DOI:10.19651/j.cnki.emt.2417531

构音障碍语音识别的解码策略研究*

樊紫岩[1] 朱耀东[2] 赵伟岚[3] 李轩逸[3]
(1.浙江理工大学计算机科学与技术学院 杭州 310018;2.嘉兴大学机械工程学院 嘉兴 314001;
3.浙江理工大学信息科学与工程学院 杭州 310018)

摘 要:构音障碍语音是神经病变导致患者发音器官产生运动障碍因而发音和韵律异常的语音,给传统自动语音识别系统带来了巨大挑战。为此,本研究提出了一种结合多层表征融合解码策略和热词增重技术的创新算法。在基于 Transformer 架构的编码器-解码器模型上,改进传统的单视角解码方式为多层次表征融合,通过 3 种表征融合策略有效增强了模型对语境和复杂句子的理解能力。同时,为进一步提升构音障碍语音的识别精度,本文将热词增重技术集成到波束搜索解码过程中,针对关键词赋予更高权重。实验在 TORGO 和 UASpeech 数据集上进行,结果表明与其他基准模型相比,该方法显著降低了 WER。其中与 Whisper 基准模型比较,在 UASpeech 数据集上 WER 从 38.31%降低至 27.18%,在 TORGO 数据集上 WER 从 16.38%降低至 12.67%,证明了该方法在提升构音障碍语音识别精度方面的有效性。

关键词:表征融合;构音障碍;语音识别;热词增重
中图分类号:TN912.34;R741 **文献标识码**:A **国家标准学科分类代码**:510.4040

Decoding strategies for dysarthric speech recognition

Fan Ziyan[1] Zhu Yaodong[2] Zhao Weilan[3] Li Xuanyi[1]
(1. School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China;
2. College of Mechanical Engineering, Jiaxing University, Jiaxing 314001, China;
3. College of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Dysarthric speech arises from neurological disorders that cause motor impairments in the articulatory organs, resulting in abnormal pronunciation and prosody, which pose significant challenges to traditional ASR systems. To address these issues, this paper proposes an innovative algorithm that combines a multi-level representation fusion decoding strategy with hotword boosting technology. Built upon the Transformer-based encoder-decoder architecture, the approach improves the conventional single-view decoding method by introducing multi-level representation fusion. This is achieved through three distinct fusion strategies, which effectively enhance the model's ability to comprehend complex sentences and contextual information. Additionally, to further improve the recognition accuracy of dysarthric speech, hotword boosting is integrated into the beam search decoding process to assign higher weights to key terms. The results demonstrate that the proposed method significantly reduces the WER compared to other baseline models. Specifically, compared to the Whisper baseline model, the WER on the UASpeech dataset decreased from 38.31% to 27.18%, and on the TORGO dataset, it decreased from 16.38% to 12.67%. This highlights the effectiveness of the proposed method in improving the accuracy of dysarthric speech recognition.

Keywords: feature fusion; transformer model; speech recognition; hotword boosting

0 引 言

构音障碍是由神经系统疾病(如帕金森病等)或创伤性损伤(如中风或头部受伤)导致的言语障碍,严重影响患者日常生活。构音障碍语音由于发音器官的精细控制能力受损,通常表现为韵律异常和发音模糊,与正常语音相比存在显著的特征差异。这使得传统自动语音识别(automatic speech recognition, ASR)系统在处理构音障碍语音时表现不佳[1]。同时,由于构音障碍语音数据的采集和标注困难,其数据规模远小于正常语音数据集,进一步加剧了 ASR 模

收稿日期:2024-12-03
* 基金项目:浙江省重点研发计划项目(2017C01043)、浙江省医学电子与数字健康重点实验室项目(MEDH202206)资助

型的适用性挑战。

早期的语音识别技术多基于隐马尔可夫模型(hidden Markov model,HMM),随后深度神经网络(deep neural network, DNN)引入构音障碍语音识别领域,例如,Xiong等[2]提出使用DNN-HMM在一定程度上改善了构音障碍语音的识别效果,张楠等[3]使用三路病理语音识别模型针对语音特征提取进行了创新。然而这些方法对数据量的依赖性极高,在构音障碍语音数据稀缺的情况下模型性能受限。为解决数据稀缺问题,张小恒等[4]提出使用多个源数据集进行病理语音数据的迁移学习来解决小样本问题,钱金阳[5]使用生成对抗网络进行语音增强,Wang等[6]基于扩散模型和文本转语音增加构音障碍语音数据。尽管这些方法在缓解数据稀缺方面发挥了作用,但存在计算成本高、合成数据多样性不足以及对单一患者过拟合的风险。

在构音障碍语音识别的研究中,构音障碍语音不同与正常语音的复杂性又是另一关键问题。Almadhor等[7]和Shahamiri[8]等将语音特征视觉化避免了音素标注和对音素的依赖。然而这些方法在极端语音样本下仍然受到视觉特征提取能力的限制,并且其依赖的视觉特征提取方式在一些极端语音样本下可能不够有效,语音视觉特征适应性仍有不足。随着自监督学习模型引入构音障碍领域,Hu等[9]提出融合自适应wav2vec2.0模型,Yu等[10]提出结合构音障碍语音音频和视觉信息的AV-HuBERT框架等。这些模型在特征生成方面表现出巨大潜力,但复杂语境下的模型鲁棒性和解码器性能仍是当前研究的难点。

