doi: 10.7690/bgzdh.2024.08.005

基于多模态注意力融合网络的英语口语自动评分

梁　珊
（陕西交通职业技术学院教务处，西安 710000）

摘要：针对传统英文口语评分效率及准确率低的问题，提出一种多模态注意力融合网络架构，加快模型训练效率及口语评分准确率。综合考虑口语发音的韵律声音特征及所答问题文本信息，从而提高网络鲁棒性。通过仿真，将该模型与 LSTM、BiLSTM、GRU 网络模型进行比较，所提出模型分数估计准确率为 96.8%，明显高于其他方法。仿真结果表明：所提方法能够大幅减少评分时间，提高评分效率。

关键词：英语口语；自动评分；深度学习；注意机制
中图分类号：TP393　　**文献标志码：**A

Automatic Rating of Spoken English Based on Multimodal Attention Fusion Network

Liang Shan
(*Academic Affairs Office*, *Shaanxi College of Communication Technology*, *Xi'an* 710000, *China*)

Abstract: Aiming at the low efficiency and accuracy of traditional oral English scoring, a multimodal attention fusion network architecture is proposed to speed up the training efficiency of the model and the accuracy of oral English scoring. The network robustness is improved by comprehensively considering the prosodic sound characteristics of the spoken language pronunciation and the text information of the answered question. Through simulation, the proposed model is compared with LSTM, BiLSTM and GRU network model, and the score estimation accuracy of the proposed model is 96. 8%, which is significantly higher than other methods. The simulation results show that the proposed method can significantly reduce the scoring time and improve the scoring efficiency.

Keywords: spoken English; automatic scoring; deep learning; attention mechanism

0　引言

在计算机辅助语言学习、自动评估和重音检测等领域，对非母语者的韵律特征进行描述越来越受到人们的关注[1-2]。韵律的一个重要组成部分是节奏，通常定义为一个人讲话中的音素、音节和字长的模式。不同的语言有着不同的自然节奏，捕捉自然节奏的能力是衡量非母语者英语水平的关键因素。

传统英语韵律特征提取主要依靠人工手段，易受主观干扰且提取效率有限[3]。随着人工智能技术不断成熟，许多学者提出利用深度网络对英语口语韵律特征进行提取，从而高效、准确地提取特征并对英语口语进行评估。文献[4]借助于快速发展起来的计算机语音合成和模仿技术，提出了建立在语音合成和模仿上的口语评估路径。文献[5]重点探讨起源于心理学的 NLP 理论在英语口语教学中的应用。文献[6]提出了一种基于卷积神经网络的英语口语流利性评分方法，从原始的时域信号输入中联合学习特征提取和评分模型。上述大部分方法进行特征提取时依赖等时性或语音特征。然而这类特征描述有限，推广性较差，应用于深度学习时无法完全发挥深度网络的优势。此外，口语评分时除韵律特征还应考虑口语回答是否得体、准确。

笔者将英语口语评分过程视为一个多模态特征提取过程，综合考虑韵律声音特征及文本信息特征。进一步，提出了一个多模态注意力融合网络架构，从而加速网络学习过程。

1　韵律特征

1.1　等时性特征

传统上，语言的自然节奏认为是由一个称为等时性的原则所控制的。在英语语言中，该节奏称为重音计时，且相邻单词的重音音节之间的时间保持不变；因此，英语中单个音节的时长变化很大，这取决于它们相对于当前和相邻单词重音的位置。基于该理论，非母语语音听起来很奇怪的一部分原因是没有与英语的重音节奏相匹配；因此，母语重音

收稿日期：2024-04-23；**修回日期：**2024-05-25
第一作者：梁　珊(1987—)，女，陕西人，硕士。

与英语重音间隔的标准差可用来度量英语口语水平。然而需注意，并不是所有英语都是重音计时的，且重音计时的变体有不同程度的差别。此外，由于未能对多种语言进行分类，等时性范式可移植性能力较差。

