基于双向 GRU+CTC 的赣南客家方言语音识别系统

沈思嘉[1]，江斌[1,2*]，魏炳辉[1]，陈诗逸[1]，庄新凯[1]
1. 赣南科技学院，江西 赣州 341000
2. 赣州市文化数字化传承与智能创新重点实验室，江西 赣州 341000
DOI: 10.61369/TACS.2026020043

摘　　要： 赣南客家方言作为客家语的重要分支，具有独特的语言特征与文化价值，但受限于语料稀缺和模型迁移困难，其语音识别研究仍处于初期。针对这一问题，本文设计了一种基于双向 GRU（Bidirectional GRU）与 CTC（Connectionist Temporal Classification）的赣南客家方言语音识别系统。系统由语音识别、方言翻译和旅游推荐三大模块构成，其中语音识别模块利用双向 GRU 提取语音前后文特征，并结合 CTC 实现端到端识别。后端采用 Python 与 Flask 框架构建 API，结合 MongoDB 数据库实现语料存储与文本查询，前端基于微信小程序实现交互。实验结果表明，该系统在低资源条件下仍能取得较高识别准确率，验证了模型在方言识别任务中的有效性。研究为方言的智能化处理与文化传承提供了新思路。

关 键 词： 双向 GRU；CTC；赣南客家方言；微信小程序

Gannan Hakka Dialect Speech Recognition System Based on Bidirectional GRU + CTC

Shen Sijia[1], Jiang Bin[1,2*], Wei Binghui[1], Chen Shiyi[1], Zhuang Xinkai[1]
1. Gannan University of Science and Technology, Ganzhou, Jiangxi 341000
2. Ganzhou Key Laboratory of Cultural Digital Inheritance and Intelligent Innovation, Ganzhou, Jiangxi 341000

Abstract: As an important branch of the Hakka language, the Gannan Hakka dialect possesses unique linguistic features and cultural values. However, restricted by the scarcity of corpus and difficulties in model migration, research on its speech recognition is still in the initial stage. To address this problem, this paper designs a Gannan Hakka dialect speech recognition system based on Bidirectional Gated Recurrent Unit (Bidirectional GRU) and Connectionist Temporal Classification (CTC). The system consists of three modules: speech recognition, dialect translation, and tourism recommendation. Among them, the speech recognition module uses Bidirectional GRU to extract contextual features of speech and combines CTC to achieve end-to-end recognition. The backend employs Python and the Flask framework to build APIs, and integrates MongoDB for corpus storage and text query, while the frontend implements interaction based on WeChat Mini Program. Experimental results show that the system can still achieve high recognition accuracy under low-resource conditions, verifying the effectiveness of the model in dialect recognition tasks. This study provides a new idea for the intelligent processing and cultural inheritance of dialects.

Keywords: bidirectional GRU; CTC; Gannan Hakka dialect; WeChat mini program

引言

语言是文化传承的核心载体，赣南客家方言保留唐宋古汉语特征，是研究汉语演变的 “活化石”，其传承对客家文化与非遗保护意义重大。受城市化、普通话普及等影响，该方言使用萎缩、传承断层，年轻群体掌握度持续下滑，数字化保护刻不容缓。但其口音复杂、语料稀缺、地域变体繁多，传统识别技术难以适配。为此，本研究提出双向 GRU 与 CTC 融合的端到端语音识别方案，可有效适配小样本方言识别场景，提升识别准确率与鲁棒性，并研发集成识别功能与红色旅游推荐模块的小程序，打通方言保护、文旅宣传与文化传播链路，为濒危方言活态传承提供技术支撑与实践参考。

项目信息：赣南科技学院大学生创新创业训练计划项目（S202413434023）。
作者简介：
沈思嘉，女，本科在读，研究方向为计算机应用；
江斌，男，通讯作者，助教，硕士研究生，研究方向为自然语言处理；
魏炳辉，男，副教授，博士研究生，研究方向为图像处理；
陈诗逸，女，本科在读，研究方向为计算机应用；
庄新凯，男，本科在读，研究方向为计算机应用。

一、技术介绍

门控循环单元是 LSTM 的变种，是相比于 LSTM 网络更简单的循环神经网络[1]。GRU 网络通过门控机制完成信息的更新过程，反观 LSTM 网络，其输入门与遗忘门具有互补属性，这使得直接采用两个独立的门结构存在一定冗余性。GRU 将输入门与和遗忘门合并成更新门（Update gate），并引入了重置门（Reset gate），同时，GRU 也不引入额外的记忆单元，直接在当前状态和历史状态之间引入线性依赖关系[2]。GRU 是 RNN 的拓展，可有效解决 RNN 存在的梯度问题[3]，双向 GRU 由正向 GRU 和反向 GRU 组成，能够同时捕捉序列数据的前向和后向依赖关系，从而更全面地理解序列的上下文信息。

