双路径优化的 HiPAMA 发音评估模型

张明雨 赵学民
郑州航空工业管理学院 计算机学院，中国 · 河南 郑州 450046

摘要：发音评估在计算机辅助语言学习中具有关键作用，其动态特征感知与多粒度分析能力直接影响着评估结果的准确性。HiPAMA 模型在多粒度发音评估中有一定的优势，但存在单词重读音感知偏差及语句完整性评估维度缺失等问题。通过解构其分层注意力架构，提出了引入双向门控循环单元和多尺度动态注意力机制的双路径优化方案，通过遗忘门与更新门双向的协同作用强化发音连续性建模提升重音边界识别精度；采用可学习权重分配的多尺度动态注意力机制实现特征通道与时间维度的自适应聚焦。在 speachocean762 基准数据集上的对比实验显示，优化的模型在推理速度保持原有水平的基础上，单词重读音评估指标与语句完整性评估指标上分别实现 4.8% 和 14.5% 的显著提升。该方案为智能语音评估系统提供了更鲁棒的特征提取框架，尤其在非母语学习者的韵律纠错场景中展现出独特优势。

关键词：发音评估；门控循环单元；注意力机制

Dual-path optimized HiPAMA pronunciation evaluation model

Zhang Mingyu, Zhao Xuemin
School of Computer Science, Zhengzhou University of Aeronautics and Industry Management, China Henan Zhengzhou 450046

Abstract: Pronunciation assessment plays a pivotal role in computer-assisted language learning, with its dynamic feature perception and multi-granularity analysis capabilities directly influencing the accuracy of assessment results. The HiPAMA model exhibits certain advantages in multi-granularity pronunciation assessment, but it suffers from issues such as perceptual bias in word stress reiteration and a lack of assessment dimensions for sentence integrity. By deconstructing its hierarchical attention architecture, a dual-path optimization scheme is proposed, incorporating bidirectional gated recurrent units and a multi-scale dynamic attention mechanism. This scheme enhances the modeling of pronunciation continuity through the bidirectional collaborative action of the forget gate and update gate, thereby improving the accuracy of stress boundary recognition. A multi-scale dynamic attention mechanism with learnable weight allocation is adopted to achieve adaptive focusing on feature channels and time dimensions. Comparative experiments on the Speechocean762 benchmark dataset show that the optimized model achieves significant improvements of 4.8% and 14.5% in word stress reiteration assessment metrics and sentence integrity assessment metrics, respectively, while maintaining the original inference speed. This scheme provides a more robust feature extraction framework for intelligent speech assessment systems, particularly demonstrating unique advantages in prosody correction scenarios for non-native learners.

Keywords: Pronunciation evaluation; Gated recurrent unit; Attention mechanism

0 引言

在全球化进程不断加速的当下，语言学习已然成为人们增强自身竞争力、拓宽国际视野的关键路径。计算机辅助语言学习（CALL）系统，作为一种高效且便捷的语言学习工具，受到了广泛的关注与应用。发音评估技术[1]作为 CALL 系统的核心构成部分，借助模拟人类听觉感知与语言处理能力，能够为学习者提供即时、客观的发音反馈，在助力学习者纠正发音错误、提升发音准确性等方面发挥着举足轻重的作用。

尽管发音评估技术历经几十年发展取得显著进步，但在单词重读音与语句完整性评估方面仍面临诸多难题。传统评估方法多聚焦于音素级发音质量评分，且常采用单独建模，对涉及较长语音序列及复杂语言现象的单词重读音和语句完整性评估能力有限[2-4]。近年来，深度学习技术为

发音评估带来新机遇，HiPAMA 模型[5] 作为先进的深度学习模型，凭借其独特的分层结构和注意力机制[6,7]，在多粒度、多方面发音评估中展现出一定优势。然而，在实际语言交流里，单词重读音与语句完整性是影响发音质量的关键要素。但 HiPAMA 模型在处理这两类信息时存在明显短板，未能充分挖掘其背后复杂的内在关联，致使在单词重读音和语句完整性评估方面效果欠佳，仍有进一步改进与优化的空间。

本文旨在深入剖析 HiPAMA 模型在单词重读音与语句完整性评估中的局限性，通过引入门控循环单元的改进策略[8] 以及优化注意力机制这两项举措，在公开可用的 speechocean762 数据集[9] 上展开实验。结果显示，改进后的模型在单词重读音与语句完整性的评估任务上，表现均显著优于基线模型 HiPAMA。