1.2　语音特征

英语中语音可分为元音和辅音间隔 2 部分[7]。因此，语音特征可统计如下：

1) %V：不考虑词的界限，句子中用于元音间隔的时间比例；

2) ΔV：元音间隔时间的标准差；

3) ΔC：辅音间隔时间的标准差。

考虑到英语中元音通常根据其在单词中的位置而缩短，因此ΔV和ΔC非常高，而%V非常低(实际上它们分别是所有测试语言中最高和最低的)。然而母语不是英语的发音者，很可能无法正确地缩短元音；因此，有学者将这一概念推广到基于成对变异指数(pairwise variability index，PVI)，该指数用来衡量连续测量之间的变异性。PVI 适用于元音的持续时间以及元音间隔。原始变坡点 rPVI 定义如下：

$$r\mathrm{PVI}=\frac{1}{m}\sum_{k=1}^{m-1}\left|d_k-d_{k+1}\right|。\tag{1}$$

式中：d_k为第 k 段的持续时间；m 为分段的数量。rPVI 的提取如图 1 所示。

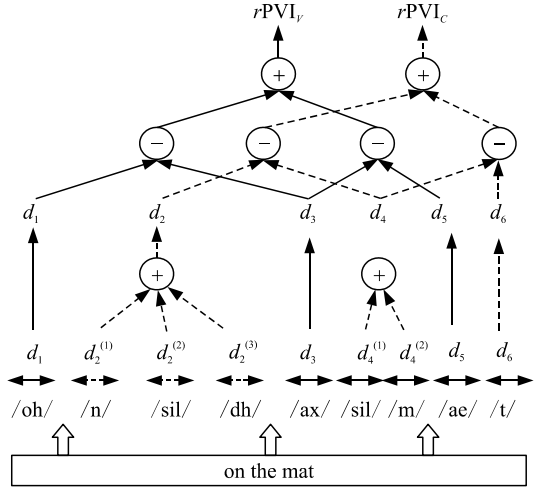

图 1　PVI 特征提取过程

进一步，PVI 的标准化 nPVI 定义如下：

$$n\mathrm{PVI}=\frac{1}{m-1}\sum_{k=1}^{m-1}\frac{\left|d_k-d_{k+1}\right|}{\left(d_k+d_{k+1}\right)/2}。\tag{2}$$

当考虑单个元音和辅音的长度，而不是元音和辅音片段的长度时，元音和辅音片段的变化是语言韵律特性的关键；因此，将每一段的持续时间除以其中包含元音数量，可计算控制补偿指数(control compensation index，CCI)：

$$\mathrm{CCI}=\frac{1}{m-1}\sum_{k=1}^{m-1}\left|d_k/n_k-d_{k+1}/n_{k+1}\right|。\tag{3}$$

式中：d_k和 n_k为第 k 次测量的持续时间和分段数；m 为测量次数。

根据上述分析中，英语语言可定义为“补偿”语言，因为相邻元音和辅音的大小不同以相互补偿，导致英语具有较高的 CCIs。然而，PVI 和 CCI 一样，仍然不可能捕捉到超出片段对水平的持续时间关系，并且求和迫使所有片段被赋予相等的权重，而实际上有些片段可能比其他片段更突出地表征节奏。

1.3　深度语音特征

随着深度学习的发展，有学者使用循环神经层和注意机制来创建一个可调的、端到端可训练的、深度学习的替代方案，以替代手工制作的特征发音特征。递归层允许捕获随时间变化的模式，而注意机制允许在将序列压缩为低维固定长度表示时权衡不同时间步的相对显著性。图 2 为笔者创建的一个深层韵律特征，该特征可替代等时性及 PVI、CCI 特征。

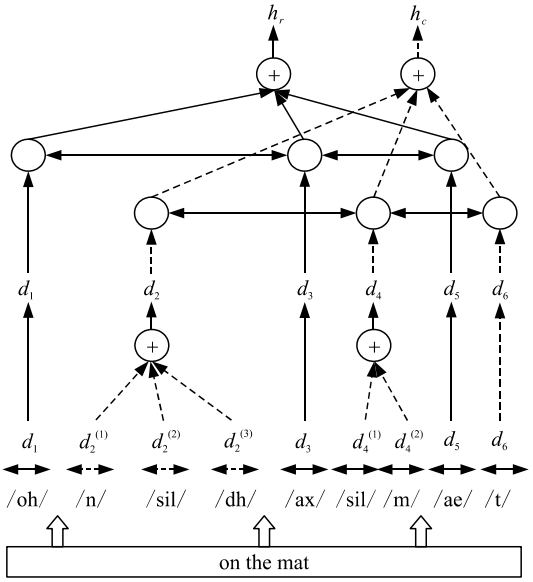

图 2　深度语音特征提取过程

考虑一个给定的英语口语音频涉及的话语，并将其分为 V 元音段和 C 元音间段(例如短语“on the mat”由元音段/oh/、/ax/和/ae/以及元音间段{/n/、/sil/、/dh/}、/m/和/t/)。

