基于多特征迁移学习的低资源临高方言语音识别方法

王忠[1,2]，曹春杰[3]，谢夏[4]，穆罕默德·艾哈迈德·拉扎[3]，陈勇青[2]，陈昱珏[5]
（1.海南大学信息与通信工程学院，海南 海口 570228；2.海南经贸职业技术学院信息技术学院，海南 海口 571127；
3.海南大学网络空间安全学院（密码学院），海南 海口 570228；4.海南大学计算机科学与技术学院，海南 海口 570228；
5.北京原点临近科技有限责任公司，北京 100074）

摘　要：针对低资源临高方言语音识别中数据稀缺、字错误率高的问题，提出了一种基于多特征迁移学习的端到端语音识别方法。以TeleSpeech-ASR1.0-large多方言预训练模型为基座，融合梅尔频率倒谱系数、滤波器组能量系数与对数梅尔谱3类互补声学特征，通过构建Conformer-LAS-CTC联合优化架构，利用深度可分离卷积和多头自注意力机制分别捕捉语音信号的局部特征与全局依赖关系，并设计融合CTC、中间层CTC与注意力机制的多任务损失函数进行联合训练。在总时长为280 h的临高方言与普通话混合语料上的实验结果表明，所提方法的字错误率降低至18.89%，显著优于基线模型，有效缓解了低资源方言面临的数据瓶颈问题，为濒危语言的数字化保护提供了可行的技术路径。

关键词：低资源语音识别；迁移学习；Conformer；多特征融合；临高方言
中图分类号：TP391.4
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025196

Low-resource Lingao dialect speech recognition method based on multi-feature transfer learning

WANG Zhong[1,2], CAO Chunjie[3], XIE Xia[4], Muhammad Ahmad Raza[3], CHEN Yongqing[2], CHEN Yujue[5]
1. School of Information and Communication Engineering, Hainan University, Haikou 570228, China
2. School of Information Technology, Hainan College of Economics and Business, Haikou 571127, China
3. School of Cyberspace Security (School of Cryptography), Hainan University, Haikou 570228, China
4. School of Computer Science and Technology, Hainan University, Haikou 570228, China
5. Beijing OriginNear Technology Co., Ltd., Beijing 100074, China

Abstract: To address the challenges of data scarcity and high character error rate in low-resource Lingao dialect automatic speech recognition, an end-to-end speech recognition method based on multi-feature transfer learning was proposed. Using TeleSpeech-ASR1.0-large multi-dialect pre-trained model as the base model, three types of complementary acoustic features—Mel-frequency cepstral coefficients, filter bank energy coefficients, and log-Mel spectrograms were fused. A Conformer-LAS-CTC joint optimization architecture was constructed, employing depthwise separable convolutions and multi-head self-attention mechanisms to capture local features and global dependencies of speech signals, respectively. A multi-task loss function integrating connectionist temporal classification, intermediate CTC, and attention mechanisms was designed for joint training. Experimental results on a 280 hour mixed corpus of Lingao dialect and Putonghua show that the proposed method reduces the character error rate to 18.89%, significantly outperforming baseline models. This method effectively alleviates the data bottleneck faced by low-resource dialects and provides a viable technical pathway for the digital preservation of endangered languages.

Keywords:low-resource automatic speech recognition, transfer learning, Conformer, multimodal feature fusion, Lingao dialect

收稿日期：2025–08–31；**修回日期：**2025–10–05
通信作者：曹春杰，caochunjie@hainanu.edu.cn
基金项目：国家自然科学基金资助项目（No.U24A20238）；国家重点研发计划基金资助项目（No.2021YFB2700600）
Foundation Items：The National Natural Science Foundation of China (No.U24A20238), The National Key Research and Development Program of China (No.2021YFB2700600)

