doi：10.13413/j.cnki.jdxblxb.2023252

# 基于带阈值的 BPE-dropout 多任务学习的端到端语音识别

马 建，朵 琳，韦贵香，唐 剑
（昆明理工大学 信息工程与自动化学院，昆明 650500）

**摘要**：针对语音识别任务中出现的未登录词问题，提出一种带阈值的 BPE-dropout 多任务学习语音识别方法. 该方法采用带随机性的字节对编码算法，在形成子词时引入带字数阈值的策略，将子词作为建模单元，编码器部分采用 Conformer 结构，与链接时序分类和注意力机制相结合. 为进一步提升模型性能，引入动态参数对损失函数进行动态调节，并同时进行多任务训练和解码. 实验结果表明，该方法采用子词作为建模单元可有效解决未登录词问题，在多任务学习框架下进一步提升了模型的识别性能. 在公开数据集 THCHS30 和 ST-CMDS 上，该模型实现了超过 95%的识别准确率.

**关键词**：语音识别；多任务学习；字节对编码；动态调节参数
**中图分类号**：TP391；TN912.3 **文献标志码**：A **文章编号**：1671-5489(2024)03-0674-09

# End-to-End Speech Recognition Based on Threshold-Based BPE-Dropout Multi-task Learning

MA Jian, DUO Lin, WEI Guixiang, TANG Jian
(*Faculty of Information Engineering and Automation, Kuming University of Science and Technology, Kunming* 650500, *China*)

**Abstract**: Aiming at the problem of unknown words in speech recognition tasks, we proposed a threshold based-BPE-dropout multi-task learning speech recognition method. This method adopted a random byte pair coding algorithm. When forming sub-words, a strategy with word number threshold was introduced. The sub-words were used as modeling units, and the encoder part adopted Conformer structure, which was combined with link timing classification and attention mechanism. In order to further improve the performance of the model, dynamic parameters were introduced to dynamically adjust the loss function, and multi-task training and decoding were performed simultaneously. The experimental results show that the proposed method can effectively solve the problem of unknown words by using sub-words as modeling units, and further improve the recognition performance of the model under the multi-task learning framework. On the public datasets THCHS30 and ST-CMDS, the model achieves more than 95% recognition accuracy.

**Keywords**: speech recognition; multi-task learning; byte pair coding; dynamic adjustment parameter

**收稿日期**：2023-06-16.

**第一作者简介**：马 建(1998—)，男，汉族，硕士研究生，从事语音识别和信号处理的研究，E-mail：2703729898@qq.com. **通信作者简介**：朵 琳(1974—)，女，彝族，博士，副教授，从事语音识别和信号处理的研究，E-mail：duolin2003@126.com.

**基金项目**：国家自然科学基金(批准号：61962032).

自动语音识别(automatic speech recognition，ASR)技术是将人类可懂语音转换为可读性的文本，是实现人机交互的关键技术[1]. 目前，ASR 产品已广泛应用于人们的日常生活，如苹果的 Siri、亚马逊的 Alexa、讯飞语音输入法、叮咚智能音箱等. 语音识别是模式识别的一部分，它主要利用模型算法检测语音向量，即语音特征参数，并最终传递识别结果[2]. 语音识别的最终目标是准确识别不同环境下的语音内容. 传统的高斯混合-隐 Markov 模型(GMM-HMM)是语音识别序列系统的主流方法. HTK[3]是一种基于隐 Markov 模型(HMM)的语音识别工具. 近年来，由于深度学习框架的崛起及神经网络处理非线性模型和大数据的能力，语音识别模型取得了许多研究成果，其中深度置信网络[4](deep belief network，DBN)与 HMM 相结合的声学模型在小词汇量连续语音识别中取得成功，深度神经网络与 HMM 结合的声学模型 DNN-HMM 在大量词汇连续语音识别(large vocabulary continuous speech recognition，LVCSR)中取得成功[5]，开启了利用深度学习进行语音识别的新时代. 同时卷积神经网络(convolutional neural network，CNN)[6]和深度神经网络(deep neural network，DNN)[7]也广泛应用于自动语音识别任务中[8]. 相比 DNN，CNN 能有效利用图像特征以及权值分布和池化操作，为模型提供更好的图像和不变性，从而提高卷积神经网络在语音识别中的准确率.

近年来，语音识别领域通过引入端到端语音识别模型取得了显著进展. 端到端语音识别模型简化了传统的神经网络结构，采用单一的结构，不再需要传统的发音字典单元，可直接预测字和单词. 这种简化结构使端到端语音识别模型在准确率方面超过了传统的语音识别模型. 目前，有两种主要算法可用于实现端到端语音识别模型：一种是基于链接时序分类(connectionist temporal classification，CTC)[9]的语音识别结构；另一种是基于注意力机制的语音识别结构.

