个循环层之后，应用概率为 0.8 的 Dropout。最后一层对每个输入音素产生三种估计：音素持续时间、音素清音的概率(即具有基频)和 20 个时间相关的 F0 值，这些值在预测持续时间内均匀采样。该模型通过最小化音素持续时间误差、基频误差、音素清音概率的负对数似然以及与 F0 相对于时间的绝对变化成比例的惩罚项来实现平滑，从而使音素持续时间误差、基频误差、音素清音概率的负对数似然以及与 F0 相对于时间的绝对变化成比例的联合损失达到最优。

音频合成模型

音频合成模型是 WaveNet 的一个变种。结合该项目特征，我们用 Wprev 和 Wcur 将卷积分解为每个时间步的两个矩阵乘法。这些层由残余连接连接起来。每一层的隐藏状态连接到一个 r 向量，用 Wskip 投影到 s skip 通道，然后用 relu 非线性进行两层 11 个卷积。WaveNet 是一种利用神经网络对原始音频波形建模的技术，仿照 PixelRNN 图像生成方式，依据之前采样点来生成下一个采样点。其相当于将典型的统计参数语音合成技术后端的声学模型和声码器合并，生成音频。我们发现，通过双向准 rnn (QRNN)层对输入进行编码，重复到所需频率进行上采样，模型会表现得更好，训练得更快，需要的参数更少。

通过对说话人声音的音高，能量以及音素持续时间的特征捕捉和训练，最终可以实现模拟说话人，合成特定的声音。

3.2.2. 算法原理

1、GE2E

Generalized end-to-end (GE2E) loss 是谷歌在论文 attention-based model (TE2E)上提出的新损失函数，与 TE2E loss 和 Triplet loss 相比，它每次更新都和多个人相比，因此能使训练时间更短，说话人验证精度更高。

其基本思路如图 7 所示，挑选 N 个人，每人 M 句话，通过图示的顺序排列组成 Batch，接着通过 LSTM 神经网络提取 N*M 句话的特征嵌入层，然后求取每个特征嵌入层和每个人平均特征嵌入层的相似度，得到一个相似度矩阵。最后通过最小化 GE2E loss 使得相似矩阵中有颜色的相似度尽可能大，灰色的相似度尽可能小，即本人的声纹特征应该和本人每句话的特征尽可能相近，和其他人的特征尽可能远离，从而训练 LSTM 网络。

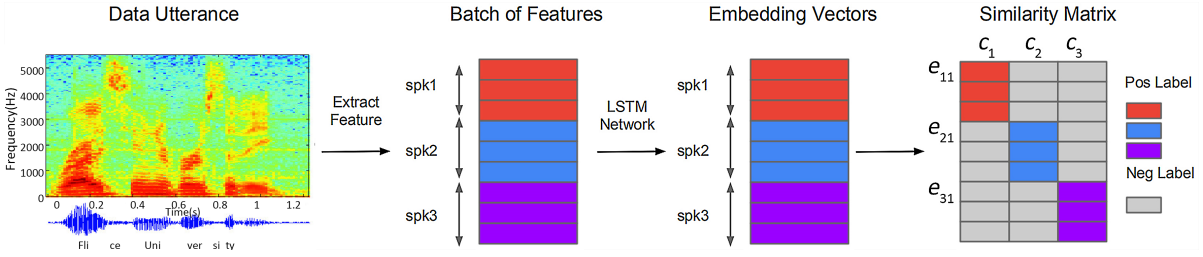

Figure 7. Basic idea of GE2E
图 7. GE2E 基本思路

相似度矩阵的定义[6]如下公式所示($1 \le j \le N, 1 \le i \le M, 1 \le k \le N$)：

$$S_{ji,k} = w \cdot \cos\left(e_{ji}, c_k\right) + b$$

其 e_{ji} 中表示第 j 人第 i 句话对应的特征嵌入层，w 和 b 是要训练的参数($w > 0$)，C_k 是第 k 人的特征嵌入层，由 M 句话的特征嵌入层求平均得到，即

$$C_k = \frac{1}{M}\sum_{m-1}^{M} e_{km}$$

为了使得相似度矩阵中有颜色的相似度尽可能大，灰色的相似度尽可能小，有两种损失函数 softmax loss 和 contrast loss，分别如下所示：

$$L_s\left(e_{ji}\right) = -\log\frac{\exp\left(S_{ji}, j\right)}{\sum_{k-1}^{M}\exp\left(S_{ji}, k\right)}$$

$$L_c\left(e_{ji}\right) = 1 - \sigma\left(S_{ji}, j\right) + \max_{1 \le k \le N, k \ne j} \sigma\left(S_{ji,j}\right)$$