0 引言

方言作为语言多样性的重要组成部分，承载着丰富的地域文化信息。然而，许多方言，特别是使用人口稀少的低资源方言，如中国海南的临高话，正面临濒危风险。利用自动语音识别（ASR，automatic speech recognition）技术进行濒危方言的数字化保护与传承，已成为语言资源保护领域的重要研究方向[1-3]。作为一种典型的低资源方言，临高话由于缺乏大规模标注数据，采用传统神经网络训练出的模型字错误率（CER，character error rate）居高不下，难以满足实际应用需要。

针对低资源方言自动语音识别问题，迁移学习[4-5]目前是一种主流的解决方案，其核心思想是将已在一个任务中学习到的知识迁移应用到另一个相关但不同的新任务中，从而提升新任务的学习效率和性能[6-9]。然而，现有基于迁移学习的方法在应用于临高话等极端低资源方言场景时，仍存在3个方面的显著局限。①特征表征单一化：多数研究仅依赖单一的声学特征如梅尔频率倒谱系数（MFCC，Mel-frequency cepstral coefficient）或滤波器组能量系数（FBANK，filter-bank energy coefficient），未能充分挖掘不同特征在频谱细节、时序动态特性及噪声鲁棒性等方面的互补潜力，限制了模型对临高话复杂声学特性的全面表征能力；②模型结构适配性不足：标准预训练模型如Transformer[10]或Conformer[11]虽在通用任务上表现优异，但其固有的注意力机制与损失函数设计，并未针对低资源方言中普遍存在的发音变异、数据稀疏等特有挑战进行专门优化，导致领域适配能力欠佳；③损失函数引导性弱：在低资源条件下，传统的单一损失函数如时序分类（CTC，connectionist temporal classification）难以在序列对齐精度与上下文语义建模之间取得平衡，容易导致模型训练不稳定且泛化能力较差。因此，亟须研发一种能够有效融合多种特征的迁移学习方法，以提升临高话等低资源方言的语音识别性能。

为解决上述挑战，本文提出了一种融合多特征迁移学习的语音识别方法，其主要特点如下。

1) 多特征融合增强表征。融合MFCC、FBANK和对数梅尔频谱（Log-Mel，log-Mel spectrogram）3种互补声学特征。通过设计特征融合模块，整合不同特征在频谱细节、时序动态性和噪声鲁棒性上的优势，构建更全面、鲁棒的临高话声学表征，有效缓解单一特征表征不足的问题。

2) 联合优化模型架构。基于Conformer编码器的混合架构，充分利用卷积神经网络（CNN）的局部特征提取能力与自注意力机制的全局依赖建模优势，同时，在解码端采用LAS（listen, attend and spell）模型进行序列生成，并引入CTC辅助分支。通过LAS-CTC-InterCTC联合损失函数进行端到端联合优化，同时约束序列对齐精度（如CTC、InterCTC）、上下文语义建模（如LAS）及中间层表示一致性（如InterCTC），提升模型在低资源条件下的训练稳定性和识别精度。

3) 迁移学习策略适配。基于大规模多方言语音识别预训练模型，针对临高话数据稀缺的特点，实施分层冻结微调策略，在保留预训练知识的同时，高效适配临高话特有的声学和语言学特性，实现知识的高效迁移与利用。

1 相关工作

1.1 低资源语音识别研究

低资源语音识别旨在解决标注数据稀缺场景下的模型性能瓶颈问题。现有研究主要从数据增强、迁移学习等方面展开。数据增强方法通过合成语音（如SpecAugment[12]）或跨语言数据迁移扩充训练数据集，但合成数据与真实方言的声学特征差异可能导致模型性能退化。迁移学习则利用高资源语言（如普通话）的预训练模型（如Wav2Vec 2.0[13]、HuBERT[14]）进行知识迁移。例如，Li等[15]通过冻结预训练模型底层参数并微调顶层，在藏语语音识别任务中将CER降低12.7%。然而，此类方法在极端低资源场景（如临高话）中仍面临领域适配不足和过拟合风险，尤其当源语言与目标方言的音系差异显著时。近期研究尝试结合多任务学习（如联合训练语音识别与音素识别[16]）提升泛化性，但如何设计适配方言特性的迁移策略仍是开放问题。