目前，基于注意力机制的 Transformer 模型[10]在各种语音识别任务中逐渐取代了传统的循环神经网络(RNN)模型. 这是因为 Transformer 模型具有捕获长距离语音特征信息和高度并行训练的能力，而卷积神经网络则擅长提取局部细粒度特征. 通过引入注意力机制，Transformer 模型能同时处理整个输入序列，而不像 RNN 模型需要按顺序逐步计算，使得 Transformer 模型能高效地并行计算，从而显著加快了训练速度和推理速度. 在此基础上研究者们提出了 Conformer 模型[11]，该模型既能捕获长距离信息又能提取局部特征信息，在端到端语音识别任务中识别性能优异.

中文作为一个汉字文字系统，具有非常庞大的词汇量，任何一个固定的词汇表都无法涵盖所有可能的中文词汇. 随着时代的发展和社会的变迁，又不断创造出新的词汇，特别是在科技、互联网和流行文化等领域. 这些新词汇可能在现有的词汇表中没有被包含，因此被称为未登录词. 此外，在实际语音交流中，人们可能会存在语音模糊、省略、连读等现象，这些现象会导致词汇的发音发生变化，使某些词汇在语音识别中被错误地认为是未登录词.

针对上述问题，本文提出一种带阈值的 BPE-dropout 多任务学习的端到端语音识别模型，该模型采用带随机性字节对编码算法，在形成子词时引入带字数阈值的策略，将子词和字作为建模单元进行多任务训练和解码，其中字任务作为辅助任务. 编码器部分采用 Conformer 结构，与链接时序分类和注意力机制相结合. 为进一步提升模型性能，引入动态参数对损失函数进行动态调节，并同时进行多任务训练和解码. 测试结果表明，本文方法能有效解决未登录词问题，并提升了语音识别性能.

## 1　语音特征提取

特征提取是语音识别的重要组成部分. 针对目前语音识别率低、收敛慢等问题，本文采用 Mel 语谱图特征作为输入特征，提取该特征的过程如图 1 所示.




**图 1　特征提取流程**

**Fig. 1　Feature extraction process**

### 1.1　Mel 语谱的生成

在语音识别领域研究中，特征参数提取非常关键. 为有效捕捉语音信号的时频特性，本文使用语

谱图作为一种有效的表示方式，它直接展示了语音信号在时域和频谱上的变化[12]. 由于人类耳蜗结构的特殊性，决定了听觉频率的空间分布呈对数关系. 因此采用Mel语谱[13]可以更好地模拟人类听觉系统对声音频率的感知方式，与人耳的听觉特征相匹配. Mel语谱特征提取步骤如下：

1) 使用汉明窗对语音信号 $x(n)$分帧加窗，分帧加窗后的第 $i$ 帧信号为 $x_i(n)$；

2) 对 $x_i(n)$进行离散Fourier变换(DFT)，有

$$S_i(k)=\sum_{n=0}^{N-1}x_i(n)\mathrm{e}^{-\mathrm{j}2\pi kn/N},$$

其中 $N$ 为离散Fourier变换的点数；

3) $S_i(k)$的能量谱密度为

$$P_i(k)=|S_i(k)|^2;$$

4) 在能量谱上应用Mel滤波器组，其公式为

$$Y_t(m)=\sum_{k=0}^{N-1}H_m(k)|P_i(k)|^2,$$

对 $Y_t(m)$取对数变换，有

$$Y_t^{\mathrm{dB}}(m)=10\lg Y_t(m),$$

其中 $k$ 表示DFT变换后的编号，$m$ 表示Mel滤波器的编号.

5) 以时间为横轴、对数化频率为纵轴，二维图像绘制的Mel语谱如图2所示.

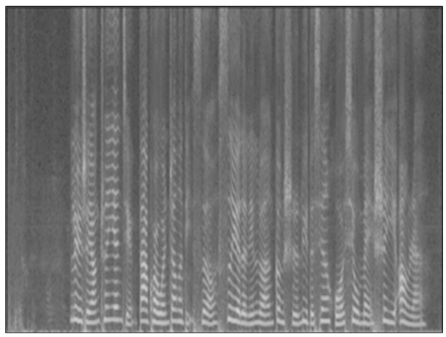


**图2 Mel语谱**

**Fig. 2 Mel spectrogram**

### 1.2 Mel滤波器组

由于人耳听到的声级与听觉频率不呈线性关系，因此Mel频率标度更符合人耳的听觉特性. Mel频率与实际频率的关系如下：

$$m=2\ 595\times\lg(1+700f), \tag{1}$$

$$f=100(10^{m/2\ 595}-1), \tag{2}$$

其中 $m$ 是以Mel为单位的感知频率，$f$ 是以Hz为单位的实际频率. 如果以e为对数，则系数为1 125. Mel滤波器组是一系列三角形滤波器，中心频率响应为1，线性度降低至0，直到两个相邻滤波器的中心频率响应为0. 计算公式为