1 相关工作

早期关于发音评估的研究，要么仅在音素层面评估单一维度的分数[10]，要么分别评估单词层面或语句层面的多个维度分数[11,12]。最近提出的 HiPAMA 模型（如图 1）利用分层结构和多方面注意力机制，在每个粒度级别预测多个方面。

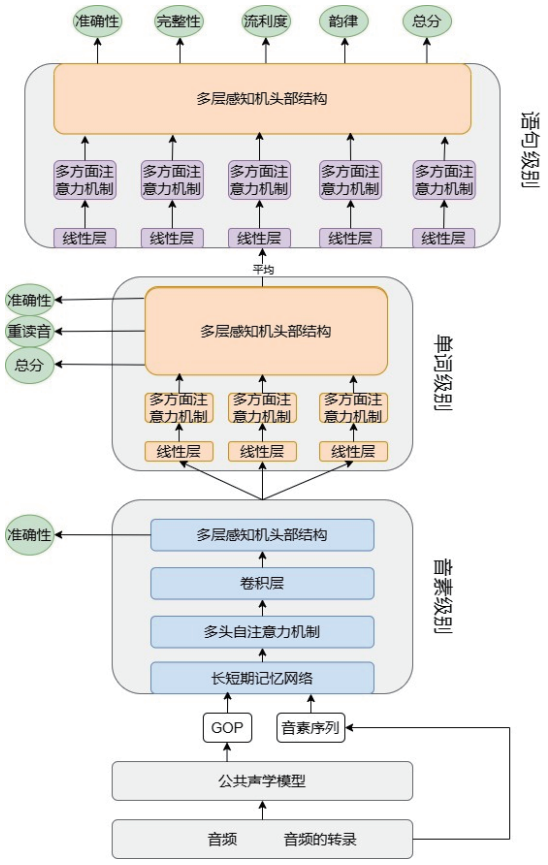

图1 HiPAMA模型架构图

尽管 HiPAMA 模型的架构从整体而言具备一定优势，然而在处理单词重读音与语句完整性信息时，仍存在一定缺陷。重读音与音素和单词之间的快速衔接以及发音的节奏感紧密相关，而 HiPAMA 模型现有的架构难以充分捕捉这些动态变化的特征；完整性与语句的语调变化以及重音分布紧密相连，HiPAMA 模型在处理完整性信息时，未能充分考量这些因素与语句结构之间的复杂关系。

2 改进措施

2.1 引入门控制循环单元

在发音评估领域，对单词重读音与语句完整性特征的分析，通常高度依赖于较长时间序列跨度内的信息。这些信息在时间维度上呈现出一定的延续性和关联性。门控循环单元（GRU）[13] 作为循环神经网络（RNN）的一种极具创新性的变体，在处理这类时序数据时展现出了独特而显著的优势。其核心优势在于，通过精妙设计的门控机制，GRU 能够有效地记忆并传递长期依赖信息，从而精准地捕捉发音过程中的连续性和节奏变化。

经过深入分析，决定在音素级别处理之后、单词级别处理之前引入门控制循环单元 (如图 2)。这样的位置选择使得门控制循环单元能够直接处理经过音素级别初步特征提取后的序列数据，从而更好地捕捉音素之间的时序关系，并将处理后的信息传递给单词级别进行下一步处理。