1.2 多特征融合的应用

传统语音识别系统多依赖单一声学特征（如MFCC或FBANK），多特征融合通过整合互补信息提升模型鲁棒性，根据融合阶段划分通常分为3类。

早期融合（特征级）：直接拼接不同特征（如MFCC+语谱图+韵律特征）作为输入。Wang等[17]在汉语方言识别中融合MFCC与Gammatone特征，

使CER降低8.3%，但高维特征易引入噪声且计算开销大。

晚期融合（决策级）：训练多个单特征模型后加权融合输出。该方法在低资源场景下因各子模型性能差异显著而效果受限[18]。

混合融合（中间层）：在模型中间层整合特征表示。Zhang等[19]在Conformer编码器中注入文本特征，通过跨模态注意力机制提升低资源语音识别性能。然而，现有研究对方言特异性特征（如临高话的喉塞音、元音松紧对立）的融合探索不足，且缺乏对特征间冗余性的有效抑制机制。

1.3　联合损失函数

语音识别模型训练中损失函数的设计直接影响模型优化方向，主流损失函数包括以下几种。

1) CTC损失：通过强制对齐实现快速训练，但存在“条件独立性假设”限制，难以建模长时间依赖关系[20]。

2) LAS注意力损失：基于编解码器架构，可捕捉全局上下文，但在低资源场景下易因注意力偏移导致错误累积[21]。

3) Transducer损失：结合CTC与LAS优势，但计算复杂度高，难以应用于实时系统[22]。

为兼顾训练效率与建模能力，联合损失函数成为研究热点，近年来的研究成果主要体现在以下几点。

1) CTC-Attention联合损失：通过加权融合（如$\lambda\mathcal{L}_{\text{ctc}}+(1-\lambda)\mathcal{L}_{\text{att}}$）优化模型。Kim等[23]在韩语语音识别任务中通过实验验证，证明该方法在低资源条件下能够有效提升模型收敛速度，但权重λ需人工调参，且两种损失的梯度冲突可能影响稳定性。

2) 多任务联合损失：引入辅助任务（如音素识别、语言模型预训练）共享编码器。例如，Chen等[24]在越南语语音识别中联合优化CTC与音素分类损失，使CER降低9.1%。然而，辅助任务与主任务的目标不一致性可能导致负迁移。

3) 层级化损失设计：InterCTC[25]在编码器中间层插入CTC监督，缓解深层网络梯度消失问题，但该方法在超低资源场景下因中间层监督信号不足而导致实际效果受限。

现有工作在低资源ASR中仍存在3个方面的局限性：在迁移学习方面，缺乏针对方言特性的特征适配策略；在多特征融合方面，未能充分挖掘方言各种声学特征的协同效应；在联合损失设计方面，未能解决多任务梯度冲突与权重自适应问题。本文提出的多特征融合迁移学习框架与LAS-CTC-Inter-CTC联合损失旨在弥合上述缺口，为低资源方言语音识别提供新范式。

