分类号：　　密　　级：公开

学　号：20222108053　　单位代码：10759

石河子大学

硕士学位论文

基于语音识别的小学低段发音评测方法研究与系统实现

学位申请人　杜肇辉

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

2025 年 6 月

分类号：　　密　　级：公开

学　号：20222108053　　单位代码：10759

石河子大学

硕 士 学 位 论 文

基于语音识别的小学低段发音评测方法研究与系统实现

学 位 申 请 人	杜肇辉
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

2025 年 6 月

Research and System Implementation of Primary School Low Grade Pronunciation Evaluation Method Based on Speech Recognition

A Dissertation Submitted to

Shihezi University

In Partial Fulfillment of the Requirements

for the Degree of

Master of Engineering

By

Du Zhao-hui

Electronic Information

Dissertation Supervisor:Prof. YuBao-hua

June,2025

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：　2025 年 5 月 27 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：　2025 年 5 月 27 日

导师签名：　　时间：　2025 年 5 月 27 日

摘要

近年来，随着人工智能技术的发展，计算机辅助语言学习成为教育领域的重要研究方向，当前主流的发音评测模型通过将语音转换为音素序列，再与目标音素进行比对，能够有效识别误读的音素，从而显著提升发音评测的准确性。然而，低龄学生的发音习惯与技巧还不成熟，发音时存在大量音素替换、插入和省略错误，导致现有发音评测模型在面对发音不规范和快速反馈等方面时仍存在诸多挑战。因此，如何优化现有的发音评测模型，以实现高效、准确和实时的评测，成为当前该领域研究的关键问题。本文围绕普通话发音评测领域展开研究，主要工作如下：

（1）基于端到端的普通话音素识别研究。针对普通话发音的复杂性和音素识别对实时性的要求，考虑到模型需要部署在资源有限的设备上，本文提出了一种基于改进 Zipformer-RNN-T(Pruned) 架构的端到端普通话音素识别模型。该模型采用 Zipformer 作为编码器，通过减少编码器层数并设计三层 Zipformer Block 结构，在降低模型参数量的同时，仍能保持较高的识别准确率。针对无状态预测网络中可能出现的神经元失活问题，本文引入 GELU 激活函数以增强模型的非线性表达能力。此外，本文提出一种混合损失函数优化策略，通过加权融合 Pruned RNN-T 损失函数与 CTC 损失函数，进一步提升模型性能。实验结果表明，本文提出的改进模型在 AISHELL1-PHONEME 数据集上表现优异，开发集词错率达到 1.92%，测试集词错率为 2.12%，实时因子仅为 0.002，参数量控制在 61.1M，均优于现有主流模型。

（2）基于迁移学习的普通话发音评测研究。本文针对小学低段学生普通话发音评测领域数据资源稀缺的难题，构建了小学学生朗读语文课文的发音评测数据集，并采用数据扩充技术增强样本多样性。在此基础上，通过迁移学习策略将自建学生语音数据融入基于 Zipformer-RNN-T(Pruned) 的预训练音素识别模型进行训练，形成适用于小学低段学生的音素识别模型。此外，本文改进了基于音素混淆的 P-NW 音素序列比对算法，并将其与小学低段学生的音素识别模型相结合，构建了发音评测模型。实验结果表明，迁移学习使音素识别的词错率降低了 26.34%，验证了该方法的有效性。同时，P-NW 在多项评测指标上均优于 NW 算法，进一步提升了发音评测的准确性。

（3）小学低段普通话发音评测系统。针对小学低段普通话发音评测中的科学性、实时性和个性化支持不足的问题，本文设计并实现了基于教师端和学生端的普通话发音评测系统。教师端支持发布与管理朗读作业功能，并通过数据统计功能跟踪学生的学习进度，为教学调整提供依据。学生端通过语音评测与反馈帮助学生改正发音问题，同时提供易错字词收藏功能，帮助学生定向巩固发音难点。该系统能够满足普通话教学需求，提升教师的教学效率和学生的发音水平。