最小化损失函数，即本人和本人的每一句话都比较相似，和其他人最相似的地方要尽可能小。GE2E loss 定义为以上两种损失函数之和：

$$L_G = \sum_{j,1}\left(L_s\left(e_{ji}\right) + Lc\left(e_{ji}\right)\right)$$

此外，为了训练的稳定性，论文中建议在计算本人和本人某句话相似度的时候，不要让该句话的特征嵌入层来参与计算本人的特征嵌入层，即实际上：

$$c_j^{(-i)} = \frac{1}{M-1}\sum_{m=1, m\ne i}^{M} e_{jm}$$

$$S_{ji,k} = \begin{cases} w \cdot \cos\left(e_{ji}, c_j^{(-i)} + b\right) \& \; if \;\; k = j; \\ w \cdot \cos\left(e_{ji}, c_j^{(-i)} + b\right) \& \; otherwise. \end{cases}$$

2、WaveGlow

随着神经网路的发展，目前常使用到的生成模型可有：自回归模型(Autoregressive model)、生成对抗网路(GAN)以及基于流的生成模型(Flow-based generative model)。尽管自回归模型在许多实验上得到了很好的效果，但这种一次生成一个样本的生成方式除了需要庞大的计算资源之外，在可平行性上也受到了限制。相对而言，生成对抗网路则免除了这种烦恼，主要透过生成器与判别器不断相互学习的迭代从而生得与真实样本接近的分布。但一直以来生成对抗网路也不免会遇到许多问题，像是生成的多样性不足以及训练过程不稳定等等。不过幸运地，基于流的生成模型有效地解决了这

些问题，而这样的生成方式，也被采用在声码器 WaveGlow 中。WaveGlow 透过分布采样生成语音，仅需一个网络及一个最大化似然的损失函数即可生成时域波形，并且在高还原度的情况下亦能即时的合成语音。利用多个可逆的变换函数组成序列，将一个简单的分布透过一系列的可逆函数转换到一个复杂的分布，并借此来模拟训练数据的分布，最后再透过最大似然准则来进行优化。其网络模型图如图 8 所示：

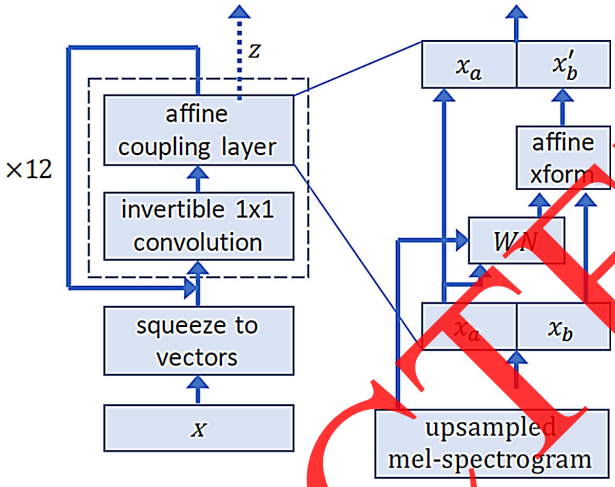

Figure 8. Basic flow chart of WaveGlow
图 8. WaveGlow 基本流程图

4. 系统设计与实现

教学视频语音提取文本系统是一个基于 HTML5，一套代码适用于多个平台，集用户 Web 端、后台管理系统端两大模块为一体[7] [8] [9] [10] [11]，并为不同端口提供了不一样的服务和权限，系统功能图如图 9 所示：