2　方法

2.1　模型架构

本文提出的基于多特征迁移学习的端到端语音识别模型的整体架构如图1所示。该模型以TeleSpeech-ASR1.0-large[26]多方言预训练模型为基座，通过Conformer-LAS-CTC联合优化框架，结合方言与普通话数据混合、多特征融合与增强等技术，有效降低模型的CER。具体而言，模型前端首先采用多尺度时频掩蔽模块（MSTFM，multi-scale time-frequency masking）[27]对输入语音信号进行增强以抑制环境噪声，随后并行提取40维MFCC、80维FBANK及80维Log-Mel这3类互补声学特征，并将其拼接为200维特征向量，再通过可学习的线性投影层映射为512维统一表征，作为编码器输入。编码器采用12层Conformer Block堆叠结构，充分利用深度可分离卷积捕捉局部特征与多头自注意力机制建模全局依赖，以增强对复杂方言音素的表征能力；解码器则联合LAS与CTC机制，其中LAS基于双向LSTM与加性注意力动态生成字符序列，CTC提供帧级对齐监督。为进一步优化训练过程，模型引入多任务联合损失函数，加权融合CTC损失、中间层CTC损失及LAS注意力损失，以平衡序列对齐精度与语义建模效率，加速低资源场景下的模型收敛。整个架构通过层次化迁移学习策略，在保留预训练通用声学知识的同时，微调顶层参数以适配临高话特有的语言学特征，从而显著提升识别性能与鲁棒性。

2.1.1　基座模型

本文选用TeleSpeech-ASR1.0-large作为基座模型，该模型基于改进的data2vec架构[28]，在多种方言混合场景中展现出优异的识别性能与无缝切换能力，为低资源临高方言语音识别提供了理想的迁移学习起点。在具体实现中，本文直接加载其预训练参数作为模型初始化，保留了其在多方言数据上学到的通用声学知识。在此基础上，通过分层迁移策略对模型进行适配：冻结底层编码器参数以保持通

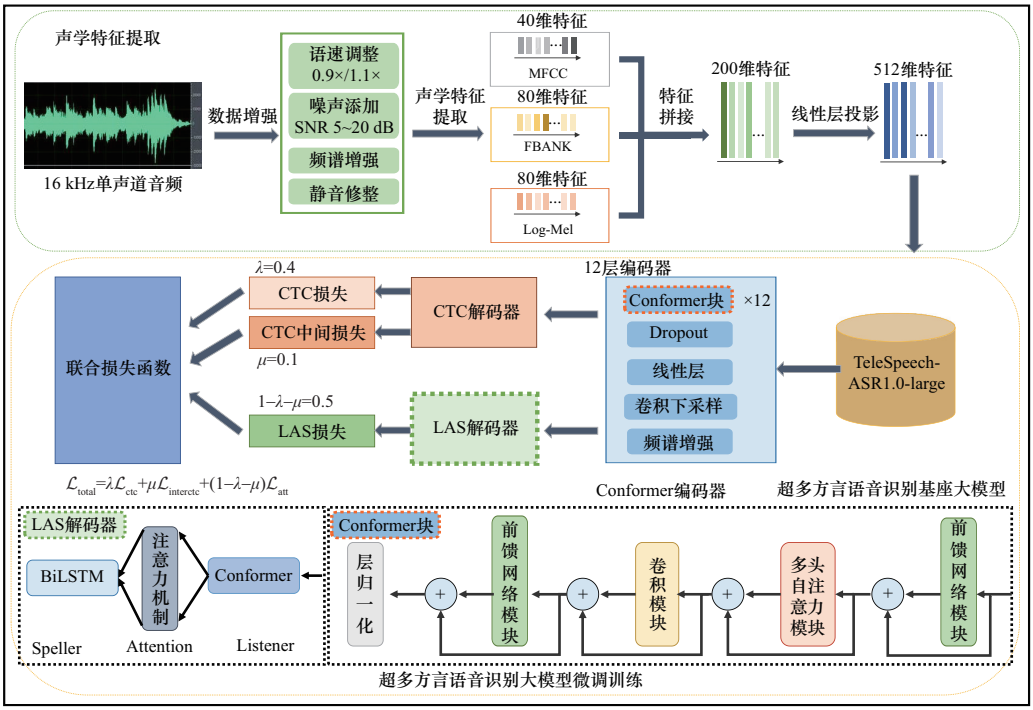

图1　模型架构

用声学特征的稳定性，同时微调顶层模块及新增的中间CTC分支，使其能够专门学习临高方言特有的音素模式和声学特性。这种基于预训练模型的迁移方式有效利用了大规模无标注数据中的语音先验知识，缓解了低资源场景下的数据稀疏问题，为后续的多特征融合和联合优化提供了强有力的声学建模基础。