针对数据多样性不足和复杂语音下模型表现不佳等问题,本文提出了基于多层表征融合的解码器改进方法,通过引入多层表征融合,增强解码器对编码器输出的理解能力[11],提高模型对构音障碍语音的适应性。具体而言,本文提出的多层分视角融合在原有解码器接受传统单视角解码的同时,额外增加了3种不同的多层次视角,使解码器能够从不同层次特征中获取更多语音信息,以应对复杂语境和不规则语音样本。此外,本文提出一种轻量不依赖于大数据的方法,通过构建基于构音障碍语音的热词库,在解码器波束搜索过程中使用热词增重(hotword boosting,HB),针对概率分布进行权重偏移,进一步提高构音障碍语音识别系统对关键词汇的识别能力。

1 提出的工作

该部分首先在1.1节回顾了编码器-解码器结构的Transformer模型架构和传统的单视图解码[12],随后在1.2节提出了3种对此改进的多层表征融合注意力,最后在1.3节介绍了HB在波束搜索中的实现。

1.1 基于Transformer的编码器-解码器架构概述

以Transformer模型为基础的编码器和解码器可以描述为,假设有一个由 T 个特征向量组成的输入语音信号 $\boldsymbol{X}_{1:T}=\{x_1,x_2,\cdots,x_T\}$,和一个由 N 个目标符号组成的目标转录 $\boldsymbol{y}_{1:N}=\{y_1,y_2,\cdots,y_N\}$,记初始语音信号 $\boldsymbol{X}_{1:T}$ 为源序列 $\boldsymbol{H}_0$。

编码器使用 N 个堆叠的编码器层对源序列进行编码,通过重复相同的过程 L 次,编码器输出源序列的表示 $\boldsymbol{H}=\{\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_L\}$,其中每层的输出序列为:

$$\boldsymbol{H}_i=f_{encoder}(\boldsymbol{H}_{i-1}),\quad i=1,2,\cdots,L \tag{1}$$

编码器层由多头注意力机制(multi-head attention,MHA)和前馈神经网络(feedforward neural network, FFN)组成,解码器使用 L 个堆叠的解码器层生成目标序列 $\boldsymbol{T}_N$。通常只有来自编码器最后一层的表示 $\boldsymbol{H}_L$ 通过注意力机制作为源与目标之间的桥梁。解码器的每一层定义为:

$$\boldsymbol{T}_i=f_{\text{decoder}}(\boldsymbol{T}_{i-1},\boldsymbol{H})_L,\quad i=1,2,\cdots,L \tag{2}$$

与编码器不同,解码器有一个处理来自编码器的表示的额外步骤。对于Transformer,解码器层通过两个MHA和一个FFN实现:

$$\begin{aligned}\hat{\boldsymbol{x}}&=\text{LN}(\boldsymbol{x}+\text{MHA}_{\text{self}}(\boldsymbol{x},\boldsymbol{x},\boldsymbol{x}))\\ \tilde{\boldsymbol{x}}&=\text{LN}(\hat{\boldsymbol{x}}+\text{MHA}_{\text{enc-dec}}(\hat{\boldsymbol{x}},\boldsymbol{y},\boldsymbol{y}))\\ f_{\text{decoder}}(\boldsymbol{x},\boldsymbol{y})&=\text{LN}(\tilde{\boldsymbol{x}}+\text{FFN}(\tilde{\boldsymbol{x}}))\end{aligned} \tag{3}$$

式中:$\boldsymbol{x}$ 是解码器上一层的输出,初始输入为目标序列的嵌入,$\boldsymbol{y}=\boldsymbol{H}_L$ 是编码器最后一层的输出,LN 代表层归一化。

解码时为了保证解码器只能依赖于之前的输出防止看到未来的信息,在计算自注意力时使用了掩码操作。解码器的FFN也是由两个线性变换和一个非线性激活函数组成。

在实验中,本文使用了Whisper大语言模型作为基准模型,Whisper模型是标准的基于Transformer的编码器-解码器架构[13]。它是一个开源的预训练“弱监督模型”,在从互联网上抓取的68万h的有噪声语音识别数据上进行了预训练,其中包括17万h的其他语言和12.5万h的英语翻译,多元化训练集使其具有高健壮性,并且对构音障碍语音的识别更有包容性。

为了训练Whisper模型,本研究使用标准交叉熵损失进行训练,其中训练模型通过最大化目标类标签的估计概率来预测实例类:

$$L_{CE}=-\sum_{i=1}^{n}\log P(\boldsymbol{y}_i\mid y<i,\boldsymbol{H}_{1:M}) \tag{4}$$

1.2 多层表征融合注意力

每一层编码器都会对输入进行不同程度的抽象和特征提取,产生对应的特征表示。通过使用不同层的表示,可以获取从低级别细节到高级别抽象的多层次信息[14]。对于时序输入,多层注意力机制可以更好地探索时间依赖性,充分利用时序信息中的局部特征和全局特征[15]。为了在序列到序列学习中使解码器充分利用来自全局和局部视角的源序列信息,本文提出通过组合具有不同抽象粗粒度的源序列的多视图来增强解码过程,鼓励解码器充分且有效地利用来自全局和局部视角的源序列信息,使模型的表达能