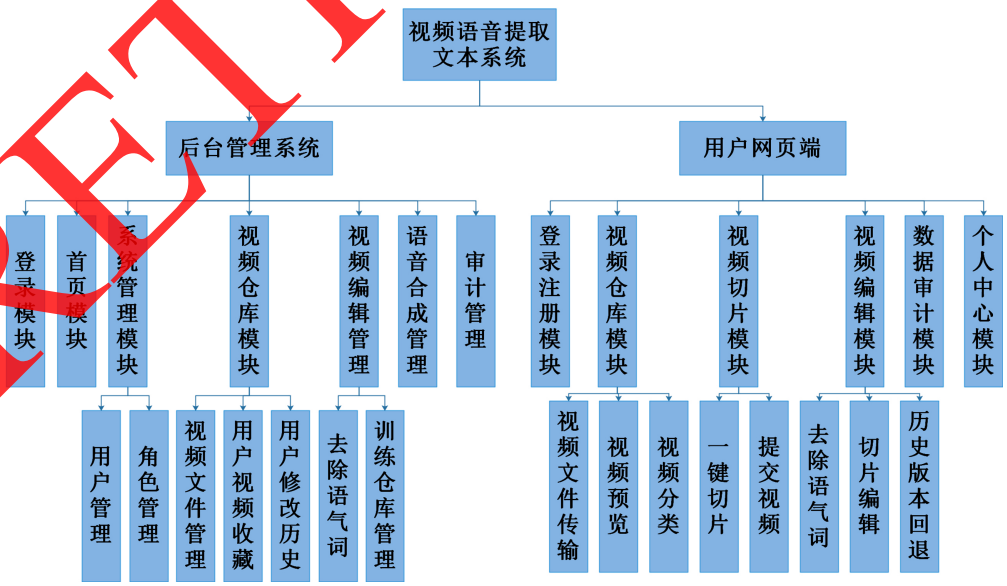

Figure 9. System function diagram
图 9. 系统功能图

4.1. 后台管理系统

1) 登录模块：在该模块中，后台管理员输入账号和密码登录进入系统，对数据进行操作。

2) 首页模块：在该模块中可以对平台系统日志进行查看和管理，将访问量、用户数量、用户在线时长、收入、消息、订单等内容通过比例图、柱形图、折线图进行可视化，简洁明了，提升了系统管理员对数据的理解。

3) 系统管理模块：系统管理分为用户管理和角色管理，审核用户资料，维护平台普通用户和会员用户的资料信息。

4) 视频仓库管理模块：本模块分为视频文件管理、用户视频收藏、用户修改历史管理三个模块。可以管理用户的各类型视频。

5) 视频编辑管理模块：分为一键去除语气词管理、切片信息管理，管理用户所编辑的视频。

6) 语音合成管理模块：系统管理员管理用户的声音模型，保证用户可以选择平台上的声音模型实现语音合成，最终使得课程视频几乎没有修改痕迹，提升学生的听课体验。

7) 审计管理：对用户的审计信息进行管理。

4.2. 用户 Web 端

1) 登录注册模块：门户网站主要服务于教师用户，以简单可视化的方式轻松调动底层复杂的程序，为用户提供了一个简单但高效的管理用户账号程序。

2) 视频仓库模块：视频仓库模块提供视频文件管理功能，用户可以进行视频文件的上传、下载、删除等基本功能，此外本模块还提供了收藏、预览、分类等其他便捷功能，实现效果如图 10 所示。