对于这 2 种类型的段，笔者定义了向量$\boldsymbol{d}_k^{(n)}$，

其中包含给定段 n 的第 k 个子段的持续时间、发音标识和其他显著信息。在 PVI 中，每个片段的子片段持续时间通过加法合并，而在 CCI 中由平均值计算。

每个子段上的自注意机制[8]用于捕获每个子段的相对显著性，并且结果与段 $\boldsymbol{d}_k^{(n)}$ 的总持续时间串联，以产生笔者所定义的关于段 n 的向量 $\boldsymbol{d}_n$：

$$\boldsymbol{d}_n=\left[\sum_{k=1}^{k(n)}\alpha_k d_k^{(n)},d_n\right]; \tag{4}$$

$$\alpha_k=\frac{\exp s(d_k^{(n)},\theta_{\text{att}})}{\sum_{j=1}^{k(n)}\exp s(d_j^{(n)},\theta_{\text{att}})}。 \tag{5}$$

式中 θ 为参数。

进一步，每个口语音频中的所有元音和所有元音间片段中的每一个向量 $\boldsymbol{d}_n$ 序列通过深度学习网络来捕获整个持续时间序列的依赖性，而不仅仅是相邻持续时间对，则有：

$$h_{1:V}^{(V)}=F(d_{1:V}^{(V)},\theta_v); \tag{6}$$

$$h_{1:C}^{(C)}=F(d_{1:C}^{(C)},\theta_c)。 \tag{7}$$

式中：$h_{1:V}^{(V)}$ 和 $h_{1:C}^{(C)}$ 为英语发音不同片段特征；F 为深度学习网络模型；θ 为网络相关参数。

注意机制将每个结果序列映射到固定长度的元音和元音间特征，以捕捉每个片段相对于整体节奏特征的相对显著性，则有：

$$\tilde{h}^{(V)}=\sum_{v=1}^{V}\alpha_v h_v^{(V)}; \tag{8}$$

$$\alpha_v=\left(\exp s(h_v^{(V)},\theta_{\text{att}})\right)\Big/\left(\sum_{j=1}^{V}\exp s(h_j^{(V)},\theta_{\text{att}})\right); \tag{9}$$

$$\tilde{h}^{(C)}=\sum_{c=1}^{C}\alpha_c h_c^{(C)}; \tag{10}$$

$$\alpha_c=\left(\exp s(h_c^{(C)},\theta_{\text{att}})\right)\Big/\left(\sum_{j=1}^{C}\exp s(h_j^{(C)},\theta_{\text{att}})\right)。 \tag{11}$$

式中：$\tilde{h}=[\tilde{h}^{(V)},\tilde{h}^{(C)}]$ 为英语发音深度特征；θ 为参数。

2 基于深度学习的评分模型

2.1 双向循环 CNN

口语音频评分标准包括语音质量、发音、流利度、重音和语调等原始特征及深度特征。为捕获音频样本的声学信息，笔者提出一种递归卷积神经网络(RCNN)[9]结构。CNN 能够提取局部特征，而 RNN 能够总结长时间的信息。

笔者将音频样本的对数标度谱图沿时间维度拆分为固定大小的帧。每个帧传递给 5 组 CNN，经最大池化处理。进一步，将从该网络获得的输出向量送入 BiLSTM[10]网络，从而捕获音频的序列结构。

2.2 双向长短时记忆

基于口语回答的内容与主题相关且适当的要求，笔者使用双向长短时记忆对回答内容评分。

首先，将口语回答文本进行预处理，单词映射到网络嵌入层。该嵌入层使用 Wikipedia 上训练的 300-D 初始化，并在训练期间进行优化。与声学模型类似，笔者在处理文本时使用 BiLSTM 网络来捕捉单词的顺序结构，并学习不同分数水平下内容。

LSTM 单元中，状态单元通过遗忘门和输入门来控制。遗忘门决定了在前一时刻 h_{n-1} 的电池状态可以保持到当前时刻 h_n 的程度。对于电池 RUL 预测，遗忘门控制前一时刻的电池容量，以生成下一时刻电池容量预测的输入参数。其中，遗忘门 f_n 由下式给出：

$$\boldsymbol{f}_n=\sigma(\boldsymbol{W}_f\bullet[\boldsymbol{h}_{n-1},\boldsymbol{x}_n]+\boldsymbol{b}_f)。 \tag{12}$$