$$H_m(k)=\begin{cases}0, & k\leqslant f(m-1),\\ \dfrac{k-f(m-1)}{f(m)-f(m-1)}, & f(m-1)\leqslant k\leqslant f(m),\\ \dfrac{f(m+1)-k}{f(m+1)-f(m)}, & f(m)\leqslant k\leqslant f(m+1),\\ 0, & k\geqslant f(m+1),\end{cases} \tag{3}$$

$$f(m)=\left(\frac{N}{f_s}\right)F_{\mathrm{Mel}}^{-1}\left(F_{\mathrm{Mel}}(f_l)+m\,\frac{F_{\mathrm{Mel}}(f_h)-F_{\mathrm{Mel}}(f_l)}{M+1}\right). \tag{4}$$

## 2 模型结构及方法

### 2.1 连接时序分类

连接时序分类(CTC)算法作为一个目标函数，可直接对标签函数和语音函数之间的映射进行建模，而无需在帧级对输入和输出信息进行预对齐. 在语音CTC的识别任务中，输入序列 $X=\{x_t|t=1,2,\cdots,T\}$，标签序列 $Z=\{z_i|i=1,2,\cdots,I\}$，其中目标序列的长度不能超过输入序列的长度. 设 $L$ 为语音识别任务建模中最小的建模单元，即字典.

在CTC网络中，除最小建模单元 $|L|$ 外，还在其输出层添加了一个“—”字符，对输入序列 $X=\{x_t|t=1,2,\cdots,T\}$中存在的重复标签和空白标签进行建模，以完成对齐任务[14]. CTC通常应用于递归神经网络上，每个RNN输出单元被解释为在特定时间观察到相应标签的概率. 标签序列 $P(\pi|x)$的

概率被建模为与网络输出的乘积条件无关：

$$P(\pi|x) \approx \prod_{t=1}^{T} P(\pi_t|x) = \prod_{t=1}^{T} q_t(\pi_t), \tag{5}$$

其中 $q_t(\pi_t)$表示时间 $t$ 时 RNN 输出层 $q$ 中 $\pi_t$ 标签的最大激活.

用 CTC 训练模型以最大化 $P(y|x)$，即所有可能标签序列的概率分布为

$$P(y|x) = \sum_{\pi \in \varphi(y')} P(\pi|x), \tag{6}$$

其中 $y'$是 $y$ 的修改标签序列，其通过在每个标签的开始和结束之间插入空白标签实现，例如若 $y=(aa,abb)$，则 $y'=(—aa—,—abb—)$. 概率分布 $P(\pi|x)$可使用向前向后算法有效地计算为

$$P(\pi|x) = \sum_{u=1}^{|y'|} \frac{a_t(u)b_t(u)}{q_t(y'_u)}, \tag{7}$$

其中：$a_t(u)$为正向变量，表示以第 $u$ 个标签结尾的所有可能($y'_1$：$u$)的总概率；$b_t(u)$是通过对任何标签(包括空标签)的损失函数求导，可使用标准反向传播对网络进行训练.

## 2.2　Conformer 模型

Conformer 模型已在当前主流的端到端语音识别任务中广泛应用，其整体结构可分为 3 个模块：前馈网络模块、多头注意力模块和卷积网络模块，如图 3 所示.




**图 3　Conformer 模型结构**

**Fig. 3　Conformer model structure**

Conformer 模型是一种序列建模架构，它融合了多个关键组件，并通过残差连接实现它们之间的连接. 这些组件包括多头注意力模块、卷积网络模块和前馈网络模块. 多头注意力模块使用类似于 Transformer-XL 的方法计算序列中的位置编码信息，可以有效捕捉输入语音特征序列中的重要语音特征信息. 在多头注意力模块前后，使用批归一化层和 Dropout 层进行正则化处理，这些正则化技术有助于提高模型的鲁棒性和泛化能力，确保模型在不同输入样本上性能稳定.