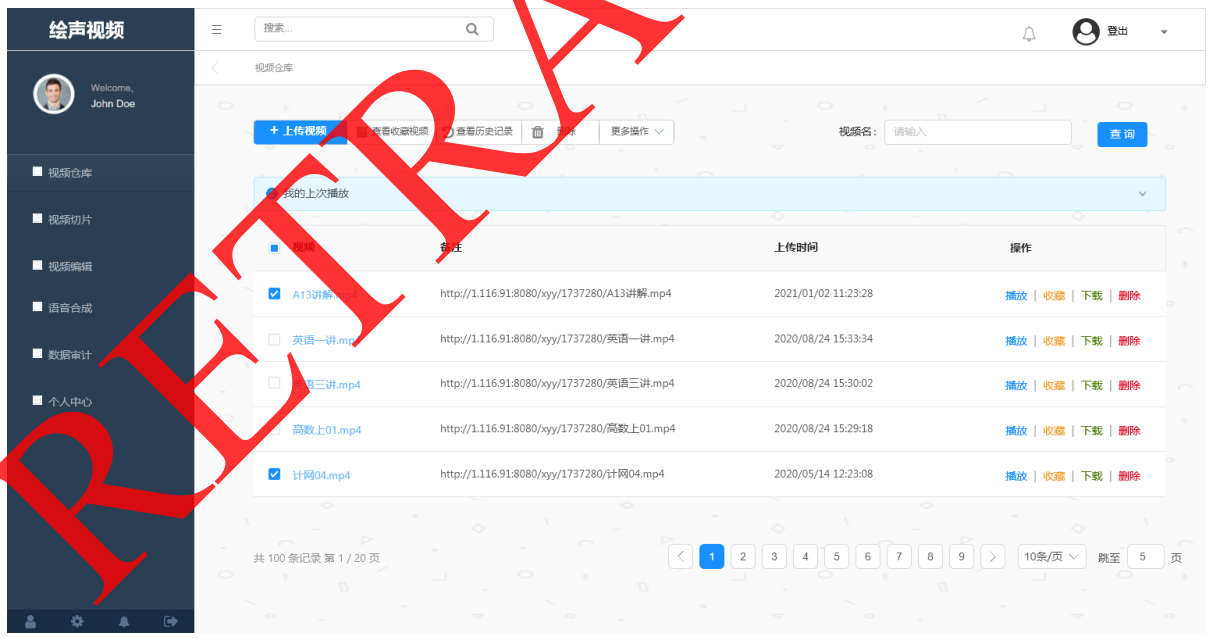

Figure 10. Video warehouse interface
图 10. 视频仓库界面

3) 视频切片模块：视频切片时，完全由用户手动切片的话工作量十分巨大，用户可以手动对视频切片，也可以在运行本模块提供的一键切片功能后进行手动调整，极大地提高了用户体验与工作效率；同时用户也可以回到总览界面对所有切片进行管理操作，进行简单的遍历和检查，实现效果如图 11~12