式中：$\boldsymbol{W}_f$ 和 $\boldsymbol{b}_f$ 分别为遗忘门的权重矩阵和偏置。$[h_{n-1},\ x_n]$表示输入矢量，该矢量由电池容量序列和循环数据组成。$\boldsymbol{W}_f$ 的维数与$[h_{n-1},\ x_n]$有关，可以重写为：

$$\left[\boldsymbol{W}_f\right]\begin{bmatrix}h_{n-1}\\x_n\end{bmatrix}=\begin{bmatrix}W_{fh}&W_{fx}\end{bmatrix}\begin{bmatrix}h_{n-1}\\x_n\end{bmatrix}。 \tag{13}$$

类似地，输入门以 S 形函数的形式控制状态单元，可描述为：

$$\boldsymbol{i}_n=\sigma(\boldsymbol{W}_i\bullet[\boldsymbol{h}_{n-1},\boldsymbol{x}_n]+\boldsymbol{b}_i)。 \tag{14}$$

式中：$\boldsymbol{i}_n$ 为输入门向量；$\boldsymbol{W}_i$ 和 $\boldsymbol{b}_i$ 分别为输入门的权重矩阵和偏置矩阵。

输入门使用 tanh 函数来计算当前输入状态单元，且由遗忘门、输入门以及电流输入序列共同确定：

$$\tilde{\boldsymbol{S}}_n=\tanh(\boldsymbol{W}_c\bullet[\boldsymbol{h}_{n-1},\boldsymbol{x}_n]+\boldsymbol{b}_c); \tag{15}$$

$$\boldsymbol{S}_n=\boldsymbol{f}_n\boldsymbol{S}_{n-1}+\boldsymbol{i}_n\tilde{\boldsymbol{S}}_n。 \tag{16}$$

式中：$\boldsymbol{S}_n$ 为状态门向量；$\tilde{\boldsymbol{S}}_n$ 为当前状态单元；$\boldsymbol{S}_{n-1}$ 为长期存储器状态单元；$\boldsymbol{W}_c$ 和 $\boldsymbol{b}_c$ 为状态门权重矩阵和偏置。LSTM 中，输入门可以防止无关信息影响记忆过程，并且可以通过遗忘门保留长期信息。同时，长期信息会影响通过输出门控制的当前输出。LSTM 的最终输出由输出门和更新状态单元确定如下：

$$\boldsymbol{O}_n = \sigma\left(\boldsymbol{W}_o\left[\boldsymbol{h}_{n-1}, \boldsymbol{x}_n\right]\right) + \boldsymbol{b}_o ; \tag{17}$$

$$\boldsymbol{h}_n = \boldsymbol{O}_n \tanh\left(\boldsymbol{S}_n\right) 。 \tag{18}$$

式中：$\boldsymbol{O}_n$ 为输出门向量；$\boldsymbol{W}_o$ 和 $\boldsymbol{b}_o$ 为状态门权重矩阵和偏置。