2.1.2　编码器

编码器基于Conformer架构（如图2所示），由12个相同的Conformer块堆叠而成，替代标准Transformer[29-30]进行深度特征抽象。每个块包含4个核心组件：前馈网络（FFN，feed-forward network）模块、多头自注意力（MHSA，multi-head self-attention）模块、卷积（Conv）模块以及层归一化（LayerNorm）。具体计算流程如下：输入信号首先经过第一个FFN模块，采用门控线性单元（GLU，gated linear unit）激活并实施半步残差连接；随后进入MHSA模块，该模块集成Transformer-XL[31]的相对正弦位置编码技术，通过相对偏移矩阵动态建模序列依赖关系；接着通过卷积模块，该模块依次执行通道扩展的逐点卷积、GLU激活、一维深度卷积和Swish激活函数，专门用于提取局部时频特征；最后经过第二个FFN和层归一化输出。整个数据处理流程严格遵循式(1)所示的层级计算关系，其中每个模块之间均采用残差连接确保梯度有效传播。特别地，MHSA模块采用相对位置编码替代绝对位置编码，直接建模音素间的相对距离；卷积模块采用深度可分离卷积结构，在降低参数量的同时增强对短时局部特征的感知能力；FFN模块中引入双线性注意力机制（Bilinear Attention）[32-33]，通

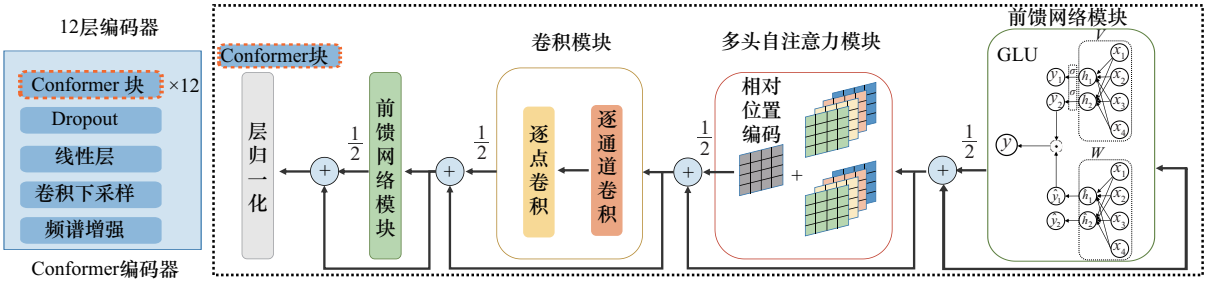

图2　Conformer架构

过构造特征间的高阶交互来增强对声调变体和辅音簇的建模能力。这种异构架构设计使得编码器能够同时捕捉语音信号的全局依赖关系和局部精细特征，为低资源方言识别提供了强有力的特征提取基础。

对于给定的第 i 个 Conformer 块输入 $\boldsymbol{x}_i$，输出 $\boldsymbol{h}_i$ 的计算式为

$$\begin{cases} \bar{\boldsymbol{x}}_i = \boldsymbol{x}_i + \dfrac{1}{2}\mathrm{FFN}(\boldsymbol{x}_i) \\ \boldsymbol{x}'_i = \bar{\boldsymbol{x}}_i + \mathrm{MHSA}(\bar{\boldsymbol{x}}_i) \\ \boldsymbol{x}''_i = \boldsymbol{x}'_i + \mathrm{Conv}(\boldsymbol{x}'_i) \\ \boldsymbol{h}_i = \mathrm{LayerNorm}\left(\boldsymbol{x}''_i + \dfrac{1}{2}\mathrm{FFN}(\boldsymbol{x}''_i)\right) \end{cases} \tag{1}$$