所示。

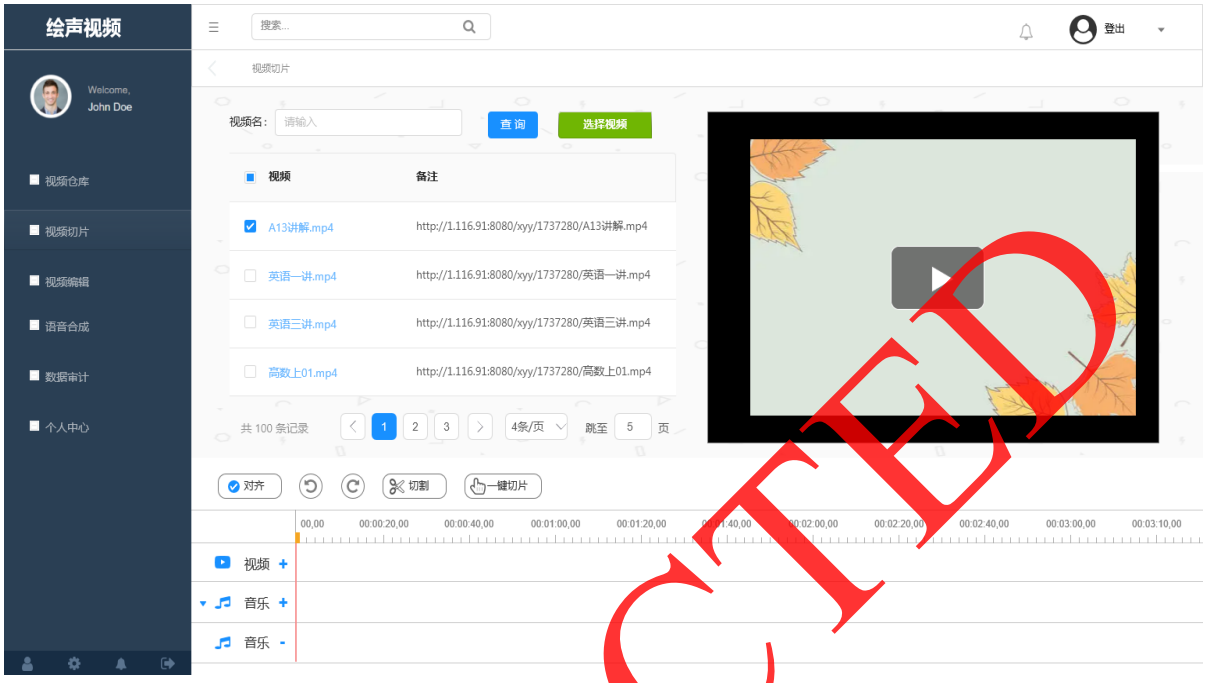

Figure 11. Video slice interface
图 11. 视频切片界面

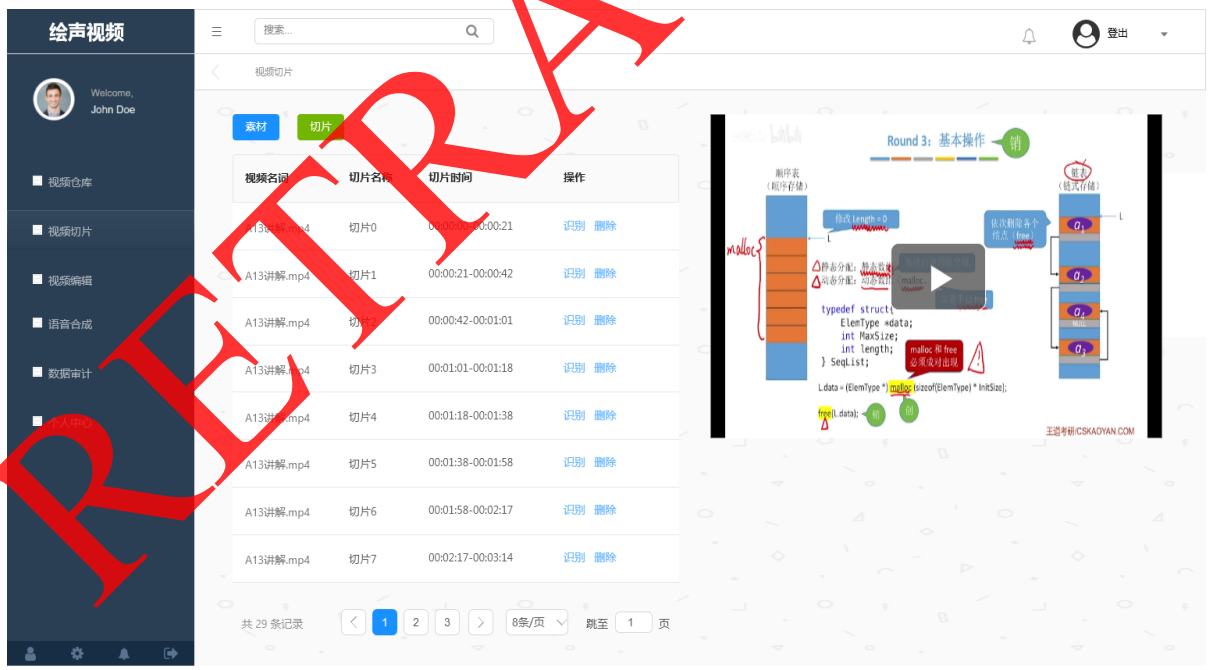

Figure 12. Video slice overview interface
图 12. 视频切片总览界面

4) 视频编辑模块：视频编辑模块是该项目的核心模块，提供对视频切片文字识别提取展示和错误音频编辑，其中错误音频修改包括文本智能语音合成和录音替换两种方式，实现效果如图 13 所示。