力能够被充分利用。

与现有使用深度表示的方法不同，本文的提案重点在于交叉视图，即在解码器层中同时提供源信息的两个不同视图，实现这一目标的方式是在最终编码器层的表示之上结合额外的源视图。最终模型的解码器由式(2)更新为：

$$\boldsymbol{T}_i = f_{decoder}(\boldsymbol{T}_{i-1}, \mathrm{LN}(g_i(\boldsymbol{S}) + \boldsymbol{S}_N)) \tag{5}$$

层归一化用于保持表示在深度神经网络中的稳定性，并允许将两个不同源视图的软集成到模型中。

在多层设置中，每个编码器层增加一个抽象级别，预计生成的表征粗粒度越来越高，即越来越抽象，并且其观察到其平均注意距离随深度[16]而增加，换句话说，不同编码层捕获的结构尺度存在一定差异。在解码器中其过程更为复杂，因为每个解码器层接收源信息和目标信息，这引发了如何正确结合不同源视图的问题，即函数 $g_i(\cdot)$ 在两个公式中，对此本文尝试了 3 种策略：抽象对应注意力(abstraction correspondence attention，ACA)、分层次注意力(hierarchical layer attention，HLA)、多层自适应注意力(multi-level adaptive attention，MLAA)，其中使用 HLA 策略的模型结构如图 1 所示，而其他模块所处位置均与 HLA 相同。

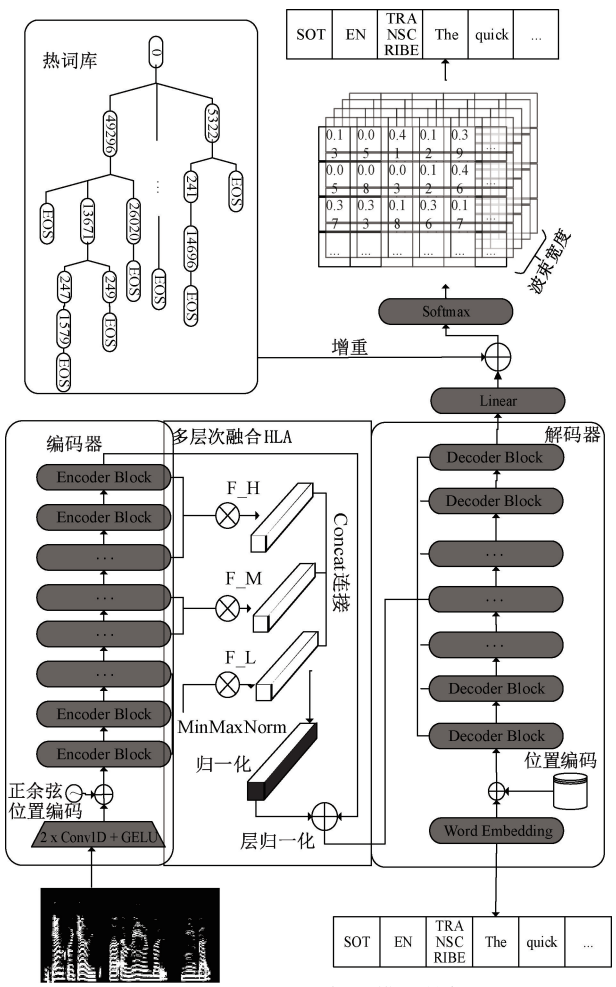


图 1　以 HLA 为例的模型结构

Fig. 1　Model structure using HLA

1)ACA

如模型使用所示，在此策略中，每一步解码被视为一个实现过程，抽象度高的表征逐渐在各层中变为具体和细化。因此，ACA 保持源视图在解码器层中的对抽象程度关注的一致性。由此，ACA 可以定义为：

$$g_i(\boldsymbol{S}) = \boldsymbol{S}_{N-i+1} \tag{6}$$

2)HLA

在这个策略中，为了整合跨多个层次的信息，本文从编码器的低层、中层和高层分别提取特征表示，记为 $\boldsymbol{F}_L$、$\boldsymbol{F}_M$ 和 $\boldsymbol{F}_H$，它们分别捕获了语音信号的细节特征、中尺度特征和全局语义信息，如图 1 中的多层次融合 HLA 模块所示。为了综合利用不同层次的特征，本文在特征维度上对 3 个层次的特征进行拼接，形成一个多层次的特征表示 $\boldsymbol{F}$：

$$\boldsymbol{F} = \mathrm{Concat}([\boldsymbol{F}_L, \boldsymbol{F}_M, \boldsymbol{F}_H]) \in \boldsymbol{R}^{N\times 3D} \tag{7}$$

式中：N 表示时间步数，D 表示每层特征的维度。通过拼接操作，每个时间步的特征维度从 D 扩展为 $3D$，综合了不同层次的信息。

对于每个层次的特征，本文分别计算注意力掩码，表示在每个时间步上，不同层次的特征对语音识别任务的贡献程度。具体地，对于 $i \in L, M, H$，注意力掩码 $\boldsymbol{a}_i$ 计算如下：

$$\boldsymbol{a}_i = \mathrm{softmax}(\boldsymbol{F}\boldsymbol{W}_i^a) \in \boldsymbol{R}^{N\times 1} \tag{8}$$

