分类号：　　密　　级：公开
学　号：20232108050　　单位代码：10759

石河子大学

硕士学位论文

少数民族学生普通话发音评测及辅助学习研究与系统实现

学位申请人　赵小锋
指导教师　于宝华 教授
申请学位类别　专业硕士
专业名称　电子信息
研究领域　计算机技术
所在学院　信息科学与技术学院

中国·新疆·石河子
2026 年 5 月

分类号：　　密　　级：公开
学　号：20232108050　　单位代码：10759

石河子大学

硕 士 学 位 论 文

少数民族学生普通话发音评测及辅助学习研究与系统实现

学 位 申 请 人　赵小锋
指 导 教 师　于宝华 教授
申请学位类别　专业硕士
专 业 名 称　电子信息
研 究 领 域　计算机技术
所 在 学 院　信息科学与技术学院

中国·新疆·石河子
2026 年 5 月

Research and System Implementation on Mandarin Pronunciation Assessment and Assisted Learning for Ethnic Minority Students

A Dissertation Submitted to

Shihezi University

In Partial Fulfillment of the Requirements

for the Degree of

Master of Engineering

By

Zhao Xiao-feng

Electronic Information

Dissertation Supervisor:Prof. Yu Bao-hua

May,2026

石河子大学学位论文独创性声明及使用授权声明

学位论文独创性声明

本人所呈交的学位论文是在我导师的指导下进行的研究工作及取得的研究成果。据我所知，除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确的说明并表示谢意。

研究生签名：  时间： 2026 年 5 月 25 日

使用授权声明

本人完全了解石河子大学有关保留、使用学位论文的规定，学校有权保留学位论文并向国家主管部门或指定机构送交论文的电子版和纸质版。有权将学位论文在学校图书馆保存并允许被查阅。有权自行或许可他人将学位论文编入有关数据库提供检索服务。有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

研究生签名：  时间： 2026 年 5 月 25 日

导师签名：  时间： 2026 年 5 月 25 日

摘要

普通话作为国家通用语言，在教育教学、社会交往和公共服务中发挥着重要作用。对于少数民族学生而言，受母语迁移、语言环境差异及教学资源分布不均等因素影响，普通话学习过程中普遍存在发音评测粒度较粗、错误定位不够精确和错误反馈形式单一等问题。传统依赖教师人工听辨的评测方式具有主观性强、效率低、反馈滞后等缺点，难以满足大规模、个性化和低成本的学习需求。基于此，本文以语音识别与语音合成技术为基础，围绕少数民族学生在普通话学习中的发音评测、示范反馈和辅助训练等需求，开展普通话发音评测方法、标准示范语音生成模型及辅助学习系统研究。主要工作如下：

（1）针对传统普通话发音评测方法难以实现细粒度偏误识别与定位的问题，为获得稳定的音素级建模能力，本文构建了 Paraformer-RNN-T 端到端普通话音素识别模型，通过前端增强、多尺度特征建模、轻量化编码和剪枝解码等结构实现学习者语音到音素序列的映射。在此基础上设计并引入音素混淆先验的 P-SW 音素序列比对算法，实现发音结果的细粒度识别与结构化标注。实验结果表明，该模型在 AISHELL1-PHONEME 数据集上的音素错误率为 4.85%，实时因子为 0.026，参数量为 66.4M；在迁移到本文自建数据 MPAD-EMS 上并微调后，音素错误率由 32.03%降低至 5.73%。基于 P-SW 的发音评测方法在精确率、召回率和 F1 值等指标上均优于对比方法，最优模型组合的 F1 值达到 0.6837。

（2）针对普通话学习过程中缺乏高质量示范读音的问题，本文以 VITS 为基线框架，构建了基于 VITS-BigVGAN 的端到端中文语音合成模型，并围绕文本建模与生成器结构两个方面对模型进行了优化设计。在文本处理阶段引入多感受野特征建模策略和位置编码机制缓解多尺度声学特征捕获能力不足等问题，在波形生成过程中采用 BigVGAN 生成器提升高频细节恢复能力和整体听感自然度。改进模型在客观指标上均优于基线模型，主观 MOS 评分由 4.24 提升至 4.36，能够为普通话学习场景提供清晰、稳定且可重复的标准示范语音。