CTC 是一种用于序列学习的损失函数，特别适用于输入和输出序列长度不一致且难以手动对齐的任务，如语音识别、手写文字识别等。它直接将输入音频序列映射到单词或其他建模单元（如音素和字符）的系统， 极大简化了语音识别模型的构建和训练[4]。CTC 的核心思想是通过引入一个特殊的空白标签，允许模型在输出序列中插入空白标签，从而自动对齐输入和输出序列。CTC 方法主要分为两个部分，一个是训练时的标签对齐和损失函数 loss 值的计算，一个是推理解码时计算得到最终的预测结果。其中，解码部分分为三种，贪婪法 (Greedy)、束搜索法 (Beam Search) 和前缀束搜索法 (Prefix Beam Search)。其中，最常用的则是简单高效的贪婪法。通过在每个时间步上寻找概率最大的值作为最终预测的输出并去掉连续重复和空白块即可。

二、语音识别系统的模块设计

赣南客家方言语音识别系统以 Pytorch 作为主要开发框架，使用 FlaskAPI 框架进行系统开发。该系统主要包括3个模块：语音识别模块、方言翻译模块以及旅游推荐模块。其中，语音识别模块负责将输入的赣南客家方言语音转化为文本；方言翻译模块用于精准解析转化后文本的语义；旅游推荐模块为用户量身定制个性化的旅游推荐方案。系统框架图具体如图1所示。

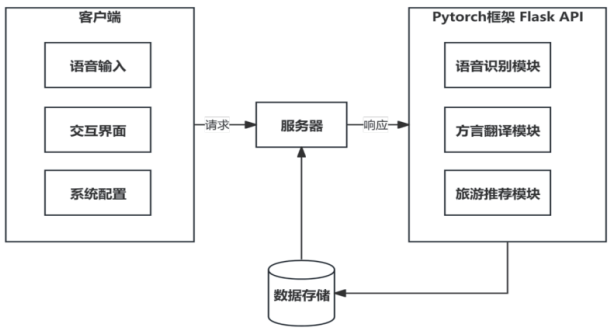

图1 系统框架图

（一）语音识别模块

针对赣南客家方言语音识别研究匮乏、数字化进程滞后的问题，本文提出一种基于双向 GRU 神经网络的语音识别方案。该方案先通过梅尔频率倒谱系数（MFCC）算法从原始语音中提取核心声学特征，再将特征输入融合连接时序分类（CTC）模型的识别网络，实现语音到文本的精准转译。

系统基于 PyTorch 框架构建，核心为 GRU 网络与 CTC 损失函数，涵盖四大功能模块：MFCC 语音特征提取模块抽取声学判别特征；CTC-GRU 训练模块经反向传播优化模型参数；CTC 贪婪解码模块将模型输出转化为可读文本；预测评估模块以准确率、词错误率等指标验证模型性能。

（二）方言翻译模块

“客译”（KeYi）是基于微信生态开发的客家方言 — 普通话互译小程序。项目以微信开发者工具为开发环境，前端采用微信原生 WXML 与 WXSS 技术实现界面布局与样式渲染；后端基于 Python 语言及 Flask 框架搭建 API 服务，完成语音识别、文本翻译与数据交互等核心逻辑。

系统选用 MongoDB 作为数据库，依托其灵活的文档型数据结构，高效存储与管理方言翻译语料，并与 Flask 后端无缝对接，实现翻译文本的快速检索、匹配与转换，保障实时准确的语言转换服务。

小程序集成语音识别、旅游推荐等功能模块，各模块遵循统一接口规范实现数据交互，保证系统架构的一致性与可维护性。

（三）旅游推荐模块

旅游推荐模块整合抖音、小红书、微博、马蜂窝等主流社交及旅游媒体资源，实现一站式内容导流。深度聚合全网热度信息，智能匹配用户偏好，打造千人千面的个性化出行灵感库。

用户可在小程序内直达各平台指定内容，提升旅游信息获取效率。模块遵循统一封装、平台适配、失败回退设计原则，兼顾用户体验、系统可扩展性与多平台兼容性。

用户通过首页 “旅游” 入口进入推荐页面，页面采用视觉化设计，搭配清新配色营造轻松浏览体验，并通过清晰导航栏实现功能快速定位。

三、系统实现

（一）语音识别模块

针对当前赣南客家方言公开标准语料库匮乏的问题，本研究立足语音识别技术需求，专项构建方言专用语料库。语料库选取方言特色突出的日常用语、亲属称谓两大核心板块，为适配语音识别场景，严格控制单条语音时长≤10秒，最终形成包含121条有效样本的小样本数据集。