其中，FFN 表示前馈网络模块，MHSA 表示多头自注意力模块，Conv 表示卷积模块，LayerNorm 表示层归一化，每个模块间都使用残差连接。

2.1.3　解码器

解码器采用 LAS[34] 模型的 Attention 和 Speller 部分与 CTC 模型联合解码，旨在将编码器输出的声学特征序列转化为目标字符序列。LAS 模型负责基于上下文感知的序列生成，CTC 模型则提供帧级对齐监督，二者通过多任务损失函数实现联合优化。

1) LAS 模型

LAS 模型基于双向长短期记忆网络（BiLSTM）并结合注意力机制构建，主要由编码器（Listener）、注意力机制（Attender）与解码器（Speller）3 个组件构成，通过联合优化实现端到端的序列生成。

编码器（Listener）：编码器（Listener）将输入的声学特征序列通过 Conformer 编码器转换为高维隐状态序列，设输入声学特征序列为 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$，其中 $\boldsymbol{X}_t \in \mathbb{R}^d$（$d$ 为特征维度）为第 t 帧的特征，Conformer 通过 MHSA 与卷积模块提取全局与局部特征

$$\boldsymbol{H}' = \mathrm{MHSA}(\boldsymbol{X}) + \boldsymbol{X}, \quad \boldsymbol{H}'' = \mathrm{Conv}(\boldsymbol{H}') + \boldsymbol{H}' \tag{2}$$

最终经 FFN 与层归一化输出高维隐状态序列 $\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_T)$

$$\boldsymbol{h}_t = \mathrm{LayerNorm}\left(\mathrm{FFN}(\boldsymbol{H}'') + \boldsymbol{H}''\right), \quad \boldsymbol{h}_t \in \mathbb{R}^{512} \tag{3}$$

注意力机制（Attention）：解码器在生成第 i 个字符 y_i 时，通过加性注意力动态对齐编码器输出

$$\boldsymbol{e}_{it} = \boldsymbol{v}^{\mathrm{T}} \tanh\left(\boldsymbol{W}_s \boldsymbol{s}_{i-1} + \boldsymbol{W}_h \boldsymbol{h}_t + \boldsymbol{b}\right) \tag{4}$$

其中，$\boldsymbol{s}_{i-1} \in \mathbb{R}^{1024}$ 为解码器上一时刻隐状态（双向 LSTM 拼接），$\boldsymbol{v}$、$\boldsymbol{W}_s$、$\boldsymbol{W}_h$、$\boldsymbol{b}$ 为可学习参数。注意力权重 α_{it} 及上下文向量 $\boldsymbol{c}_i$ 计算式为

$$a_{it} = \frac{\exp\left(\boldsymbol{e}_{it}\right)}{\sum_{t'=1}^{T}\exp\left(\boldsymbol{e}_{it'}\right)}, \boldsymbol{c}_i = \sum_{t=1}^{T} a_{it}\boldsymbol{h}_t \tag{5}$$

解码器（Speller）：由 2 层隐藏维度为 512 的双向 LSTM 构成，每层通过自注意力机制动态捕捉前后文依赖关系，输入为上一时刻字符嵌入 $\boldsymbol{e}_{i-1}\left(\boldsymbol{e}_{i-1} \in \mathbb{R}^{\epsilon}\right)$ 与上下文向量 $\boldsymbol{c}_i$ 的拼接，更新当前隐状态 $\boldsymbol{s}_i$

$$\boldsymbol{s}_i = \mathrm{BiLSTM}_{\mathrm{dec}}^{(2\times 512)}\left(\left[\boldsymbol{e}_{i-1};\boldsymbol{c}_i\right],\boldsymbol{s}_{i-1}\right), \boldsymbol{s}_i \in \mathbb{R}^{1024} \tag{6}$$