（3）针对普通话发音学习中评测、反馈与示范等环节相互割裂的问题，本文设计并实现了一套面向少数民族学生的普通话发音辅助学习系统。系统采用 Web 管理端与微信小程序学习端相结合的架构，实现了发音练习、自动评测、错误反馈和标准读音示范等核心功能，系统集成本文提出的发音评测模型与语音合成模型，形成从练习输入到自动评测、错误分析和示范反馈的完整闭环。系统测试结果表明，该系统运行稳定、功能完整，能够满足少数民族学生在普通话学习中的实际应用需求，验证了本文研究成果的工程可行性与应用价值。

关键词：普通话发音评测；音素识别；语音合成

Abstract

Mandarin, as the national common language of China, plays an important role in education, social interaction, and public services. For ethnic minority students, Mandarin learning is often affected by mother tongue transfer, differences in language environment, and the uneven distribution of teaching resources. As a result, common problems include coarse-grained pronunciation assessment, inaccurate error localization, and limited forms of error feedback. Traditional assessment methods that rely on teachers' auditory judgment suffer from strong subjectivity, low efficiency, and delayed feedback, making it difficult to meet the needs of large-scale, personalized, and low-cost learning. To address these issues, this thesis is based on speech recognition and speech synthesis technologies, and focuses on the needs of pronunciation assessment, demonstration feedback, and assisted training for ethnic minority students in Mandarin learning. The main work is summarized as follows.

(1) To address the difficulty of achieving fine-grained error recognition and localization in traditional Mandarin pronunciation assessment, a Paraformer-RNN-T end-to-end Mandarin phoneme recognition model is constructed to obtain stable phoneme-level modeling capability. The model maps learner speech to phoneme sequences through front-end enhancement, multi-scale feature modeling, lightweight encoding, and pruned decoding. On this basis, a P-SW phoneme sequence alignment algorithm incorporating phoneme confusion priors is designed to achieve fine-grained recognition and structured annotation of pronunciation results. Experimental results show that the proposed model achieves a phoneme error rate of 4.85%, a real-time factor of 0.026, and 66.4M parameters on the AISHELL1-PHONEME dataset. After being transferred to the self-built MPAD-EMS dataset and fine-tuned, the phoneme error rate decreases from 32.03% to 5.73%. The P-SW-based pronunciation assessment method outperforms comparative methods in precision, recall, and F1 score, with the best model combination achieving an F1 score of 0.6837.

(2) To address the lack of high-quality standard pronunciation demonstrations in Mandarin learning, this thesis takes VITS as the baseline framework and constructs an end-to-end Chinese speech synthesis model based on VITS-BigVGAN. The model is optimized from two aspects, namely text modeling and generator structure. In the text processing stage, a multi-receptive-field feature modeling strategy and a positional encoding mechanism are introduced to alleviate the insufficient capability of capturing multi-scale acoustic features. In the waveform generation stage, the BigVGAN generator is adopted to improve high-frequency detail restoration and the overall naturalness of listening quality. The improved

model outperforms the baseline model on objective metrics, and the subjective MOS score increases from 4.24 to 4.36, enabling it to provide clear, stable, and repeatable standard demonstration speech for Mandarin learning scenarios.

(3) To address the fragmentation among assessment, feedback, and demonstration in Mandarin pronunciation learning, this thesis designs and implements a Mandarin pronunciation-assisted learning system for ethnic minority students. The system adopts an architecture combining a Web-based management terminal and a WeChat Mini Program learning terminal, and implements core functions including pronunciation practice, automatic assessment, error feedback, and standard pronunciation demonstration. By integrating the proposed pronunciation assessment model and speech synthesis model, the system forms a complete closed loop from practice input to automatic assessment, error analysis, and demonstration feedback. System testing results show that the system runs stably and provides complete functionality. It can meet the practical needs of ethnic minority students in Mandarin learning, verifying the engineering feasibility and application value of the proposed research.