该语料库遵循标准化流程搭建，全程严把质量关：第一步为音频预处理，依托专业工具完成格式标准化转换，结合语义完整性、语音韵律规范精准切割音频，确保单条语料语义连贯、节奏贴合识别要求；第二步为双重标注与规范统一，选聘兼具赣南客家方言与普通话专业素养的学者，同步开展方言拼音标注、普通话译文转写，并制定统一标注细则，最大限度消减标注偏差；第三步为质量管控核验，采用多人交叉核查、随机抽样复核的双重质控模式，校验标注精准度，为后续客家方言语音识别研究提供高质量、高可靠性的数据支撑。

为直观展示赣南客家话与普通话的对应关系，表 1 列出部分语料示例，涵盖客家话原文、客家话发音标注、对应普通话原文及普通话发音四方面信息：

表 1 部分客家话与普通话发音对比

赣南客家话	客家话发音标注	对应普通话	普通话发音
阿姐	a1 zia3	姐姐	jie3 jie3
阿公	a1 gong4	爷爷	ye2 ye2
赖子	lai4 zi1	儿子	er2 zi1
食朝	shik1 zhao1	吃早饭	chi1 zao3 fan4
得闲嘛	de2 xian2 ma1	有空吗	you3 kong4 ma1

语音特征提取是方言语音识别系统的核心环节，其目标是从原始音频信号中提取能够有效表征方言语音特性的关键信息。针对赣南客家方言的复杂音韵系统（如古汉语残留声母、入声韵尾、多声调变化等），本研究设计了多层次、多维度的特征提取方案，具体包括基础声学特征提取、特征增强与融合策略，以及针对方言特性的优化方法。本文将梅尔滤波器的数量由传统的 26 个扩充至 40 个，旨在更精准地捕捉赣南客家方言中高频入声韵尾的细微特征。同时，融合一阶差分（Δ - MFCC）与二阶差分（Δ^2 - MFCC），以此刻画语音特征随时间变化的动态信息，例如声调的转折以及连读时的变调现象。此外，对提取的特征实施均值方差归一化处理，以此消除不同说话人以及录音环境所带来的音量差异。

本文基于自建赣南客家话语音数据集，采用双向 GRU 网络结合语速扰动、频谱掩码策略构建完整识别模型，并以词错率（WER）作为核心评估指标开展实验验证。词错率是语音识别、机器翻译等自然语言处理任务的通用量化评估指标，用于表征模型输出序列与人工标注参考序列的偏差程度，通过统计模型输出错误词数占参考文本总词数的比值实现偏差量化。整体语音识别流程如图2所示。

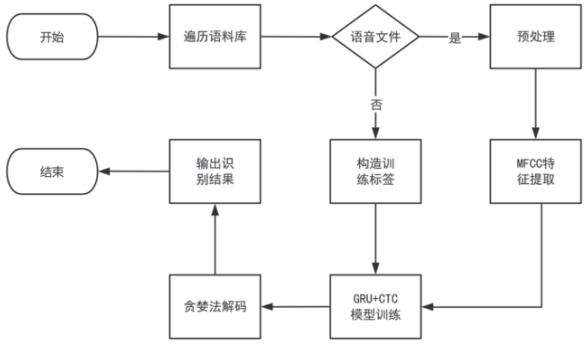

图2 语音识别流程图

实验统计结果显示，该模型在赣南客家方言数据集上的词错率（WER）为35.67%，整体识别准确性较好，针对常用生活词汇与短句的识别表现尤为稳定。

针对数据集内“嗦粉”“酿豆腐”“圩日”等方言特色词汇，模型召回率达91.3%；针对10字以上长句语音，召回率为88.6%，表明模型可精准捕捉方言标志性词汇，且语速扰动技术有效缓解了长句语速波动带来的识别偏差，较好保留长语音序列信息完整性。

模型整体精确率为90.5%，其中纯净无噪语音场景精确率达94.8%，含集市、家庭等轻微环境噪声的语音精确率为87.2%；频谱掩码技术的引入，大幅提升了模型噪声适配能力，有效降低噪声干扰引发的误识别率。

（二）方言翻译模块

相对于传统的 APP，微信小程序由于无需下载安装，不占手机存储空间，受到用户欢迎，因此，项目选择微信小程序开发。为了确保小程序的功能稳定和用户体验良好，在项目中选择了以下一系列工具和技术：

(1) 小程序框架设计：WXML 定义小程序页面结构与内容层级，WXSS 控制视觉样式与布局，JavaScript 实现交互逻辑，三者结合实现小程序页面灵活布局、样式设计及动态交互响应，保障前端功能高效稳定。