关键词：发音评测；语音识别；注意力机制；深度学习

Abstract

In recent years, with the development of artificial intelligence technology, computer-assisted language learning has become an important research direction in the field of education, and the current mainstream pronunciation assessment model can effectively identify mispronounced phonemes by converting speech into a sequence of phonemes and then comparing it with the target phonemes, thus significantly improving the accuracy of pronunciation assessment. However, the pronunciation habits and skills of younger students are still immature, and there are a large number of phoneme substitution, insertion, and omission errors in pronunciation, resulting in the existing pronunciation assessment models still facing many challenges in terms of pronunciation irregularities and rapid feedback. Therefore, how to optimise the existing pronunciation assessment models to achieve efficient, accurate and real-time assessment has become a key issue in the current research in this field. This thesis focuses on the field of Mandarin pronunciation evaluation, and the main work is as follows:

(1) Research on end-to-end Mandarin phoneme recognition. Aiming at the complexity of Mandarin pronunciation and the requirement of real-time phoneme recognition, and considering that the model needs to be deployed on devices with limited resources, this thesis proposes an end-to-end Mandarin phoneme recognition model based on an improved Zipformer-RNN-T(Pruned) architecture. The model adopts Zipformer as the encoder, and by reducing the number of encoder layers and designing a three-layer Zipformer Block structure, the number of model parameters is reduced while still maintaining a high recognition accuracy. Aiming at the neuron inactivation problem that may occur in stateless prediction networks, this thesis introduces the GELU activation function to enhance the nonlinear expression ability of the model. In addition, this thesis proposes a hybrid loss function optimisation strategy to further enhance the model performance by weighted fusion of Pruned RNN-T loss function and CTC loss function. The experimental results show that the improved model proposed in this thesis performs well on the AISHELL1-PHONEME dataset, with a word error rate of 1.92% in the development set, a word error rate of 2.12% in the test set, a real-time factor of only 0.002, and a parameter count control of 61.1M, which are all better than the existing mainstream models.

(2) Mandarin Pronunciation Assessment Based on Transfer Learning. To address the challenge of data scarcity in Mandarin pronunciation assessment for young students, this study constructs a pronunciation assessment model specifically designed for early primary school students. A Mandarin pronunciation assessment dataset was collected and processed based on students' oral readings of

Chinese textbooks, and data augmentation techniques were applied to enhance sample diversity. By adopting transfer learning, the student speech dataset was integrated into a pre-trained Zipformer-RNN-T (Pruned) phoneme recognition model to develop a phoneme recognition model tailored for early primary school students. Furthermore, this study proposes a P-NW phoneme sequence alignment algorithm based on phoneme confusion patterns, which is incorporated into the phoneme recognition model to construct the pronunciation assessment framework. Experimental results demonstrate that transfer learning reduces the phoneme recognition WER by 26.34%, confirming its effectiveness. Meanwhile, P-NW outperforms NW in multiple evaluation metrics, further improving pronunciation assessment accuracy.

(3) Mandarin Pronunciation Assessment System for Lower Elementary School. Aiming at the lack of scientific, real-time and personalised support in Mandarin pronunciation assessment in the lower primary school, this thesis designs and implements a Mandarin pronunciation assessment system based on the teacher's end and the student's end. The teacher's end supports the function of publishing and managing reading assignments, and tracks the students' learning progress through the statistical function to provide a basis for teaching adjustment. On the student's side, the system helps students to correct pronunciation problems through voice evaluation and feedback, and provides a collection function for easy-to-write words and phrases to help students consolidate difficult pronunciation points. The system can meet the needs of Mandarin teaching and improve the teaching efficiency of teachers and the pronunciation level of students.

Key words: Pronunciation Assessment; Speech Recognition; Attention Mechanism; Deep Learning

目录

第 1 章 绪论

1.1 研究背景及意义

语言是人类社会交流的重要工具，它不仅是人际沟通和信息传播的桥梁，也是文化传承的重要载体。我国是多民族、多语言、多方言的人口大国，普通话作为中国国家通用语言文字的标准发音，具有增强社会凝聚力，消除语言隔阂的重要作用。同时，学习和使用普通话也是传承中华优秀传统文化、提升文化自信的有效途径。在普通话推广的过程中，语言教学的质量起着至关重要的作用，直接影响到普通话推广的效果。