最终通过线性层和 Softmax 预测字符概率分布

$$P(y_i \mid y_{<i}, \boldsymbol{X}) = \mathrm{Softmax}\left(\boldsymbol{W}_o \boldsymbol{s}_i + \boldsymbol{b}_o\right) \tag{7}$$

其中，$\boldsymbol{W}_o \in \mathbb{R}^{|V|\times 1024}$ 为输出权重矩阵，V 为词汇表大小。

2) CTC 模型

CTC 模型用于处理未分段序列的分类，自动对齐语音帧序列与文字序列。在训练过程中，CTC 模型通过符号合并规则实现高效对齐，通过引入空白符号 ϵ，定义所有可能的对齐路径集合。

设输入语音特征序列为 $X=[x_1,x_2,\cdots,x_T]$，神经网络输出标签序列为 $Y=[y_1,y_2,\cdots,y_U]$，初始解码路径为 $\pi=\{\pi_1,\pi_2,\cdots,\pi_T\}$。路径概率定义为

$$p(\pi \mid \boldsymbol{X}) = \prod_{t=1}^{T} p\left(\pi_t \mid \boldsymbol{x}_t\right) \tag{8}$$

因不同路径可能对应相同解码结果，需对路径概率求和得到标注序列 l 的后验概率

$$p(l \mid \boldsymbol{X}) = \sum_{\pi \in \Phi(l)} p(\pi \mid \boldsymbol{X}) \tag{9}$$

其中，$\Phi(l)$ 为所有可能解码路径的集合。引入空白符号后，标注长度变 $|l'| = 2N+1$，N 为原始符号数。前向概率 $\alpha(t,n)$ 和后向概率 $\beta(t,n)$ 分别通过路径集合 $F(t,n)$ 和 $B(t,n)$ 计算

$$\alpha(t,n) = \sum_{\pi \in F(t,n)} \prod_{i=1}^{t} p\left(\pi_i \mid \boldsymbol{x}_i\right) \tag{10}$$

$$\beta(t,n) = \sum_{\pi \in B(t,n)} \prod_{i=t}^{T} p\left(\pi_i \mid \boldsymbol{x}_i\right) \tag{11}$$

最终后验概率为

$$p(l|\boldsymbol{X}) = \sum_{n=1}^{l'} \alpha(t,n)\beta(t,n) \tag{12}$$

训练损失函数采用负对数似然

$$\mathcal{L}(S) = -\sum_{(\boldsymbol{x},l)\in S} \log p(l|\boldsymbol{x}) \tag{13}$$

2.2　多任务联合优化

模型采用多任务联合优化策略，通过融合CTC损失、中间层CTC（InterCTC）损失及LAS注意力损失，提升模型收敛速度与识别精度。多任务联合优化总损失函数定义为

$$\mathcal{L}_{\text{total}} = \lambda\mathcal{L}_{\text{ctc}} + \mu\mathcal{L}_{\text{interctc}} + (1-\lambda-\mu)\mathcal{L}_{\text{att}} \tag{14}$$

其中，$\lambda = 0.4$为CTC损失权重，$\mu = 0.1$为中间层CTC损失权重，$(1-\lambda-\mu) = 0.5$为注意力损失权重。

CTC损失：优化音频帧序列$\boldsymbol{X} = [\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T]$与目标文本序列$l$的单调对齐，解决方言中音素与文本的非严格时序匹配问题，通过自动对齐机制避免人工标注时间边界，加速训练收敛，损失函数为

$$\mathcal{L}_{\text{ctc}} = -\log p(l|\boldsymbol{X}) \tag{15}$$

中间层CTC（InterCTC）损失：在编码器第6层插入辅助监督分支，对中间特征$\boldsymbol{h} = [\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_T]$进行音素级约束

$$\mathcal{L}_{\text{interctc}} = -\log p(l|\boldsymbol{h}) \tag{16}$$

中间层CTC（InterCTC）损失以音素级标签监督底层声学特征学习，通过低层特征规范化，增强模型对复杂音素的表征能力，其权重设为0.1以平衡对底层特征的约束强度。