(2) 数据库设计与语料管理：小程序后端采用 MongoDB 存储管理客家方言与普通话对照语料，语音识别文本结果生成后，系统调用该数据库快速检索匹配，返回对应翻译文本至前端。

(3)API 接口设计：后端采用 Python 语言与 Flask 框架构建。通过 Flask，系统能够后端基于 Python+Flask 构建，通过 Flask 创建多 API 端点，接收前端语音识别请求、执行翻译查询、调用数据库并返回结果；结合 Python 第三方库，集成机器学习 / 自然语言处理模块，实现高效的语音识别与文本翻译。

在“客译”小程序中，翻译流程图如图3所示：

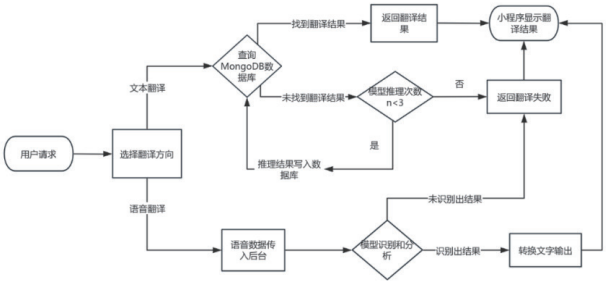

图3 翻译流程图

(1) 语音识别：用户点击小程序 “开始录音” 按钮，调用 wx.getRecorderManager () 采集麦克风声音并保存为 wav 格式音频，经 wx.uploadFile () 上传至后端 POST /voice 接口；核心的客家话语音识别模块基于 GRU+CTC 深度学习模型（经大量客家话语音数据训练），通过 Flask+RESTful API（路径 /voice）、Nginx+uWSGI 部署，将语音转文本后传入语料库对比模块。

(2) 对比语料库：识别文本与 MongoDB 语料库匹配，匹配成功返回翻译结果，失败则提示翻译失败。

(3) 翻译结果处理：匹配成功时，小程序通过 wx.request () 调用 /voice 接口接收 JSON 结果，经 this.setData () 更新 WXML 页面输出翻译结果；匹配失败则返回 “翻译错误” 提示。

（三）旅游推荐模块

本模块阐述 “客译” 小程序旅游信息搜索浏览流程，核心运用 JavaScript、微信小程序 API、事件绑定、数据驱动模型，通过函数封装与配置化设计实现可扩展的旅游攻略平台跳转系统，

逻辑层支撑事件响应、平台访问控制等核心功能，是实现多平台内容聚合的关键。系统采用 Page () 构造器模块化封装页面逻辑，依托数据驱动视图渲染思想，定义 platforms 数据对象存储平台标识与访问路径；用户点击首页 “旅游” 入口时，系统通过 bindtap 属性绑定点击事件与逻辑层函数，实现界面层与逻辑层双向事件通信。系统流程图如图 4 所示：

图 4 旅游景点推荐流程图

在浏览旅游推荐列表后，用户可以选择感兴趣的旅游网站。用户选择旅游网站后，小程序会将该网站的相关信息复制到剪贴板，并且自动跳转到浏览器打开，这些信息包括网站链接：旅游网站的 URL 地址；网站名称：旅游网站的名称。

这个模块与“客译”小程序的翻译功能紧密结合，用户可以在浏览旅游信息的同时，使用“客译”小程序的翻译功能，解决语言障碍，更好地了解旅游目的地的文化和风俗。

四、结语

本文提出了一种基于双向 GRU+CTC 的赣南客家方言语音识别系统，并且详细介绍了系统的三个主要模块的设计与实现。系统后端通过 Python 和 FlaskAPI 框架实现，在语音识别模块中，采用了深度学习技术，实现了相应的 GRU 和 CTC 模型，提高了识别的准确率。

参考文献

[1]Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J/OL].ArXiv Preprin (2018-08-13) [2020-12-29]. http://arxiv.org/abs/1412.3555

[2] 张德政 , 范欣欣 , 谢永红 , 蒋彦钊 . 基于 ALBERT 与双向 GRU 的中医脏腑定位模型 [J]. 工程科学学报 , 2021, 43(9): 1182-1189. DOI: 10.13374/j.issn2095-9389.2021.01.13.002

[3] 杨国亮 , 习浩 , 龚家仁 , 温钧林 . 基于 Transformer 的短时交通流时空预测 [J]. 计算机应用与软件 , 2024, 41(3): 169-173,225. DOI: 10.3969/j.issn.1000-386x.2024.03.026

[4] 谢旭康 , 陈戈 , 孙俊 , 等 . TCN-Transformer-CTC 的端到端语音识别 [J]. 计算机应用研究 , 2022, 39 (3): 699-703.