卷积网络模块由逐点卷积网络、ReLU 激活函数和一维深度卷积网络组成，它能有效捕捉输入特征序列中的局部细节语音特征信息. 逐点卷积网络通过在输入序列上进行逐点操作，提取出每个位置的特征表示；ReLU 激活函数则引入非线性性质，增强模型对不同语音特征的区分能力；一维深度卷积网络通过多层卷积操作，逐渐扩大感受野，学习到输入序列中更复杂和抽象的特征表示. 同理，在卷积网络模块内部也使用批归一化层和 Dropout 层进行正则化处理.

前馈网络模块在 Conformer 模型中扮演重要角色，它由两个线性变换层和 Swish 激活函数构成. 该模块引入了非线性变换，以更好地捕捉输入特征的复杂关系. 为提高模型的鲁棒性和泛化能力，在前馈网络模块中采用层归一化和 Dropout 层进行正则化处理. 这些正则化技术有助于平衡模型的输入分布，并减少过拟合的风险. Conformer 模型借鉴了 Macaron-Net 网络结构的思想，将前馈网络模块分别放置在多头注意力模块之前和卷积网络模块之后. 这种设计使模型充分利用多头注意力模块对全局上下文的建模能力，以及卷积网络模块对局部细节的建模能力. 同时，通过在各模块之间添加残差连接，有助于信息的传递并减轻梯度消失问题.

## 2.3　建模单元

在语音识别任务中，所选择的建模单元对模型的识别性能具有重要影响，因为它可能导致大量未登录词的出现. 如果选择以字为建模单元，可能会遇到序列过长的问题，使模型难以学习到序列之间的依赖关系. 针对该问题，本文选择子词作为建模单元. 这种选择既有助于减小词汇量的规模，又能有效解决未登录词的问题. 通过将语言单位划分为更小的子词，能更好地捕捉单词内部和单词之间的语义信息，从而提高模型对复杂文本的理解能力.

目前广泛使用的方法是采用字节对编码(byte pair encoding, BPE)算法生成子词. 字节对编码是一种用于数据压缩和自然语言处理的算法, BPE 的基本思想是通过反复合并出现频率最高的字节对(byte pairs)构建一个编码表, 从而实现对数据的压缩或表示. 字节对编码算法的步骤如下:

1) 初始化, 将输入数据按字节(byte)进行划分, 形成初始的词汇表;

2) 构建频率表, 统计所有字节对的出现频率, 字节对可以是相邻的字节或字符;

3) 选择合并, 选择出现频率最高的字节对, 并将它们合并成一个新的字节或符号;

4) 重复步骤 3), 直至达到指定的合并次数或者达到某个停止条件;

5) 生成编码表, 将输入数据中的每个字节映射到对应的合并后的字节或符号中.

本文使用改进的 BPE 算法——BPE-dropout, 它在字节对编码算法中引入了一定的随机性元素, 通过构造高频字符片段, 使这些子词在语料库中的出现频率足够高, 从而能有效训练并生成最佳的翻译模型. 在进行迭代合并频繁出现的字符对时, BPE-dropout 会以一定的概率 $p$ 丢弃某些合并操作. 当 $p=0$ 时, 采用普通字节编码算法的方式形成子词; 当 $p=1$ 时, 不进行任何合并操作, 只保留原始的字.

首先初始化词汇表, 将训练数据中的字符作为初始词汇表的元素, 每个字符都是一个独立的词, 然后统计每个字符对(由相邻字符组成)的出现频率. 从字符对频率表中选择出现频率最高的字符对, 将两个字符合并成一个新的字符. 同时更新词汇表, 将合并后的字符加入到词汇表中, 逐步增加词表. 更新字符对频率表, 将原始字符对替换为合并后的字符, 重新计算字符对的频率.

通过分析使用传统 BPE 算法生成的子词长度分布, 发现在处理汉文时, 最长的子词长度为 21 字节. 因此在实验中设置了一个子词字数阈值, 只提取长度不超过该阈值的子词. 带阈值的 DPE-dropout算法流程如图 4 所示.