Key words: Mandarin Pronunciation Evaluation; Phoneme Recognition; Speech Synthesis

目录

摘要 I

Abstract II

第 1 章 绪论 1

1.1 研究背景及意义 1

1.2 国内外研究现状 2

1.2.1 语音识别技术研究现状 2

1.2.2 发音评测技术研究现状 5

1.2.3 语音合成技术研究现状 7

1.3 本文研究内容及技术路线 9

1.3.1 研究内容 9

1.3.2 技术路线 10

1.4 组织结构 11

第 2 章 相关理论与关键技术 13

2.1 发音评测方法 13

2.1.1 特征比对 13

2.1.2 GOP 算法 13

2.1.3 语音识别 14

2.2 语音识别技术 15

2.2.1 Transformer 15

2.2.2 Conformer 16

2.2.3 Paraformer 17

2.3 语音合成技术 18

2.3.1 Tacotron 系列 18

2.3.2 FastSpeech 系列 20

2.3.3 VITS 模型 22

2.4 本章小结 23

第 3 章 基于音素识别的普通话发音评测方法研究 24

3.1 引言 24

3.2 MPAD-EMS 数据集构建 24

3.2.1 语料采集设计 26

3.2.2 音频与文本处理流程 26

3.2.3 数据标注方案设计 27
3.3 Paraformer-RNN-T 模型构建 28
3.3.1 音频前端增强模块 29
3.3.2 多尺度特征提取与编码器设计 31
3.3.3 解码器与损失函数设计 34
3.4 P-SW 音素序列比对算法设计 37
3.4.1 得分矩阵构建与评分策略 37
3.4.2 音素混淆判定与替换惩罚建模 38
3.4.3 算法执行流程与伪代码实现 39
3.5 少数民族学生普通话发音评测模型工作流程设计 41
3.6 实验结果与分析 42
3.6.1 实验数据集 42
3.6.2 实验环境与配置 43
3.6.3 评价指标 43
3.6.4 音素识别实验 44
3.6.5 发音评测实验 47
3.7 本章小结 49
第 4 章 基于 VITS-BigVGAN 的端到端中文语音合成模型研究 50
4.1 引言 50
4.2 VITS-BigVGAN 模型架构设计 51
4.2.1 多感受野块 52
4.2.2 位置编码 53
4.2.3 BigVGAN 生成器 55
4.3 实验分析 57
4.3.1 数据准备 57
4.3.2 训练配置 57
4.3.3 评价指标 58
4.3.4 实验结果对比 60
4.4 消融实验 63
4.4.1 消融实验设置 64
4.4.2 单模块对模型性能的影响分析 64
4.4.3 多模块联合效果与协同分析 65
4.4.4 消融实验小结 66
4.5 本章小结 66

第 5 章 普通话发音辅助学习系统设计与实现 68
5.1 引言 68
5.2 系统需求分析 68
5.3 系统总体设计 69
5.3.1 系统架构设计 69
5.3.2 功能模块设计 71
5.3.3 数据库设计 72
5.4 系统功能实现 73
5.4.1 基础信息管理模块 73
5.4.2 发音评测与反馈模块 75
5.4.3 系统功能管理模块 76
5.4.4 系统开发及部署环境 78
5.5 系统功能测试 79
5.5.1 基础信息管理模块测试 79
5.5.2 发音评测与反馈模块测试 79
5.5.3 系统功能管理模块测试 80
5.6 本章小结 81
第 6 章 总结与展望 82
6.1 全文总结 82
6.2 未来展望 83
参考文献 84
致谢 90
作者简介 91

第 1 章 绪论

1.1 研究背景及意义

语言能力是个体参与社会交流和获取教育资源的重要基础。随着经济社会的快速发展，国家通用语言在教育教学、社会交往和公共服务中的作用日益凸显。我国是统一的多民族、多语言、多方言国家，普通话作为国家通用语言的重要载体，在促进沟通交流、增强社会凝聚力以及消除语言隔阂方面具有重要意义。习近平总书记在中央第七次西藏工作座谈会和第三次中央新疆工作座谈会上均强调了在民族地区加强国家通用语言文字教育的重要作用，分别指出“国家通用语言文字教育要从娃娃抓起”“要把加强国家通用语言文字教育作为关键性、基础性工作来抓，作为做好民族地区工作的长久之策、固本之举来抓”[1]。因此在基础教育阶段开展普通话能力培养，并建立相对客观、可重复的发音训练与评价方式，具有明确的现实需求与应用价值。