2.3　多模态注意力融合网络

注意机制能够加权在每个时间步学习的上下文信息，允许模型确定哪些状态需要注意。笔者使用注意力融合结合文本和音频模态的特征来分级口语评分，具体模型如图 3 所示。

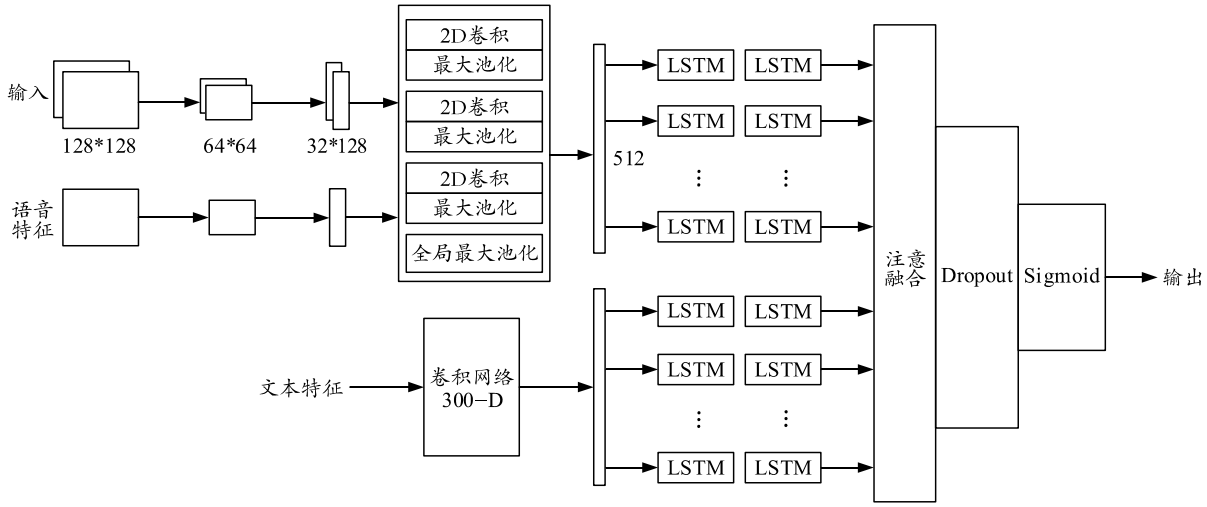

图 3　多模态注意力融合网络架构

令来自音频模型和文本模型的双向时间隐藏状态为 h_t 和 h_a，则口语评分多模特状态 h^m 定义为：

$$h^m = [h^a, h^t] 。 \tag{19}$$

式中 $[\bullet]$ 为状态融合。

进一步，上下文信息 c^m 定义如下：

$$e_t = h_t^m \boldsymbol{w}_a ; \tag{20}$$

$$a_t^m = \frac{\exp(e_t)}{\sum_{i=1}^{T} \exp(e_i)} ; \tag{21}$$

$$c^m = \sum_{t=1}^{T} a_t^m h_t^m 。 \tag{22}$$

式中：h_t^m 为时间 t 时词汇/声音的多模态表示；$\boldsymbol{w}_a$ 为注意层的权重矩阵；e_t 为多模态注意重要性得分；此外，笔者使用多模态注意重要性得分作为权重，计算线索的上下文信息 c^m，并作为所有时间步的加权总和。

3　仿真与分析

3.1　数据集与仿真环境

仿真所用数据集为某高校口语水平面试所收集的数据。每位考生都会在电脑屏幕上看到一张由 6 个不同难度的题目组成的表格，并记录他们的回答。然后由 2 个专家评分员对这些回答进行独立评分。第三方专家解决分歧(如果有的话)。要回答这些问题，考生需要具备解释和论证能力。根据他们在这些任务中的表现，他们的英语口语水平从最高等级 A2 到最低等级 C1 不等(获得的分数与题目的难度相关)。考生多为 17～22 岁在校大学生，约 8 000 人，大多数录制的音频的大小都小于 120 s。表 1 所示为数据集部分相关统计信息。双重评分与干预的三分制和评估准则，确保评分过程中不偏向无关的细节，如性别、年龄等。作为参考，所有提示的得分和文本长度之间的相关性为 0.35。

表 1　数据集部分相关统计信息

题目序号	难度	平均时长/s	分数统计		
			A1	低于 B1	高于 B1
1	B1	57.7	279	1 575	6 154
2	B1	58.7	546	3 122	4 492
3	B2	81.5	121	672	3 535
4	C1	104.2	123	729	3 573
5	C1	106.1	112	557	3 061
6	B1	55.9	121	1 045	6 992

仿真时环境设置如下：编译环境为 python+pytorch；硬件环境为联想服务器，显卡 NVIDIA rtx2080，16 GB 内存。

3.2　训练过程

将数据集按照 70:10:20 的比例分成训练、验证和测试集。为训练模型，首先，将音频响应降采样到 16 kHz，并生成它们的对数标度 Mel 频谱图。然后对这些频谱图进行归一化和零填充处理，以获得最大响应长度。最后，将输出沿时间维度拆分为大

小为 128×128 的帧。

笔者将回答的得分视为一个回归问题。与响应相关的分数(*N* 级)映射到范围[0, *N*−1]，将这些分数标准化到[0, 1]的范围内进行训练。在测试过程中，将模型输出重新缩放到原始的分数范围，并对性能进行度量。表 2 所示为仿真训练时超参数。