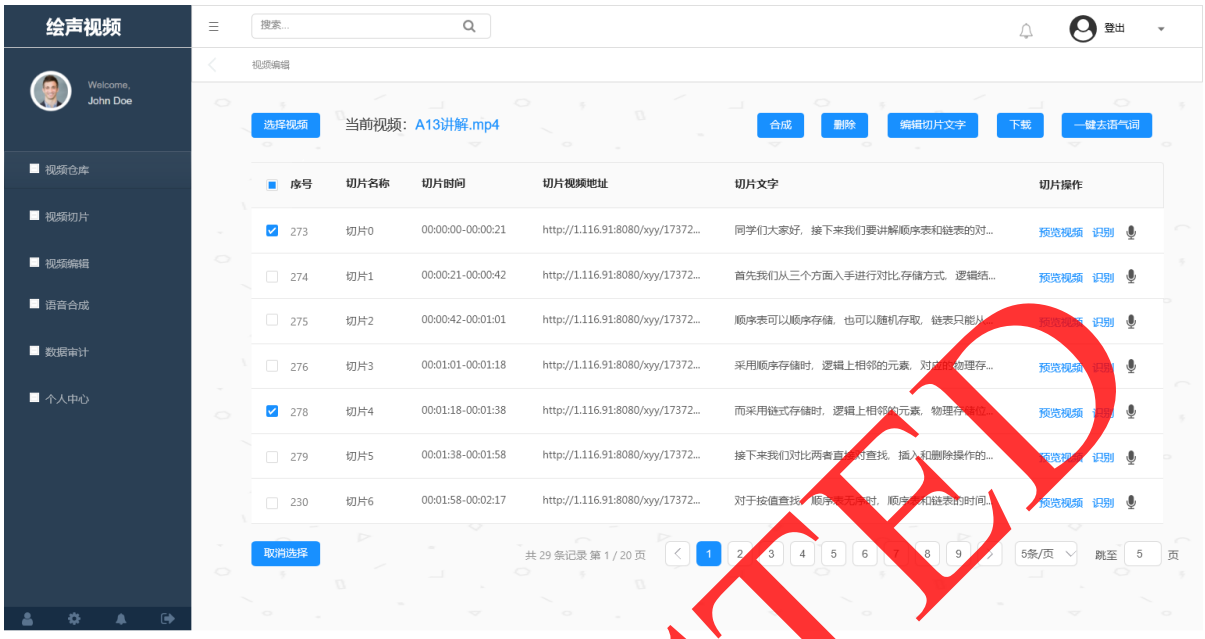

Figure 13. Video editing interface
图 13. 视频编辑界面

5) 语音合成模块：普通的文字转语音存在机器语音明显的情况，为了更好的符合用户体验，使用语音合成技术，提供多种智能语音模型供用户选择，从而实现更加逼真的替换效果。

6) 数据审计模块：可语教学视频语音提取文本系统中，讲师可以对上传的教学视频进行审计，审计内容为视频修改内容，审计功能以报表形式和 Web 图形化形式体现，并支持批量审计。

7) 个人中心模块：用户在个人中心模块可以查看个人信息，当遇到问题时，可以在帮助中心向平台助手咨询相关问题。

5. 结束语

该文章主要提出了当前视频编辑平台的一些痛点问题以及给出了项目的整体设计方案和具体的语音识别和语音合成的实现方案。结合 k8s 和服务网格实现云上方便快捷部署，真正做到语音识别精准高效，实时进行音文转换，视频传输格式多样，一键切片方便快捷，多元展示审计信息，用户数据形象直观，平台助手引导提示，用户使用简便高效等特点。系统在设计上高内聚，低耦合，可以横向扩展，分布式部署。该平台的实现必将提升网络教学视频的质量，推动教育行业信息化、数字化和智能化转型，以全流程数据聚合及智能运用，实现高效互联互通[12]。

参考文献

[1] 杨辰雨, 庄磊. 语音合成技术及其在金融场景下的应用[J]. 中国金融电脑, 2021(6): 43-46.

[2] 潘丽鹏. 嵌入式英语语音识别控制系统研究[J]. 微型电脑应用, 2021, 37(6): 73-75.

[3] 卢林, 王东. 浅谈声音识别模型发展趋势[J]. 汽车实用技术, 2021, 46(12): 186-188.

[4] 鱼昆, 张绍阳, 侯佳正, 张少博. 语音识别及端到端技术现状及展望[J]. 计算机系统应用, 2021, 30(3):14-23.

[5] Yao, K.S. and Zweig, G. (2015) Sequence-to-Sequence Neural Net Models for Grapheme-to-Phoneme Conversion. https://dblp.uni-trier.de/rec/journals/corr/YaoZ15.html

[6] 胡亚军. 基于神经网络的统计参数语音合成方法研究[D]: [博士学术论文]. 合肥: 中国科学技术大学, 2018.

[7] 胡新月. 语音识别技术在软件工程中的应用[J]. 电子技术与软件工程, 2021(4): 240-241.
[8] 魏伟华. 语音合成技术综述及研究现状[J]. 软件, 2020, 41(12): 214-217.
[9] 吴大非. MOOC 管理平台的设计与开发[J]. 电脑知识与技术, 2018, 14(27): 47-49+52.
[10] 潘丽鹏. 嵌入式英语语音识别控制系统研究[J]. 微型电脑应用, 2021, 37(6): 73-75.
[11] 马莉, 朱永胜, 王晓刚. 基于语音识别技术的复杂超声检查报告智能生成系统设计[J]. 现代医用影像学, 2021, 30(5): 928-930.
[12] 冯君. 基于 Android 平台的语音识别技术应用研究[J]. 铜陵职业技术学院学报, 2021, 20(1): 62-65+82.