式中：$\boldsymbol{W}_i^a \in R^{3D\times 1}$ 是需要学习的权重矩阵，softmax 函数在时间步维度上进行归一化。

为了获得综合的注意力表示，本文将不同层次的注意力掩码进行归一化和融合，因此 HLA 定义为：

$$g_i(\boldsymbol{S}) = \mathrm{MinMaxNorm}\left(\sum_i \mathrm{MinMaxNorm}(\boldsymbol{a}_i)\right) \tag{9}$$

式中：MinMaxNorm 表示最小-最大归一化函数，确保注意力值在可比较的范围内。

3)MLAA

MLAA 通过注意力机制自适应地注入不同粗粒度级别的信息，进一步增强模型对不同粒度信息的利用能力，从而提高解码器对多层次特征的整合能力。

MLAA 的核心思想是为每个解码器层构建一个独立的向量，以预测注意力权重，从而实现不同层次信息的自适应匹配，定义为：

$$g_i(\boldsymbol{S}) = \sum_{j=1}^{N} \boldsymbol{\alpha}_{ij}\boldsymbol{S}_j \tag{10}$$

式中：$\boldsymbol{S} \in \boldsymbol{R}^{N\times D}$ 表示输入的特征序列，$g_i(\boldsymbol{S})$ 是第 i 个解码器层的输出，$\boldsymbol{\alpha}_{ij}$ 是由注意力机制计算得到的注意力权重，满足 $\sum_{j=1}^{N}\boldsymbol{\alpha}_{ij} = 1$。

对于每个解码器层 i，本文构建一个独立的注意力权重向量 $\boldsymbol{\alpha}_i$。注意力权重的计算过程如下：

$$\boldsymbol{\alpha}_i = \mathrm{softmax}(\boldsymbol{S}\boldsymbol{W}_i^a) \in \boldsymbol{R}^{N\times 1} \tag{11}$$

式中：$\boldsymbol{W}_i^a \in \boldsymbol{R}^{D\times 1}$ 是第 i 个解码器层的可学习权重矩阵，

softmax 函数确保所有注意力权重的和为 1。

将 MLAA 集成到解码器中,可以使每个解码器层根据自身需求动态地聚焦于不同粒度的信息。具体地,解码器的每一层通过 MLAA 计算得到的 $g_i(\boldsymbol{S})$ 被用作该层的上下文向量,进而指导注意力机制的计算。

1.3 添加 HB

1)构建 HB 向量

实现 HB 首先搭建了一个构音障碍热词库,热词库包含一组需要在识别过程中被增强的词汇,记为 $\boldsymbol{W}_{HW}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_M\}$。搭建构音障碍热词库使用预训练的大语言模型 Whisper,如 2.1 节所述,每个输入 $(\boldsymbol{X}_{1:T},\boldsymbol{Y}_{1:N})$ 都是一个(音频,文本)对,经过语音识别后输出 $\boldsymbol{y}_{1:N}$,如果 $\boldsymbol{y}_i\neq\boldsymbol{Y}_i$,则将 $\boldsymbol{Y}_i$ 添加入热词库。在实现上,热词库存储的是每个词汇的词汇表编码,以 0 为根节点,99 999 为终止符(EOS)。可以将整个热词库可以看成一个前缀树,当出现 $(\boldsymbol{y}_i\neq\boldsymbol{Y}_i,\cdots,\boldsymbol{y}_{i+n}\neq\boldsymbol{Y}_{i+n})$ 的情况时,将短语的词汇表编码以前缀树的形式存进热词库,如模型结构图 1 中的热词库所示。

每个热词 w_m 在词汇表中的索引为 v_m。为每个热词 w_m 分配一个增重分数 $\boldsymbol{S}_w(w_m)$,称为热词得分(hotword score, HS),其他词汇的分数为零。将分数赋值到对应的词汇表索引位置,构建 HB 向量 $\boldsymbol{S}_w$:

$$\boldsymbol{S}_w=\begin{cases}s, & \boldsymbol{y}\in W_{HB}\\ 0, & \text{其他}\end{cases} \tag{12}$$

在解码器输出词汇表概率分布后,将热词库的增重向量 $\boldsymbol{S}_w$ 添加到概率分布矩阵中,表示为:

$$\boldsymbol{P}(\boldsymbol{y}_t\mid\boldsymbol{y}_{<t},H)=\text{softmax}(\boldsymbol{W}\cdot d_t+b+\boldsymbol{S}_w) \tag{13}$$

2)在波束搜索中应用 HB

波束搜索是一种常用于序列生成问题中的启发式搜索算法[17],通过在每个时间步上选择 k 条累积概率最高的序列,计算它们的联合概率,找到更有可能的解,从而避免陷入局部最优解,其中称 k 为波束宽度(beam size,BS)。

在波束搜索中应用 HB 具体算法流程如下:

(1)初始化:从初始词汇 w_0 开始,设置 BS=k。

(2)添加 HB:在每个时间步 t 解码器计算词汇表中每个词概率分布,随后检查前 $2k$ 个序列是否存在 $\boldsymbol{y}\in\boldsymbol{W}_{HW}$,如果存在,则将热词库词汇表的额外权重概率添加到输出的概率分布中得到加重偏置后概率分布矩阵。需要注意的,当热词为连续单词时,在第一个单词进行概率加权,即仅对限定的词汇进行一次加重。