为了加强普通话的普及和推广，提高普通话教学质量，国家制定了多个相关政策和标准。2020 年，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”。2016 年，教育部、国家语委发布《国家语言文字事业“十三五”发展规划》，进一步提出要“加快民族地区国家通用语言文字普及”，并“在农村和民族地区开展国家通用语言文字普及攻坚”。2020 年，国务院办公厅印发《关于全面加强新时代语言文字工作的意见》，进一步明确指出“全面加强民族地区国家通用语言文字教育”，并对新时代国家通用语言文字推广普及工作作出系统部署。2021 年，教育部、国家语委部署语言文字事业“十四五”重点任务，提出“全面加强国家通用语言文字教育”。在基础教育阶段，学生的语言学习是重要的教学任务之一，普通话教学质量与效果直接影响着学生未来的语言发展。教育部也对各学段学生普通话教学做出了指示，在《义务教育语文课程标准》（2022 年版）的课程目标中对不同学段都提出了普通话教学的要求[1]。一系列重要批示及政策的发布表明国家对语言文字教学工作的高度重视，尤其是少数民族地区国语教学的战略意义巨大。

然而，对语言学习的四项基本技能“听说读写”而言，学校内进行的普通话学习受限于课时量。教师往往更关注学生在语法、写作等方面的练习，在口语朗读方面相关的教学投入时间较少，且教师评测较为主观。同时我国各地方言混杂，不少地方仍然存在用方言教学这一客观现实，使得朗读发音学习的评测及反馈难以进行。为此，研究者针对普通话发音测评进行大量的研究，制定了包括拼写、用词选择、句子语法等方面的规则，旨在客观地评价学生的普通话发音水平。而这种基于规则的评测方法在推广应用的

过程中，面临人力投入大、测评技术及方法自动化程度低等问题。

为了解决上述问题，众多学者将人工智能技术应用于计算机辅助语言学习（computer - assisted language learning, CALL）系统，以此来有效的规范、提升用户的普通话水平。计算机辅助语言学习技术是运用计算机替代或辅助教师完成语言教学任务，帮助学习者解决发音问题并给出针对性的训练计划。CALL 系统包含自动发音错误检测模块，该模块运用语音识别技术，对学习者的发音先进行识别，然后提取语料中关键特征，根据这些关键特征完成对学习者的发音分析，并定位发音错误的位置。尽管如此，普通话发音评测技术也存在着较大的挑战，如图 1-1 所示。

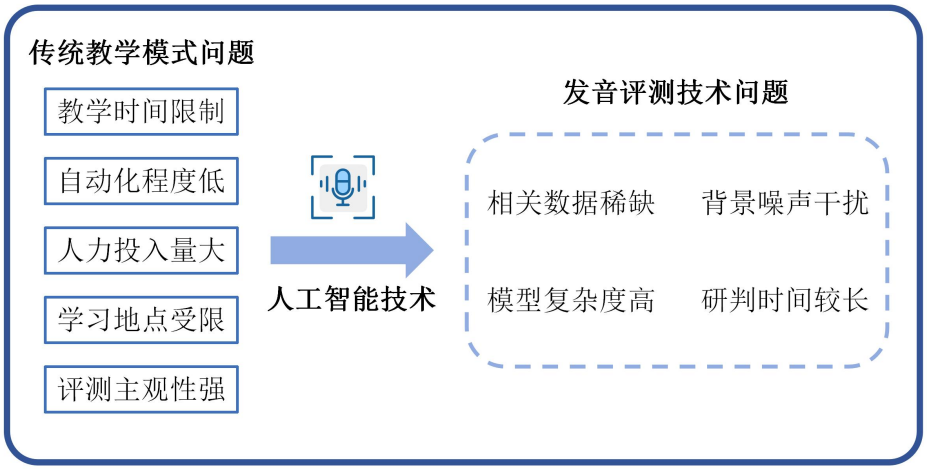

图 1-1 普通话教学面临的问题

Figure 1-1 Problems Faced in Mandarin Teaching

目前，普通话发音评测任务面临的挑战主要包括：（1）相关数据稀缺。由于录制和标注语音评测数据需要大量的人力和时间成本，导致普通话发音评测的高质量数据集较为稀缺。（2）背景噪声干扰。在学生评测环境中，背景噪声的类型多样，包括说话噪声、交通噪声、设备噪声等，对普通话发音的识别产生严重的干扰。（3）模型复杂度高。现有的发音评测模型为提高准确率，常采用复杂的网络结构和算法，这种复杂性不仅阻碍了模型的快速迭代与优化，还增加了评测系统的部署成本和维护难度。（4）研判时间较长。普通话发音的自动评测涉及复杂的声学特征分析，这种多层次的语音处理流程增加了系统的响应延迟。