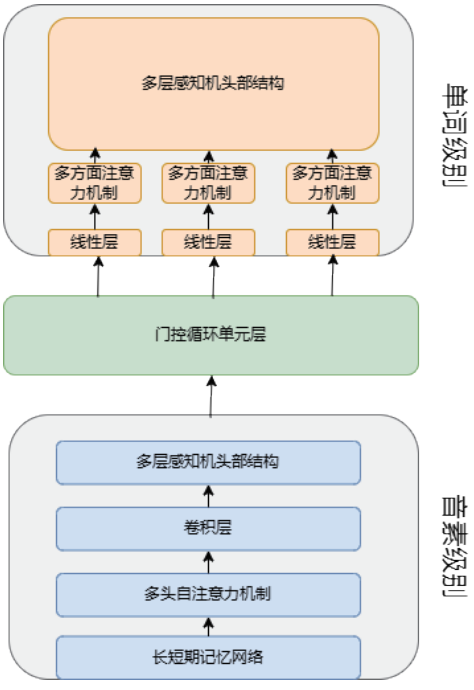

图2 引入门控制循环单元示意图

2.2 优化注意力机制

2.2.1 自适应权重初始化

在模型参数初始化中，传统固定标准差随机初始化法[14] 虽有通用性，但面对多样化输入特征时，难以洞察其内在

结构，导致其处理复杂特征较为困难。为此，本文采用自适应权重初始化方案。它依据输入特征统计，调配注意力机制初始参数。先计算输入特征 X 各维度方差，计算公式为：

$$\sigma_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2, j = 1,\ldots,d \quad (1)$$

其中，$\overline{X_j}$ 是第 j 个特征维度的均值，即$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$。

随后，根据方差初始化 att_v 和 att_w 参数，给方差大的特征维度更大权重。这就像给模型配了导航，训练初期就能聚焦关键特征，避免盲目搜索，加快收敛速度。

2.2.2 多尺度注意力权重

在模型注意力机制设计中，引入 k 个不同尺度分支，各分支采用大小 s_k(k=1,...,K) 的卷积核，对输入特征 X 进行卷积操作，固定卷积核密度为 w, 得到不同尺度特征表示 $X^{(k)}$，卷积公式为：

$$X_{ij}^{(k)} = \sum_{m=0}^{s_k-1}\sum_{n=0}^{w-1} X_{j+m,j+n} \cdot C_{mn}^{(k)}, i = 1,\ldots,n-s_k+1, j = 1 \quad (2)$$

其中 $C^{(k)}\epsilon Rs^{k*w}$ 是第 k 个分支的卷积核参数。对每个尺度分支 k, 分别计算其注意力权重 $weights^{(k)}$，通过引入可学习的融合权重 β_k(满足$\sum_{k=1}^{K}\beta_k = 1$)，采用自适应加权融合的方式，将各分支注意力权重融合为最终的注意力权重 weights,，计算公式为：

$$\text{weights} = \sum_{k=1}^{K}\beta_k \cdot weights^{(k)} \quad (3)$$

相较于单一维度的注意力权重计算，引入不同尺度的注意力权重计算方法[15], 增设并行分支 , 以不同卷积核处理输入特征 , 分别计算注意力权重后 , 最后将这些不同尺度下得到的注意力加权结果进行融合。

3 实验设置

3.1 数据集

本文选用 speechocean762 数据集，该数据集具有丰富的多粒度多方面评分标签。每个样本都有专业语言教师标注的详细评分，每个语句有准确性、流畅性、完整性、韵律、总分（0 – 10 分）五个维度评分；每个单词有准确性、重读音、总分（0 – 10 分）三个维度评分；每个音素则有准确性分数（0 – 2 分），能充分满足发音评估模型的验证需求。

3.2 损失函数

本文使用均方误差（MSE）损失作为发音评估任务的损失函数，该方法在发音评估任务中较为常用。其计算公式为：

$$L_{total} = \sum_{m=1}^{M}\frac{1}{N}\sum_{n=1}^{N} L_{mn} \quad (4)$$

其中，M 代表粒度级别总数，本文涉及音素、单词、语句三个粒度级别（M = 3）；N 表示每个粒度级别下的方面总数，如音素级别仅准确性 1 个方面（N = 1），语句级别有准确性、完整性等 5 个方面（N = 5）。先计算各级别内各个方面损失 L_{mn} 的总和，再求其平均值，最后将所有粒度级别的平均损失相加得到总损失 L_{total}。模型通过最小化总损失，调整自身参数，使不同粒度和方面的预测更接近真实评分，提高发音评估准确性。

3.3 评估指标

为全面、客观评估模型性能，选用皮尔逊相关系数（PCC）和均方误差（MSE）作为主要指标。PCC 衡量模型预测值与真实值的线性相关性，取值范围 [−1, 1]。接近 1 时，表明模型能精准评估发音；接近 − 1 时，模型评估与实际相悖；接近 0 时，模型未有效学习数据规律。MSE 衡量预测值与真实值误差平方的平均值，值越小，模型预测越准确。