(3)扩展序列:每次自回归都从概率分布矩阵中选择 k 条概率最高的序列,由此并行产生了 k 条自回归路线。

(4)回溯序列:一旦当前进行了 HB,则一定要走到热词树前缀树的终点,因此一旦路线正处于 HB 的过程,需要检查当前路线上是否脱离热词库,如果检查当前词汇不符合已存在的连续词汇,则进行回溯,需要回溯到被判别为热词最开始的地方,并且减去为这个支线增加的分数,然后进行概率重排。

(5)终止条件:当所有候选序列达到 EOS 时,停止搜索。

2 实验设置

2.1 数据集

本文使用了两个英文的构音障碍数据集,其中:TORGO 数据集[18]是一个具有声学发音平行记录的无序语音语料库,包含 8 个构音障碍者和 7 个对照者。所有 7 名对照说话者的数据与 8 名构音障碍的说话者的 2/3 的数据合并为训练集(11.7 h),其余构音障碍的说话者的数据作为测试数据(1.79 h)。

UASpeech 语料库是最大的公开可用的无序语音语料库[19],由 15 名构音障碍者和 13 名对照说话者组成。来自所有 29 个说话者的块 1 和块 3 的数据用作训练集(69.1 h 的音频,总共 99 195 个话语),而从所有 16 个构音障碍说话者收集的块 2 的数据用作测试数据集(22.6 h 的音频,总共 26 520 个话语)。在去除语音音频段的开始和结束处的过度沉默之后,来自块 1 和块 3 的组合的总共 30.6 h 的音频数据被用作训练集,而来自块 2 的 9 h 的语音被用于测试。

2.2 实验参数设置

在本实验中,使用 NVIDIA GeForce GTX 3070Ti,在 Python3.9 版本使用 PyTorch 1.10.0 框架完成模型训练。模型训练使用 Adam 作为参数优化器,学习率设置为 1×10^{-5},批次大小均设置为最大化利用显存。

模型的语音特征输入使用 80 维的对数梅尔频谱图(Log-Mel spectrogram)。由于绝大多数音频都在 10 s 以内,因此所有音频被截断或者填充至十秒。所有音频重采样为 16 kHz,并通过短时傅里叶变换获得频谱,然后使用 Mel 滤波器组将频率从线性刻度转换到 Mel 频率刻度来模拟人耳听觉特性,随后对 Mel 频谱图的幅度值取对数,得到 Log-Mel 频谱图,便于模拟人耳对声音强度的感知。一名构音患者的音频时域图和 log-Mel 频谱图如图 2 所示。

2.3 评价指标

为了评估语音识别模型的性能,本文采用了常用的评价指标——字错误率(word error rate, WER)。

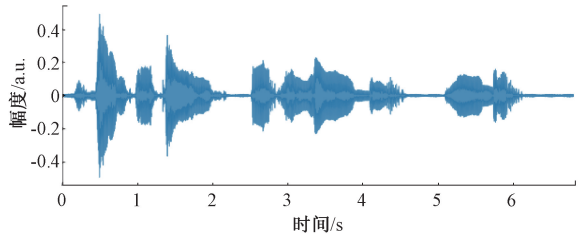
WER 衡量模型生成的转录与参考文本之间的差异,它是根据编辑距离计算的,即需要通过替换、删除、插入操作将识别结果转换为参考文本的最小操作数。

WER 通过式(4)计算:

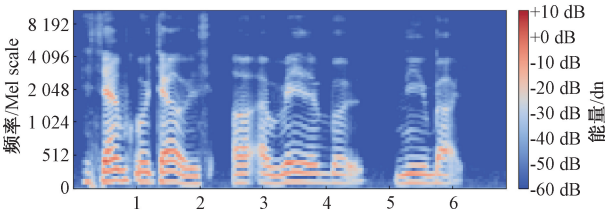
$$WER=\frac{S+D+I}{N} \tag{14}$$

式中:S 表示替换错误的数量,即错误地将一个词识别为另一个词的次数;D 表示删除错误的数量,即参考文本中的词未被正确识别的次数;I 表示插入错误的数量,即识别结果中多余的词;N 是参考文本中的总词数。

WER 值越低,表示模型的识别精度越高。WER 为



（a）构音障碍患者音频时域图
（a）Time-domain waveform of dysarthric speech



（b）构音障碍患者音频的log-Mel频谱图
（b）Log-Mel spectrogram of dysarthric speech

图 2　TORGO 数据集 M02 语音-"Some hotels are available nearby"

Fig. 2　TORGO dataset M02 utterance-"Some hotels are available nearby"

0 时，表示识别结果与参考文本完全一致。该指标广泛用于 ASR 系统的评估，能够有效反映模型的实际使用性能。

3　实验结果

3.1　对比实验

为验证所提出模型的性能优越性，本研究在测试集上对比了 5 种代表性模型，并与本文提出的改进模型（Whisper＋HLA＋HB）进行性能评估。这些模型包括 HMM-DNN、CTC、HuBERT、Wav2vec2 和 Whisper 模型。其中 HMM-DNN 和 CTC 模型基于 Kaldi 工具包实现，使用与本文一致的预处理流程和训练集；HuBERT、Wav2vec2 和 Whisper 模型从 Huggingface 开源模型库下载，分别对其官方预训练版本进行了全参微调，所有模型均经过适配以确保在 TORGO 和 UASpeech 数据集上的公平对比。