传统的普通话评测往往是教师的主观评价，存在效率较低、实施成本高等问题，不利于大规模推广。普通发音习惯的养成和发音技巧的练习主要在小学低段学生完成，学生在学习汉语的过程中，难以发现自己的发音错误，导致错过最佳发音学习阶段。因此，本文利用人工智能技术研究基于语音识别技术的发音检测方法，为小学朗读教学带来更客观、有效、个性化的解决方案，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

随着人工智能以及互联网通信的快速发展，语音识别作为人类与智能设备交互的重要技术被应用在社会的各个领域。在教育领域，发音评测是语言学习中至关重要的一环，在语音识别技术的推动下正逐渐展现出巨大的潜力。理想的发音评测系统在接收学习者语音后，借助先进的语音识别技术精准识别出学习者发音的音素序列，并通过对比标准音素序列与待检测语音音素序列间的差别，从而完成评测任务。因此，本节将围绕（1）计算机辅助语言学习；（2）语音识别模型；（3）发音评测模型三个方面，系统性总结语音识别技术在普通话发音评测领域的研究现状。

1.2.1 计算机辅助语言学习研究现状

互联网的飞速发展在过去几十年间对教育体系产生了深远的影响[2]。在此背景下，计算机辅助语言学习（Computer-Assisted Language Learning, CALL）和移动辅助语言学习（Mobile-Assisted Language Learning, MALL）的应用日益广泛，并对教育质量的提升发挥了重要作用[3]。

CALL 作为一种前景广阔的教学方法，在提高英语作为外语学习者的语言能力方面具有显著效果[4]。关于 CALL 是否能够有效促进二语（Second Language, L2）学习的问题引起了学者的广泛关注，研究表明其在提升 L2 学习方面具有积极作用[5]。CALL 可用于多种教学目的，包括语言练习、教学方法创新及课堂讨论等，还能够让教师与学习者按照自身节奏进行教学与学习[6]。近年来，研究者更加关注如何通过优化认知负荷设计来提升 CALL 系统的有效性。Bahari 等人提出，从认知负荷视角审视 CALL 系统设计可显著提升 L2 学习效率，并总结了包括视频反馈、交互式词汇标注和多屏协作等十二种有效策略[7]。与此同时，随着人工智能的发展，Intelligent CALL（ICALL）正在兴起。Namaziandost 与 Rezai 的研究表明，ICALL 环境下学习者的情绪调节和正念水平对其学习动机与自主性具有显著影响，凸显了 AI 环境中情感因素对语言学习效果的重要性[8]。此外，CALL 还可以促进在线课堂互动，提供多样化的学习形式，为学生表现提供建设性反馈，鼓励合作学习，增强学习者的自我调节能力，并为其提供丰富的学习资源[9]。

在 CALL 推动 L2 学习的同时，MALL 作为技术发展的延伸，为语言学习带来了新的可能性。随着移动设备成为人们日常生活的不可或缺的一部分[10]，它们对学习方式的影响日益显著[11]。移动设备在 L2 教育中的应用十分广泛，研究表明其能够有效支持语言学习[12]。相较于禁止 L2 学习者在课堂上使用智能手机，教师更应积极探索如何将其融入教学，以帮助学生在真实环境中进行语言学习[13,14]。Xu 和 Peng 将 MALL 定义为“利用移动工具加速 L2 学习和教学”[15]。不同于传统课堂学习，MALL 使学习者能够通过

移动设备随时随地进行学习[16]，这不仅将教师和学习者从课堂环境延伸至真实世界，还能构建更加丰富的学习体验。MALL 的学习资源涵盖广泛，包括音频磁带、书籍、音频 CD、DVD 播放机以及便携式收音机等[17]。由于 MALL 突破了时间和空间的限制，它能够显著提升 L2 学习者的学习动机，使其在学习过程中承担更多责任，并增强其对学习进程的控制感[18,19]。