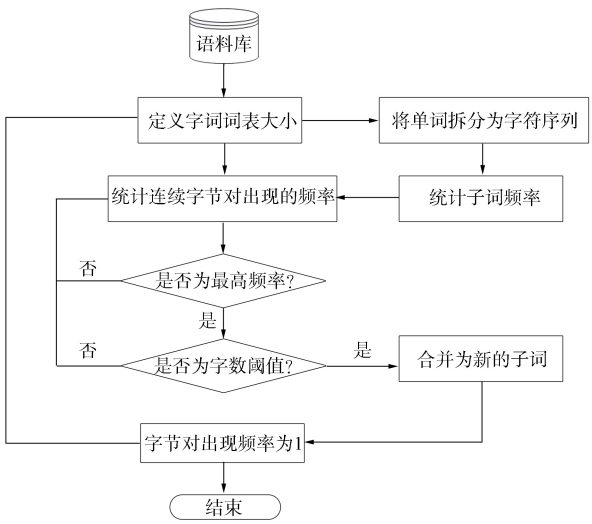



**图 4 带阈值的 DPE-dropout 算法流程**

**Fig. 4 Flow chart of DPE-dropout algorithm with threshold**

## 3 多任务学习模型

本文提出一种多任务学习模型, 其结构如图 5 所示. 该模型通过结合 Conformer 和 CTC-Attention的多层编码器-解码器网络构建而成. 每个编码器层包括多头注意力机制、卷积网络和前馈网络. 通过在不同层之间进行归一化和随机失活处理, 可提高模型的稳健性和泛化能力. 解码器部分采用链接时序分类和注意力机制, 以实现快速训练和联合解码.

该过程首先对输入的语音信号进行特征提取, 并对其进行降采样. 然后通过一个线性层将特征 $x=\{x_1,x_2,\cdots,x_T\}$送入到 Conformer 编码器, 在编码器中, 经过前馈网络、多头注意力和卷积网络的处理, 产生一个中间特征序列 $h=\{h_1,h_2,\cdots,h_T\}$. 最后将中间特征序列 $h$ 分别送入不同解码器中进行

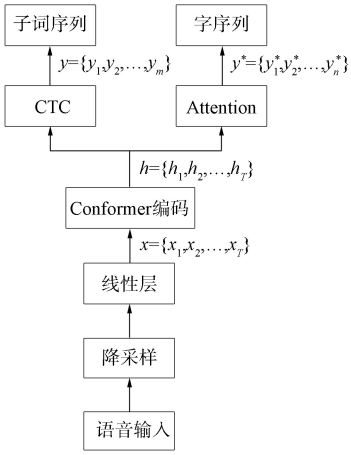



**图 5　多任务学习模型流程**

**Fig. 5　Multi-task learning model process**

处理，在每个子词解码器和字解码器中分别输出标签 $y=\{y_1,y_2,\cdots,y_m\}$ 和 $y^*=\{y_1^*,y_2^*,\cdots,y_n^*\}$，其中 $y_i$ 表示子词建模单元，$y_i^*$ 表示字建模单元.

在本文实验中，将子词任务作为主任务，字任务作为辅助任务同时训练多任务模型，使用 CTC 解码器和 Attention 解码器联合解码优化模型参数，因此在训练过程中采用多任务框架定义损失函数为

$$L_{\mathrm{MTL}}=\lambda L_{\mathrm{CTC}}+(1-\lambda)L_{\mathrm{Attention}},\ 0\leqslant\lambda\leqslant 1, \tag{8}$$

其中 $\lambda$ 表示用于设置损失函数的参数. 在解码过程中，采用波速搜索算法对得到的 Attention 和 CTC 的分量进行解码[15]. 波束大小为 20，以降低计算成本. 在式(8)中，加权系数的参数 $\lambda$ 须在训练开始前手动调整，并且不能在整个训练过程中更改，因此本文提出了动态权重调节的方法解决该问题：

$$\lambda=\ln\ \mathrm{sig}(L_{\mathrm{Attention}}-L_{\mathrm{CTC}})=\frac{1}{1.0+\exp\{L_{\mathrm{CTC}}-L_{\mathrm{Attention}}\}}. \tag{9}$$

改进后，当 $\lambda=0$ 时为 BiLSTM 架构，当 $\lambda=1$ 时为 CTC 架构. 在训练期间无需手动调整，且动态持续调整可更好地支持校准过程.

## 4　实验配置

### 4.1　数据集

利用 PyCharm 进行仿真实验，使用的软件为 TensorFlow1.15 版，Window10 操作系统，12 GB 内存，处理器为 Intel-i5-12400F. 本文使用的数据集是中文数据集 THCHS30 和 ST-CMDS，其中数据集 THCHS30 总持续时间超过 30 h，采样频率为 16 kHz，采样大小为 16 bit. 数据集 ST-CMDS 是一个人工智能数据公司发布的中文语音数据集，约为 100 h 的语音数据，数据内容主要包括在线语音聊天和智能语音控制指令. 数据集 THCHS30 信息列于表 1.

**表 1　中文语音数据集 THCHS30 信息**

**Table 1　Information of THCHS30 Chinese speech dataset**

| 数据集 | 语音时长 | 句子数 | 词数 |
|---|---|---|---|
| Train | 25 h | 10 000 | 198 252 |
| Dev | 2 h 14 min | 893 | 17 743 |
| Test | 6 h 15 min | 2 459 | 49 085 |

为使识别出的单词序列与默认序列匹配，必须插入、替换或删除某些单词. 插入、替换或删除的总字数除以标准序列中总字数的百分数即为词错误率(word error rate，WER)，本文采用 WER 作为评价指标，其计算公式为

$$\mathrm{WER}=\frac{\mathrm{S}+\mathrm{D}+\mathrm{I}}{N}\times 100\%, \tag{10}$$

其中 S 表示替换，D 表示删除，I 表示插入，$N$ 表示词数目.