表 1 展示了不同模型在 TORGO 和 UASpeech 数据集上的整体字 WER 对比。传统机器学习模型 HMM-DNN 在与普通语音相近程度更高的 TORGO 数据集上的 WER 为 37.42%，深度学习模型的效果更佳，其中 Whisper 模型的 WER 降至 16.38%，改进后的方法则又将 WER 降低了 3.7%。UASpeech 数据集有更难理解的发音，HMM-DNN 方法的 WER 仅为 69.79%，而深度学习方法的效果有大幅提升，HuBERT 与 Wav2vec2 模型识别效果均优于传统语音识别，特别是 Whisper 模型达到了 38.31%的 WER，表现了深度学习在处理构音障碍语音时的优势。本文改进模型进一步将 WER 降低了 11%，表明多层次表征融合解码策略对端到端的语音识别模型的提升效果。

表 1　对比各方法在 TORGO 和 UASpeech 上的 WER

Table 1　WER of methods on TORGO and UASpeech

%

方法	TORGO	UASpeech
Whisper＋HLA＋HB	**12.67**	**27.18**
HMM-DNN	37.42	69.79
CTC	27.33	49.91
HuBERT	24.07	52.83
Wav2vec2	21.64	42.45
Whisper	16.38	38.31

表 2 进一步分析了各个深度学习模型对于不同可理解水平的语音识别表现，本文提出的改进模型在所有可理解度水平下均取得最佳表现，尤其是在中低理解水平的语音上有较大提高。在高可理解度语音，即发音较清晰且韵律较规则的语音上，各模型表现相对接近。在低可理解度语音中，本文模型的 WER 为 23.91%，相比 Whisper 模型的 42.36%，显著降低了 18.45%。在很低可理解度语音，即发音严重扭曲的语音上，本文模型的 WER 为 51.32%，显著优于 Whisper 模型的 71.43%，进一步验证了本文解码策略在极端语境下的鲁棒性和泛化能力。这表明本文采用的 Whisper 模型，以及对其进行的定制化解码器优化改进在构音障碍语音识别任务中具有强大的适应性和泛化能力。

表 2　对比不同方法在 UASpeech 不同可理解度划分的 WER

Table 2　WER of methods at various intelligibility levels on the UASpeech dataset

%

方法	UASpeech 不同可理解度水平			
	高	中	低	很低
Whisper＋HLA＋HB	**13.12**	**21.73**	**23.91**	**51.32**
HuBERT	28.76	43.50	58.85	85.41
Wav2vec2	19.88	37.75	45.71	71.73
Whisper	14.01	30.58	42.36	71.43

3.2　消融实验

在本节中，本文单独分析了每一部分对模型改进的影响。其中基准模型采取标准的 whisper-large-v2 模型，其中 BS＝5；Whisper＋HLA 模型，在基准模型的解码器部分采取多层表征融合策略中的 HLA 策略；Whisper＋HB 模型，在基准模型使用 HB；Whisper＋HLA＋HB 模型，将多层表征融合和 HB 都应用于基准模型。实验结果如表 3 所示。

从表 3 的实验结果中可以看到，HLA 与 HB 技术对两个数据集的影响效果不一样，其共同作用使两个数据集的

表 3 对比不同模块对 WER 的影响

Table 3 Impact of different modules on WER %

方法	TORGO	UASpeech
Whisper+HLA+HB	**12.67**	**27.18**
Whisper	16.38	38.31
Whisepr+HLA	15.17	30.12
Whisepr+HB	13.90	36.72

WER 取得了最低。结果显示在 TORGO 数据集上,HB 的效果更好,而在 UASpeech 数据集上,表征融合注意力带来的效果提升更大。

使用 Whisper+HLA 模型在两个数据集上 WER 均有所降低。特别是在 UASpeech 数据集上降低了 8%,这表明 HLA 的多层次融合注意力策略可以很好地应对构音障碍语音的复杂性。这种策略通过逐层细化对语音信息的抽象,使得解码器可以在不同的粒度级别上结合全局和局部信息,从而优化模型对模糊发音和多音节词汇的解码能力。

使用 HB 技术使 TORGO 数据集的 WER 从 16.38% 降至 13.90%,这可以说明特定词汇在句子中的上下文增强改善了模型的整体识别率;在 UASpeech 数据集上,尽管 HB 效果不如 TORGO 显著,但仍然略微降低了识别错误率。这表明,即使在孤立词的情况下,某些常见词汇的加权也能带来小幅提升。分析其效果差别可能是因为 TORGO 数据集提供了包含完整句子的音频样本,上下文信息充足,HB 技术能够充分发挥其增强语境的作用。而在 UASpeech 数据集上,由于该数据集大部分是孤立单词,缺乏上下文信息,使得 HB 在流畅度提升上的效果有限。

3.3 不同多层次融合策略的对比实验

在本节实验中本文为多层融合解码策略引入了不同解码路径以最大化模型的整体性能。在实验中,本文测试了 3 种解码策略,包括 ACA、HLA 和 MLAA。这些策略分别从不同的视角处理解码过程,以最大化模型的性能,实验结果如表 4 所示。