在移动技术推动语言学习的背景下，基于 APP 的朗读教学逐渐成为语言教学中的重要手段。随着信息化水平的提升，人工智能技术在学生朗读训练中发挥着越来越重要的作用。陈凡提出，教师应引入趣味性、引导性和实践性兼具的朗读 APP，并充分利用课堂内外的时间，引导学生利用多种朗读 APP 进行练习[20]。此外，王文宇等人建议利用社交类语音 APP 搭建朗读平台，以促进学生间的交流与互动[21]。蒋岩指出，诗歌朗读长期以来未受到充分重视，而配乐朗读在手机 APP 上的应用使其重新获得关注，标志着诗歌声音表现形式的创新[22]。白梦群尝试利用“配音秀”APP 激发粤方言区学生学习普通话的兴趣，并探索更高效的普通话学习方式[23]。王梦玉强调，基于移动终端的英语口语类 APP 不仅能提供丰富的练习材料，还具备创新的训练模式，并可随时随地使用。学生能够获得即时反馈，从而提升学习效果。对于教师而言，英语口语类 APP 的数据分析功能有助于全面了解学生的个体及整体学习情况，从而为教学决策提供科学依据，使口语教学更加精准高效[24]。

综上所述，随着信息化的普及，计算机技术逐渐融入传统课堂，同时衍生出各类自动测评系统，助力学生个性化培养。其中，借助信息化技术手段检测学生发音问题的朗读 APP 也逐渐出现在人们的视野，朗读 APP 可利用课外时间对学生进行发音训练，客观的评价学生发音水平。然而，以 APP 开发设计的英语口语居多，朗读类 APP 在语文教学应用中较少。

1.2.2 语音识别模型研究现状

语音识别是一种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。早期的语音识别技术主要以 GMM-HMM[25]为核心的概率统计模型为主，然而由于这类模型存在无法充分利用帧间背景信息和深度非线性特征转换的缺陷逐渐被淘汰。近年来，在深度学习技术的发展下，语音识别技术逐渐以神经网络模型和端到端框架为主，模型结构越发简单，目前国内外学者对此已经开展了大量研究，并取得了丰富的研究成果。

（1）基于神经网络的语音识别

传统神经网络模型通过改进网络架构与特征融合策略，显著提升了语音识别系统的鲁棒性与识别效率。在循环神经网络领域，Li 等人[26]提出基于 Grid-LSTM 的多通道语

音识别系统，结合自适应去混响前端和联合多通道处理机制，使 Google Home 系统的词错误率（Word Error Rate, WER）相对降低 8%-28%。针对频域特征建模，Passricha 等人[27]设计 CNN-BiLSTM 混合模型，利用 CNN 的局部特征提取能力与 BiLSTM 的时序建模优势，在连续语音识别任务中较纯 DNN 系统降低 10%相对 WER。Schneider 等人[28]开发的 wav2vec 模型通过卷积神经网络的无监督预训练策略，在仅少量标注数据下将 WSJ 数据集的 WER 降低 36%，证明了预训练对低资源场景的有效性。全卷积网络方面，Zhang 等人[29]采用深度全卷积网络处理振动信号谱图，在轴承故障识别任务中达到 99.22%准确率，验证了 CNN 在时序信号分类中的通用性。

（2）基于 CTC 结构的语音识别

在 CTC（Connectionist Temporal Classification）优化方面，Higuchi 等人[30]提出 Mask-CTC 非自回归框架，通过掩码预测机制修正 CTC 输出，在 WSJ 数据集上将 WER 从 17.9%降至 9.1%，推理速度达 0.07 实时因子。Nozaki 等人[31]通过中间层 CTC 损失引入条件依赖，使 WSJ 任务 WER 相对降低 20%。Higuchi 等人[32]进一步构建分层条件 CTC 模型，利用多粒度子词单元逐步优化表征，在 LibriSpeech-100h 上实现 3.0% WER。

（3）基于 RNN-T 结构的语音识别

在 RNN-T（Recurrent Neural Network Transducer）模型方面，He 等人[33]开发基于音素/字素预测的流式 RNN-T 关键词检测系统，显著超越传统 CTC 填充器基线模型。Li 等人[34]通过最小词错误率训练与 LAS 二次打分，将流式 RNN-T 的延迟降低 160ms，WER 相对减少 18.7%。Saon 等人[35]提出的乘法融合编码器-预测网络在 Switchboard 任务中达到 5.9% WER，较传统模型提升显著。Sainath 等人[36]结合多领域数据与重口音训练，实现流式 RNN-T+LAS 模型超越传统混合系统。Zhang 等人[37]对比 RNN-T、CTC 与 LF-MMI 在流式 ASR 中的效率，证实了 RNN-T 的精度优势与 CTC 的推理速度优势。