为保证公平比较，除改进部分外，其他配置与 HiPAMA 模型一致，采用 Librispeech 960 小时数据训练的自动语音识别（ASR）声学模型[16]，使用 Adam 优化器训练 100 轮，初始学习率 1e−3，批量大小 25，多头注意力设为 4 个头，以 PCC 作为评估指标，音素级别部分用 MSE 作为评估指标。

4 实验结果与分析

4.1 实验结果

本文对改进模型与 HiPAMA 模型展开了全面细致的对比分析，具体数据见表 1。实验结果清晰直观地凸显出改进模型的显著优势。在各类评估任务中，改进模型展现出强大的性能，其 PCC 分数在绝大多数情况下都处于较

表1 实验结果统计表

	Phoneme Score		Word Score(PCC)			Utterance Score(PCC)				
	MSE ↓	PCC ↑	Accuracy ↑	Stress ↑	Total ↑	Accuracy ↑	Completeness ↑	Fluency ↑	Prosody ↑	Total ↑
HiPAMA	0.087	0.601	0.553	0.305	0.571	0.735	0.312	0.744	0.751	0.755
HiPAMA (改进)	0.083	0.617	0.572	0.353	0.587	0.735	0.457	0.752	0.753	0.758

高水平。在单词重读音评估方面，改进后的模型相较于原模型，提升幅度达到了 4.8%；在语句完整性评估上，改进后的模型表现尤其出色，与原模型相比，提升幅度高达 14.5%。

4.2 消融研究

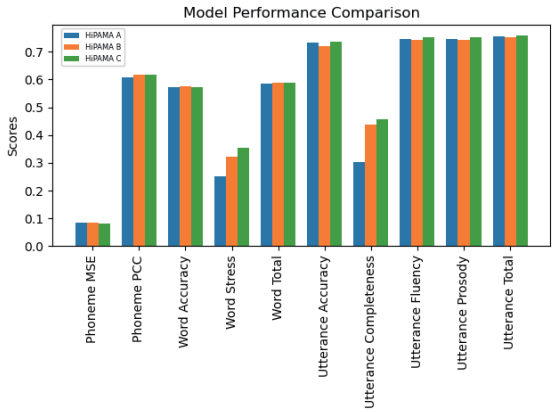

图3 消融实验结果图

为深入探究改进模型中各组件的具体作用，开展了消融实验，实验结果如图 3 所示。实验首先构建并测试了一个基础模型（HiPAMA A），该模型既未引入门控循环单元，也未采用优化注意力机制。在此基础上，对第二个模型（HiPAMA B）仅优化了注意力机制，用于评估注意力机制单独发挥的效能。最终的改进模型（HiPAMA C）则是在第二个模型的基础上，进一步引入门控循环单元。实验数据表明，优化注意力机制后，在单词重读音和语句完整性这两个以往较难评估的方面，性能提升幅度显著高于其他方面。这意味着优化后的注意力机制能够精准聚焦关键信息，有效攻克单词重读音和语句完整性评估的难题。不仅如此，引入门控循环单元后，单词级和语句级的整体评估性能也得到了提升，这表明门控循环单元在提升模型整体的语言处理能力中发挥了重要作用。

5 结语

本文针对 HiPAMA 模型在单词重读音和语句完整性评估方面的不足，提出了引入双向门控循环单元和多尺度动态注意力机制的双路径优化方案。在 speechocean762 数据集上的实验显示，改进后的模型在单词重读音评估指标提升 4.8%，语句完整性评估指标提升 14.5%，效果显著优于原模型，能够提供更为精准、可靠的评估结果，为语言学习者提供更贴合实际的发音反馈，助力其发音水平的高效提升。

参考文献：

[1] Lin B, Wang L, Feng X, et al. Automatic Scoring at Multi-Granularity for L2 Pronunciation[C]//Interspeech. 2020: 3022-3026.

[2] Shi J, Huo N, Jin Q. Context-aware goodness of pronunciation for computer-assisted pronunciation training[J]. arXiv preprint arXiv:2008.08647, 2020.

[3] 赵倾国. 基于 DIVA 模型的英语辅音发音错误自动校正方法[J]. 信息技术，2023, (12): 162-166+171.

[4] Leung W K, Liu X, Meng H. CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 8132-8136.

[5] Do H, Kim Y, Lee G G. Hierarchical pronunciation assessment with multi-aspect attention[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

作者简介：张明雨（1998.03-），男，汉，河南周口人，硕士在读，研究方向：语音处理。