表 4 对比不同表征融合策略的 WER

Table 4 Impact of different feature fusion strategies on WER %

方法	TROGO	UASpeech
Whisper	16.38	38.31
Whisper+ACA	15.76	34.19
Whisper+HLA	**15.17**	**30.12**
Whisper+MLAA	16.11	34.67

ACA 策略能够在解码器中保持抽象一致性,使得信息在各层解码时信息逐步细化和完善,从而降低 WER。该策略对数据集的识别效果有所提升,尤其是在 UASpeech 数据集上取得了显著的性能改善。这可能是因为 ACA 能够有效地捕获不同层次的抽象信息,适用于具有复杂语音特征的 UASpeech 数据集。然而,由于缺乏全局信息的整合能力,ACA 在处理句子全局结构上表现略逊,因此在 TORGO 数据集上的提升不如在 UASpeech 数据集上明显。

HLA 策略既考虑了编码器和解码器抽象程度的一一对应,又通过高中低 3 个层次的划分使得周围的编码器表征能够局部融合,使解码器不至于仅关注单一表征而忽视全局信息。实验结果显示,这种折中的策略对效果提升最佳,无论是在 TORGO 数据集还是 UASpeech 数据集上,HLA 都有效地降低了词错误率,取得了最优的性能。这表明 HLA 能够更全面地利用多层次特征,增强模型对不同粒度信息的整合能力。

MLAA 策略旨在通过注意力机制自适应地注入不同粒度级别的信息,进一步增强模型对多层次信息的利用能力。但是如表 3 所示,MLAA 的性能并未超越 HLA。尽管 MLAA 在两个数据集上均较基线模型有所提升,但在 TORGO 数据集上仅将 WER 从 16.38%降至 16.11%,提升幅度有限;在 UASpeech 数据集上,WER 降至 34.67%,略低于 ACA 的 34.19%。导致 MLAA 提升有限的原因可能在于其自适应机制引入了额外的复杂度和参数开销,增加了模型训练的难度。自适应注意力需要模型学习到最优的注意力权重分配,对于数据量相对有限的语音识别任务,可能无法充分发挥其优势。此外,MLAA 对超参数的敏感度较高,可能需要更精细的调优才能达到最佳性能。

综上所述,实验结果表明,将多层次表征融入解码器能够提升语音识别模型的性能。在所测试的 3 种策略中,HLA 展现了最为显著的效果提升。这说明在解码过程中,平衡地结合不同抽象层次的信息,有助于解码器更准确地捕获语音信号中的复杂结构,提升对全局和局部信息的建模能力。

3.4 不同波束搜索设置的实验结果

本文在实验中比较了不同的 BS 与不同的 HS,在 Whisper+HLA+HB 模型下以 TORGO 数据集为例,不同 BS 与 HS 对 WER 的影响的实验结果如图 3 所示。可以看到随着 BS 的增加,语音识别变得更加精准,最终 BS 增大带来的效果提升逐渐趋缓。理论上 BS 越高,在进行语音识别时容错率越高,但是对于显存消耗也越高,因此本文基于 BS=5 研究。同时模型采用不同的 HS 进行加重也有不同的反应,实验结果显示 HS=3 时效果最好,HS 增加或减少都会使 WER 上升。同时两个数据集对 HB 的适应性不同。结合表 2 来看,由于 TORGO 数据集的句子和短语混杂,使得 HB 在其上的优化效果更好,而 UASpeech 数据集主要是孤立单词,导致在 HB 优化时效果有所限制。故 HS 的设置需要根据数据集特性进行调整,以适应不同数据类型的需求。

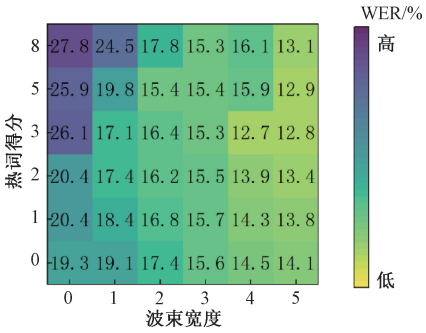


图 3　HB 参数对 WER 的影响

Fig. 3　The impact of HB parameters

4　结　　论

本文研究了提高构音障碍语音识别准确性的方法，提出了一种基于多层融合解码策略和热词增重技术的创新算法。实验在 TORGO 数据集和 UASpeech 数据集进行，结果显示表征融合注意力和 HB 技术对两个数据集的提升效果有所不同，但其共同作用使两个数据集的 WER 都达到了最低。本文探索了 3 种不同多层表征融合注意力策略，这 3 种策略都可以改善构音障碍语音识别，其中分层次表征融合注意力策略在构音障碍语音识别中取得了最好的成果，通过平衡地结合不同抽象层次的信息，帮助解码器更准确地捕获语音信号中的复杂结构，提升了对全局和局部信息的建模能力，能够在复杂语境和不规则发音中取得更好的识别效果。同时通过热词库的引入，低成本的提高了构音障碍语音识别的准确性。HB 技术对两个数据集的识别效果受数据集影响较大，TORGO 数据集提供了包含完整句子的音频样本，所以 HB 技术能够充分发挥其增强语境的作用，而 UASpeech 数据集主要由孤立单词组成，缺乏上下文信息，HB 的效果相对有限。本文的模型在两个构音障碍语音数据集上均取得了优秀表现，验证了所提出方法在处理复杂语音任务中的有效性和适应性。这一方法不仅提升了现有模型在特定数据集上的识别率，还为构音障碍语音识别领域提供了新的思路和方向。未来将更深入探索构音障碍语音识别，探索建立更精准的热词库与可自适应的 HB 技术，研究解码器使用更低资源获取更全面的表征信息。