（4）基于 Transformer 的语音识别

在 Transformer 架构中，Zhang 等人[38]设计可流式解码的 Transformer Transducer，在 LibriSpeech 上实现 2.1% WER（无语言模型）。Wang 等人[39]的混合 Transformer 模型通过位置编码改进，在 LibriSpeech 上较传统混合系统提升 19%-26%相对性能。Chang 等[40]提出多说话人 Transformer 模型，结合 WPE 去混响技术，在混响场景下实现 15.2% WER。Moritz 等人[41]采用时间受限自注意力与触发注意力机制，在 LibriSpeech 流式任务中达到 2.8% WER。Huang 等人[42]开发的 Conv-Transformer Transducer 通过跨步卷积降采样，在低延迟条件下实现 3.6% WER。

（5）基于 Conformer 的语音识别

Conformer 作为卷积增强型 Transformer，Gulati 等人[43]在 LibriSpeech 上取得 1.9% WER（带外部语言模型），较纯 Transformer 模型参数效率提升 30%。Kim 等人[44]提出 SE-Conformer 用于语音增强，在 VCTK 数据集上超越传统基线模型。Zhang 等人[45]的

MFA-Conformer 通过多尺度特征聚合提升说话人验证性能，EER 低至 0.64%。Chen 等人[46]将 Conformer 用于连续语音分离，在 LibriCSS 数据集上实现最佳性能。Xiao 等人[47]提出 SLA-Conformer，通过移位线性注意力优化视听语音识别，在 LRS2/3 数据集上 WER 达 1.5%。Wu 等人[48]设计 MPSA-Conformer-CTC/Attention 模型，结合最大熵优化与稀疏注意力，在藏语识别任务中 WER 降低 10.68%。Li 等人[49]将 Conformer 引入流式 RNN-T 编码器，结合级联编码器策略实现质量-延迟帕累托前沿优化。Pan 等人[50]开发的 SRU++模型融合快速循环与注意力机制，在长语音输入任务中超越 Conformer。

（6）基于 Zipformer 的语音识别

Zipformer 作为新型高效架构，Yao 等人[51]通过 U-Net 分层编码与 ScaledAdam 优化器，在 LibriSpeech 上较 Conformer 提速 2 倍的同时 WER 降低 5%相对值。Cui 等人[52]融合 SSL 离散特征，在 GigaSpeech 任务中实现 11.14% WER。Yang 等人[53]提出 k2SSL 框架，将 Zipformer 作为 SSL 主干网络，在 Libri-Light 上 WER 相对降低 34.8%。Wang 等人[54]设计 ZipEnhancer 模型，通过双路径降采样在语音增强任务中 PESQ 达 3.69。

综上所述，语音识别技术经历了从基于 HMM 到神经网络，再到 CTC、RNN-T、Transformer、Conformer 和 Zipformer 端到端框架的演变。目前，端到端语音识别方法在解决语音序列与标签序列对齐问题、实时解码能力以及上下文建模上具备明显优势，已经成为当前研究的重点。

1.2.3 发音评测模型研究现状

随着全球化进程的加速和语言学习需求的增长，计算机辅助发音评测技术已成为语音处理领域的重要研究方向。传统评测方法主要依赖专家人工评分，效率低且主观性强，而基于人工智能的自动评测技术通过建模声学特征与发音质量间的映射关系，显著提升了评估的客观性和可扩展性。近年来，随着深度学习技术的发展，发音评测模型逐渐从基于后验概率的统计方法，演进为基于端到端的模型，评测性能取得显著突破。

（1）基于 GOP 的发音评测模型

Sudhakara 等人[55]提出在 DNN-HMM 框架中引入隐马尔可夫模型转移概率的改进 GOP 算法，在 Kaldi 工具包中实现后，使评测分数与专家评分的相关性提升 14.89%。Cheng 等人[56]构建 ASR-free 评分方法，通过语音信号边际分布建模缓解音素竞争问题，与 GOP 结合后效果优于单 GOP 基线模型。Chao 团队[57]提出 3M 模型，融合韵律特征、自监督特征及音位编码，在 Speechocean762 数据集上显著提升流畅度和韵律评估指标。Duan 等人[58]通过跨语言迁移学习构建非母语音素发音模型，在日语母语者的英语评测中同时提升识别准确率和 GOP 检测效果。

（2）基于深度神经网络的发音评测模型