### 4.2　模型参数配置

在本文给定模型下，通过优化学习率确定模型的最佳训练系数，并研究学习系数对模型识别精度的影响，结果如图 6 所示. 由图 6 可见，当学习率从 0.000 7 提高到 0.001 时，发现学习率为 0.000 8 时，在数据集 THCHS30 中模型的识别准确率最高，而当学习率为 0.000 85 时，在数据集 ST-CMDS 中模型的识别准确率最高. 而当学习率过低或过高时，模型易出现过拟合现象，因此需要进行适当的调整，以取得最佳的模型性能.

然后进行模型训练迭代周期的调试，以降低损失率. 图 7 为迭代周期与损失函数的变化关系. 由

图7可见，从第28次迭代开始曲线变得平缓，表示模型已经收敛. 迭代周期从25～33次相对合理，继续增加将导致过拟合. 因此，本文选择训练迭代周期为30次.

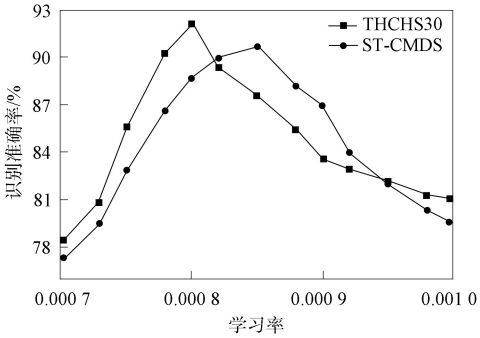



**图6 调整学习率对模型识别准确率的影响**

**Fig. 6 Effect of adjusting learning rate on recognition accuracy of model**

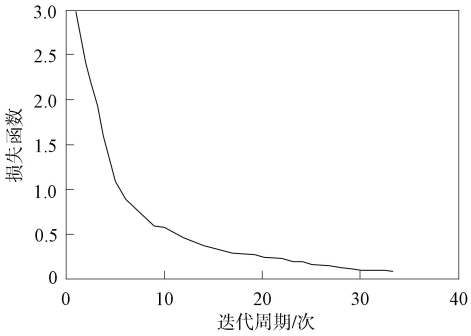



**图7 迭代周期与损失函数的变化关系**

**Fig. 7 Relationship between iteration period and loss function**

本文采用Conformer为编码器模型，以子词任务为主任务，字任务为辅助任务，采用多任务(MOL)框架优化损失函数，提升系统鲁棒性. 在式(8)中，将加权系数$\lambda$作为一个调整参数，对两个目标函数进行调整，结果如图8所示.

不同的$\lambda$取值对WER的影响不同. 由图8可见，当$\lambda$取值接近0.5时，语音识别模型的错误率最低. 实验结果表明，使用多任务学习框架能有效提高语音识别准确率.

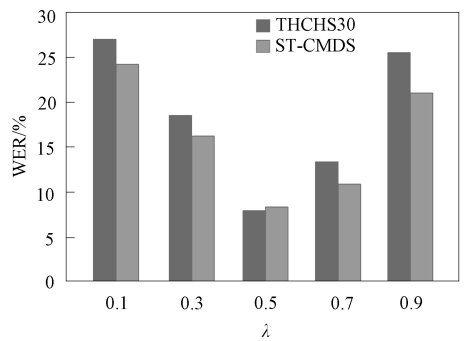



**图8 $\lambda$取值与模型在数据集上测试WER的关系**

**Fig. 8 Relationship between $\lambda$ value and testing WER by model on dataset**

### 4.3 不同模型下的识别性能

在语音识别领域，GMM-HMM[16]，DNN-HMM[17]和CNN-HMM[18]等都是目前流行的语音识别模型. 在真实环境下都采用Fbank特征作为语音输入特征. GMM训练速度快，声学模型较小，但无法利用帧的上下文信息，每一步迭代的计算量均较大；DNN能利用上下文信息，但无法利用历史信息辅助当前任务；CNN-HMM为DNN-HMM的改进优化模型，CNN比其他神经网络更能捕捉到特征的不变性. 本文为测试Conformer-CTC-Attention多任务学习模型与其他不同模型的性能，设计7组实验，验证Conformer-CTC-Attention多任务学习模型的识别效果. 实验结果列于表2.