表 2 仿真超参数

参数	学习率	学习率衰减周期	学习率衰减倍数	批大小	最大迭代次数	Dropout 率
值	10^{-4}	20	0.1	16	150	0.3

3.3 性能测试

3.3.1 不同模型性能

进一步，将笔者所提多模态注意力融合网络模型与 LSTM、BiLSTM、GRU 网络模型进行比较，测试准确率如图 4 所示。可以看出，所提出的多模态注意力融合网络模型分数估计准确率可达到 96.8%，优于 LSTM 的 95.4%、GRU 的 95.6%和 BiLSTM 的 95.8%。仿真结果进一步验证了所提方法的有效性。

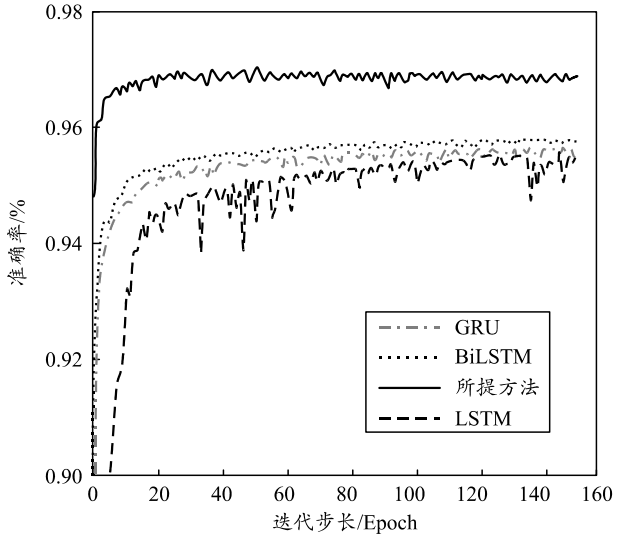

图 4 不同模型测试准确率对比结果

3.3.2 效率对比

在数据集中随机抽取 30 组口语测试语音，分别由笔者所提模型和人工口语测评进行对比。为减少由于偶然因素造成的随机误差，取 30 组平均值作为最终测试结果。表 3 为不同方法对比统计情况。可以看出，所提模型平均耗时为 0.1 s，几乎在学生答完题即刻给出评分，且准确率为 96.7%。人工评分时，受主观因素影响，多位教师评分偶尔出现争议，此时需要进行协商，这将影响评分速度。仿真结果进一步验证了所提方法较人工测评相比，能够大幅减少评分时间，提高测试效率。

表 3 不同方法对比统计情况

方法	平均耗时/s	准确率/%
所示方法	0.1	96.7
人工测评	5.6	83.3

4 结论

笔者对英语口语中自动评分进行了研究与分析。首先，对英语口语韵律特征进行了研究，提出了适用于深度学习的深度声音特征；进一步，通过对声音、文本多模态进行分析，提出了基于多模态注意力融合网络的英语口语评分模型，该模型能够解放人力资源，大幅减少评分时间，提高测试效率。该模型为非英语专业大学生口语自动评分研究奠定了一定基础。

参考文献：

[1] 姚炎清. 国外韵律特征研究现状[J]. 英语广场, 2021(3): 25−27.

[2] 陈红. 英汉韵律特征对比与英语语音教学[J]. 高教学刊, 2019(26): 112−114.

[3] 薛锦, 聂亚如, 李斑斑. 英语韵律特征和初始习得年龄的关系研究[J]. 外语学刊, 2019(6): 79−86.

[4] 刘祖斌. 人工智能语音识别英语口语评估路径探讨[J]. 信息记录材料, 2019, 20(11): 92−95.

[5] 郭红梅, 梁媛元. 基于 NLP 的英语口语教学模式探究[J]. 教育现代化, 2019, 6(72): 56−59.

[6] 吴丹, 梁琪瑛, 王民意. 基于卷积神经网络的英语口语打分[J]. 信息技术, 2020, 44(11): 34−38, 44.

[7] 黄秋文, 袁志明, 田润芬. 基于语料库的本科生英语元音习得实证研究[J]. 江西理工大学学报, 2020, 41(6): 77−82.

[8] 张林明. “注意”机制在大学英语词汇附带习得中的作用研究[J]. 福建商学院学报, 2019(6): 90−95.

[9] 杨德志, 柯显信, 余其超, 等. 基于 RCNN 的问题相似度计算方法[J]. 计算机工程与科学, 2021, 43(6): 1076−1080.

[10] 李文亮, 杨秋翔, 秦权. 多特征混合模型文本情感分析方法[J]. 计算机工程与应用, 2021, 57(19): 1−12.