然而，学习者的母语或地方方言与普通话在音系与韵律特征上往往存在较大差异，普通话学习过程中容易出现由母语迁移引发的系统性发音偏误。此类偏误若缺乏及时、准确地纠正与持续反馈，可能在学习初期逐步固化，进而影响普通话口语表达的整体水平。同时传统学校教学场景中发音训练与评测通常依赖教师听辨与经验判断，受师资力量、课堂时长与个体化辅导成本等因素制约，难以在大规模教学条件下持续提供细粒度、可量化且可回溯的纠错反馈[2]。尤其对于低年级学习者而言，持续的发音训练与自我纠错能力尚不成熟，更需要外部反馈帮助其建立稳定的发音对照与纠正路径。因此，亟需面向学习者的自动化发音评测与反馈机制研究。

近年来，人工智能的深度应用进一步推动了语音识别与语音合成技术在建模能力方面的研究进展，为计算机辅助语言学习提供了可行的技术基础。对于发音评测任务而言，现有系统在实际教学应用中仍面临两类关键挑战：其一，评测结果需要具备足够的可解释性与可定位性，能够在较细粒度上指出错误位置，以支持后续纠正；其二，评测结果需要与学习反馈有效衔接，使学习者在获得诊断信息后能够及时进行对照与重复练习，从而提升学习效率。此外，针对少数民族学习者还需要考虑学习者语音与标准普通话语料之间的分布差异带来的适配问题，避免模型在目标场景中泛化能力不足而影响评测稳定性。

除发音错误诊断外，标准读音示范也是普通话发音训练中的重要环节。仅指出学习者发音错误位置并不足以完成有效纠错，还需要为其提供稳定、清晰且可重复的标准发

音参考，帮助其在对照模仿中不断调整发音。语音合成技术能够即时生成标准普通话读音，弥补传统教学中示范资源不足、反馈链条不完整的问题。因此，本文将自动发音评测与标准语音合成协同设计，为普通话教学带来更加客观高效的解决方案，对培养学生的语言沟通能力和提高普通话发音水平具有重要的现实意义。

1.2 国内外研究现状

随着人工智能与互联网通信技术的发展，语音识别已成为人机交互的重要基础能力，并在教育、医疗、媒体传播与智能终端等场景中得到广泛应用。在语言教育领域，计算机辅助语言学习（Computer-Assisted Language Learning, CALL）逐步从传统的规则驱动与人工评测模式转向数据驱动的自动评测与反馈模式。发音评测技术依托声学建模与序列识别能力，可为学习者提供自动化、可量化且可复现的语音能力评价结果，因而成为 CALL 研究中的重要方向之一。

在普通话发音评测研究中，评测建模粒度的选择直接影响偏误定位与纠错反馈的精确度。相较于字符或词级建模，音素作为语音系统中更细粒度的对齐单元，能够更直接地刻画学习者在音位替换、插入、删除以及相近音素混淆等方面的发音差异，更适用于分析由母语音位体系迁移所引发的结构性偏误现象[3]。基于音素序列的发音评测方法通常包含两个关键推理环节：首先通过声学模型将学习者语音解码为音素序列；随后采用序列对齐与相似度计算等方法建立学习者发音与标准发音之间的对应关系，从而实现偏误检测、类型标注与评分输出。

此外，为提升评测结果的教学可解释性与学习反馈效率，近年来相关研究开始关注将标准读音示范引入评测系统，通过语音合成（Text-to-Speech, TTS）生成稳定、可重复的标准普通话语音，支持学习者对照练习与即时纠正，从而构建评测—示范—反馈的闭环学习流程。本节将从模型构建与系统落地两个视角出发，围绕语音识别模型、发音评测模型和语音合成模型三个方向，对国内外研究现状与技术演进进行梳理，为面向民族地区的普通话发音评测与辅助学习系统设计提供理论依据与技术参考。

1.2.1 语音识别技术研究现状

语音识别是将人类语音信号转换为文本的技术，广泛应用于语音评测、智能助手和自动翻译等领域。早期的语音识别系统主要基于以 GMM-HMM（Gaussian Mixture Model-Hidden Markov Model）为核心的概率统计模型[4]。然而，由于此类模型存在无法充分利用语音信号帧间上下文信息以及缺乏深度非线性特征转换能力等缺陷，性能提升受限。近年来，随着深度学习技术的飞速发展，语音识别逐渐转向以神经网络模型和端