**表2 两个数据集在不同模型下的WER**

**Table 2 WER of two datasets under different models**

| 系统模型 | WER/% | |
|---|---|---|
| | THCHS30 | ST-CMDS |
| GMM-HMM | 20.75 | 20.19 |
| DNN-HMM | 18.87 | 18.18 |
| CNN-HMM | 17.95 | 16.92 |
| BiLSTM-CTC/Attention | 11.25 | 11.64 |
| Transformer-CTC | 9.14 | 10.78 |
| Conformer-CTC | 7.95 | 8.35 |
| Conformer-CTC-Attention-MTL | 6.26 | 6.88 |

由表2可见，CNN-HMM相比于GMM-HMM,DNN-HMM模型，在两个数据集上的识别准确率分别提高了2.80,0.92个百分点和3.27,1.26个百分点.

BiLSTM-CTC/Attention模型[19]由编码器-解码器网络组成，编码器由3层双向LSTM和全连接层组成，解码器将CTC和多头注意力连接进行联合训练解码. 由表2可见，相比于CNN-HMM模型，

在两种数据集上的识别准确率分别提高了 6.70，5.28 个百分点. Transformer-CTC 模型[10]采用编码器-解码器结构，Transformer 编码器由多头注意力、前馈网络和归一化网络组成. Conformer-CTC 模型[20]同样采用编码器-解码器结构，Conformer 编码器内包含多头注意力、卷积网络、前馈网络、归一化网络. 二者皆以链接时序分类 CTC 作为解码器，解码过程中采用集束搜索算法. Conformer-CTC-Attention-MTL 模型采用多任务学习框架，采用 39 维 Mel 特征作为输入，以 BPE-dropout 方法形成子词作为主任务建模单元，动态地调整词汇表，以更好地适应具体任务或数据集的特点，以字任务作为辅助任务. 采用 Conformer 结构作为编码器，以链接时序分类和注意力机制进行联合解码，引入动态权重调节算法以提升模型性能. 实验结果表明，多任务学习模型的性能明显高于传统模型性能，相比于 GMM-HMM 模型，错误率分别降低了 14.49，13.31 个百分点. 这是由于共享层注意力和卷积神经网络，能有效捕捉长距离和细粒度特征信息.

表 3 列出了模型深度对识别性能(WER)的影响. 由表 3 可见，随着编码器层数的增加，模型的性能也相应提升，进一步验证了网络层数对模型学习辅助特征信息的影响. 当编码器层数达到 6 时，模型识别性能较好，而当编码器层数增加到 9 时，模型性能开始下降. 这是因为随着网络层数的增加，模型所需训练的参数数量也随之增加，导致模型易出现过拟合，识别性能下降.

**表 3　模型深度对识别性能(WER)的影响**

**Table 3　Effect of model depth on recognition performance (WER)**

| 模型层数 | WER/% | |
|---|---|---|
| | THCHS30 | ST-CMDS |
| MLT($L$=3) | 7.64 | 8.05 |
| MLT($L$=6) | 6.26 | 6.88 |
| MLT($L$=9) | 6.87 | 7.32 |

表 4 列出了 BPE 和 BPE-dropout 形成的建模单元对识别性能(WER)的影响. 由表 4 可见，使用 BPE-dropout 方法生成的子词作为建模单元，识别性能较好. 相比于普通的字节编码算法(BPE)，BPE-dropout 方法在不同数据集上错误率分别下降了 1.98，0.75 个百分点. 此外，丢弃率对识别性能也有一定的影响.

**表 4　BPE 和 BPE-dropout 形成的建模单元对识别性能(WER)的影响**

**Table 4　Effect of modeling unit formed by BPE and BPE-dropout on recognition performance (WER)**

| 建模单元 | WER/% | |
|---|---|---|
| | THCHS30 | ST-CMDS |
| BPE | 8.24 | 7.63 |
| BPE-dropout(0.001) | 6.26 | 6.88 |
| BPE-dropout(0.01) | 12.51 | 10.45 |

综上所述，针对中文语音识别中经常出现未登录词的问题，本文提出了一种基于带阈值的 BPE-dropout 多任务学习的端到端语音识别模型. 该模型采用带随机性字节对编码算法，在形成子词时引入了带字数阈值的策略，将子词和字作为建模单元进行多任务训练和解码，其中字任务作为辅助任务. 编码器部分采用了 Conformer 结构，与链接时序分类和注意力机制相结合. 为进一步提升模型性能，引入了动态参数对损失函数进行动态调节，并同时进行多任务训练和解码. 测试结果表明，本文方法能有效解决未登录词问题，提升了语音识别性能，并指出了丢弃率对识别性能的影响.