参考文献

[1]　宋伟，张杨豪. 构音障碍语音识别算法研究综述[J]. 计算机工程与应用，2024，60(11)：62-74.
SONG W，ZHANG Y H. A survey of specific speech recognition algorithms for dysarthria[J]. Computer Engineering and Applications，2024，60(11)：62-74.

[2]　XIONG F F，BARKER J，CHRISTENSEN H. Deep learning of articulatory-based representations and applications for improving dysarthric speech recognition[C]. Speech Communication，13th ITG-Symposium，2018：1-5.

[3]　张楠，陈媛媛，陈鑫钰，等. 基于 LMD 改进特征提取的三路病理语音识别[J]. 电子测量技术，2024，47(12)：140-147.
ZHANG N，CHEN Y Y，CHEN X Y，et al. Three channel pathological speech recognition based on LMD improved feature extraction [J]. Electronic Measurement Technology，2024，47(12)：140-147.

[4]　张小恒，李勇明，王品. 双阶段帕金森病语音聚类包络卷积稀疏迁移学习算法[J]. 仪器仪表学报，2022，43(11)：151-161.
ZHANG X H，LI Y M，WANG P. Two-stage PD speech clustering envelope and convolution sparse transfer learning algorithm[J]. Chinese Journal of Scientific Instrument，2022，43(11)：151-161.

[5]　钱金阳. 生成对抗网络数据扩增下的病理语音识别研究[D]. 苏州：苏州大学，2023.
QIAN J Y. Pathological voice recognition based on data augmentation with generative adversarial networks[D]. Suzhou：Suzhou University，2023.

[6]　WANG H L，RAVICHANDRAN V，RAO M，et al. Improving fairness for spoken language understanding in atypical speech with text-to-speech[J]. ArXiv preprint arXiv：2311.10149，2023.

[7]　ALMADHOR A，IRFAN R，GAO J CH，et al. E2E-DASR：End-to-end deep learning-based dysarthric automatic speech recognition[J]. Expert Systems with Applications，2023，222：119797.

[8]　SHAHAMIRI S R. Speech vision：An end-to-end deep learning-based dysarthric automatic speech recognition system[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering，2021，29：852-861.

[9]　HU SH J，XIE X R，JIN Z R，et al. Exploring self-supervised pre-trained asr models for dysarthric and elderly speech recognition[C]. IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2023：1-5.

[10]　YU CH CH，SU X S，QIAN ZH P. Multi-stage audio-visual fusion for dysarthric speech recognition with pre-trained models[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering，2023，31：1912-1921.

[11]　PADI B，MOHAN A，GANAPATHY S. End-to-end language recognition using attention based hierarchical gated recurrent unit models[C]. IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). IEEE，2019：5966-5970.

[12] VASWANI A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30, DOI:10.48550/arXiv.1706.03762.

[13] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision [C]. International Conference on Machine Learning. PMLR, 2023: 28492-28518.

[14] ZHAO Z P, BAO ZH T, ZHANG Z X, et al. Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 14(2): 423-434.

[15] 张丽彩, 李鸿燕, 司马飞扬, 等. 基于多层注意力机制的4DC-BGRU脑电情感识别[J]. 电子测量技术, 2023,46(8):134-141.
ZHANG L C, LI H Y, SIMA F Y, et al. EEG emotion recognition by 4DC-BGRU based on multi-level attention mechanism[J]. Electronic Measurement Technology,2023,46(8):134-141.

[16] WANG R T, SHEN Y Q, ZUO W L, et al. Transvpr: Transformer-based place recognition with multi-level attention aggregation [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 13648-13657.

[17] VIJAYAKUMAR A K, COGSWELL M, SELVARAJU R R, et al. Diverse beam search: Decoding diverse solutions from neural sequence models[J]. ArXiv preprint arXiv:1610.02424, 2016.

[18] RUDZICZ F, NAMASIVAYAM A K, WOLFF T. The TORGO database of acoustic and articulatory speech from speakers with dysarthria[J]. Language Resources and Evaluation, 2012, 46: 523-541.

[19] KIM H, HASEGAWA J M, PERLMAN A, et al. Dysarthric speech database for universal access research [C]. Interspeech 2008, 9th Annual Conference of the International Speech Communication Association, 2008: 1741-1744.

作者简介

樊紫岩,硕士研究生,主要研究方向为深度学习和智能应用等。

E-mail:202230603058@mails.zstu.edu.cn

朱耀东(通信作者),博士,教授,主要研究方向为智能机器人与健康物联网。

E-mail:zhuyaodong@163.com

赵伟岚,硕士研究生,主要研究方向为深度学习与智能应用等。

E-mail:2484497912@qq.com

李轩逸,硕士研究生,主要研究方向为机器学习与智能应用等。

E-mail:5756978937@qq.com