## 参　考　文　献

[1] ALDARMAKI H, ULLAH A, RAM S, et al. Unsupervised Automatic Speech Recognition: A Review [J]. Speech Communication, 2022, 139: 76-91.

[2] JUNG H Y, LEE S Y. On the Temporal Decorrelation of Feature Parameters for Noise-Robust Speech Recognition [J]. IEEE Transactions on Speech & Audio Processing, 2000, 8(4): 407-416.

[3] YOUNG S. The HTK Hidden Markov Model Toolkit: Design and Philosophy [R]. Cambidge: Cambridge

University, 1993.
[4] GEORGESCU A L, CUCU H, BURILEANU C. Kaldi-Based DNN Architectures for Speech Recognition in Romanian [C]//2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). Piscataway, NJ: IEEE, 2019: 1-6.
[5] SAMIN A M, KOBIR M H, KIBRIA S, et al. Deep Learning Based Large Vocabulary Continuous Speech Recognition of an Under-Resourced Language Bangladeshi Bangla [J]. Acoustical Science and Technology, 2021, 42(5): 252-260.
[6] YALTA N, WATANABE S, HORI T, et al. CNN-Based Multichannel End-to-End Speech Recognition for Everyday Home Environments [C]//2019 27th European Signal Processing Conference (EUSIPCO). Piscataway, NJ: IEEE, 2019: 1-5.
[7] CHAN W, LANE I. Deep Convolutional Neural Networks for Acoustic Modeling in Low Resource Languages [C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2015: 2056-2060.
[8] DAI Y J, WU Z W. Mobile-Assisted Pronunciation Learning with Feedback from Peers and/or Automatic Speech Recognition: A Mixed-Methods Study [J]. Computer Assisted Language Learning, 2023, 36(5/6): 861-884.
[9] GRAVES A, MOHAMED A R, HINTON G. Speech Recognition with Deep Recurrent Neural Networks [C]// IEEE International Conference on Acoustics. Piscataway, NJ: IEEE, 2013: 6645-6649.
[10] MIAO H R, CHENG G F, GAO C F, et al. Transformer-Based Online CTC/Attention End-to-End Speech Recognition Architecture [C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2020: 6084-6088.
[11] ZHANG Y, PUVVADA K C, LAVRUKHIN V, et al. Conformer-Based Target-Speaker Automatic Speech Recognition for Single-Channel Audio [C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2023: 1-5.
[12] LI J L, WANG H X, HE P S, et al. Long-Term Variable Q Transform: A Novel Time-Frequency Transform Algorithm for Synthetic Speech Detection [J]. Digital Signal Processing, 2022, 120: 103256-103267.
[13] GAO T, PAN Q, ZHOU J, et al. A Novel Attention-Guided Generative Adversarial Network for Whisper-to-Normal Speech Conversion [J]. Cognitive Computation, 2023, 15: 773-792.
[14] ZHANG Z Y, ZHOU P. End-to-End Contextual ASR Based on Posterior Distribution Adaptation for Hybrid CTC/Attention System [EB/OL]. (2022-02-18)[2023-01-25]. https://arxiv.org/abs/2202.09003.
[15] KIM S, HORI T, WATANABE S. Joint CTC-Attention Based End-to-End Speech Recognition Using Multi-task Learning [C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2017: 4835-4839.
[16] AMOOLYA G, HANS A S A, LAKKAVALLI V R, et al. Automatic Speech Recognition for Tulu Language Using GMM-HMM and DNN-HMM Techniques [C]//2022 International Conference on Advanced Computing Technologies and Applications (ICACTA). Piscataway, NJ: IEEE, 2022: 1-6.
[17] ZHAO J, LI X, LIU W H, et al. DNN-HMM Based Acoustic Model for Continuous Pig Cough Sound Recognition [J]. International Journal of Agricultural and Biological Engineering, 2020, 13(3): 186-193.
[18] LI T X. Study on a CNN-HMM Approach for Audio-Based Musical Chord Recognition [J]. Journal of Physics: Conference Series, 2021, 1802(3): 032033-1-032033-8.
[19] ZHANG X Y, WANG J P, JIN L W, et al. CMT-Co: Contrastive Learning with Character Movement Task for Handwritten Text Recognition [C]//Proceedings of the Asian Conference on Computer Vision. Berlin: Springer, 2022: 3104-3120.
[20] GUO H Z, CHEN Y S, XIE X K, et al. Efficient Conformer-Based CTC Model for Intelligent Cockpit Speech Recognition [C]//2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). Piscataway, NJ: IEEE, 2022: 522-526.

(责任编辑: 韩 啸)