LAS注意力损失：基于解码器双向LSTM与注意力机制，动态捕捉长距离依赖关系，优化目标字符序列$\boldsymbol{Y} = [\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_U]$的生成精度

$$\mathcal{L}_{\text{att}} = -\sum_{i=1}^{U} \log p(y_i|y_{<i},\boldsymbol{X}) \tag{17}$$

联合优化过程中，主CTC损失主导训练初期的对齐效率，注意力损失在训练后期精细化解码过程，中间CTC损失则贯穿全程稳定底层特征分布，形成层次化的训练目标。多任务联合损失函数3个部分通过平衡序列对齐精度与序列建模效率，实现模型快速收敛。

2.3　多特征融合与训练策略

在完成噪声消除等信号预处理后，需要从语音片段中解析关键声学特征，将原始语音转化为机器可处理的特征向量，使声音特征适配人类的听觉感知，本文中提取的特征有以下3类。

1) FBANK：设功率谱为$P(k)$，梅尔滤波器组传递函数为$H_m(k)$ $(m = 1,2,\cdots,M)$，则其计算式为

$$\text{FBANK}[m] = \ln\left(\sum_{k=0}^{N-1} P(k)\cdot H_m(k)\right) \tag{18}$$

其中，M为滤波器数量（如80维），N为FFT点数。

2) Log-Mel：其完整保留了梅尔频域的局部细节和能量分布，特征维度更高（通常为40~128维），计算式为

$$\text{Log-Mel}[m] = \ln\left(\sum_{k=0}^{N-1} P(k)\cdot H_m(k)\right) = \text{FBank}[m] \tag{19}$$

3) MFCC：通过模拟人耳非线性听觉感知提取语音信号关键特征，将频谱转换至倒谱域，有效压缩数据并突出声道谱包络，抑制激励源与噪声干扰。其计算式为

$$\text{MFCC}[n] = \sum_{m=1}^{M} S[m]\cdot\cos\left(\frac{\pi n(m-0.5)}{M}\right),\quad n = 0,1,\cdots,L-1 \tag{20}$$

其中，$S[m]$为对数梅尔谱，L为保留系数（如静态13维+Δ13维+ΔΔ14维=40维）。

2.3.1　特征融合

为了提升模型在低资源方言场景下的鲁棒性与泛化能力，本文中采用多特征融合策略，整合3种互补声学特征：MFCC、FBANK与Log-Mel。

1) 特征并行提取

提取MFCC（40维）特征：通过Kaldi工具提取，包含静态系数（13维）+一阶差分（Δ,13维）+二阶差分（ΔΔ,14维）。经梅尔滤波器组滤波、对数能量计算及DCT压缩，突出声调差异，计算式为

$$\{\text{MFCC}_{\text{final}} = [\text{MFCC}_{\text{static}},\Delta\text{MFCC},\Delta\Delta\text{MFCC}] \in \mathbb{R}^{40} \tag{21}$$

提取FBANK（80维）特征：跳过DCT，直接输出梅尔滤波器组对数能量（静态特征），保留完整频谱分布，适配非线性音素（如爆破音、擦音），计算式为

$$\text{FBank} \in \mathbb{R}^{80},\quad \text{FBank}[m] = \ln\left(\sum_{k} P(k)\cdot\boldsymbol{H}_m(k)\right) \tag{22}$$

提取Log-Mel（80维）特征：在静态Log-Mel

（40 维）基础上拼接时域一阶/二阶动态差分特征（Δ/ΔΔ 各 20 维），增强频带不变性，计算式为

$$\text{Log-Mel}_{\text{final}} = [\text{Log-Mel}_{\text{static}}, \Delta\text{Log-Mel}, \Delta\Delta\text{Log-Mel}] \in \mathbb{R}^{80} \quad (23)$$

动态特征计算式为

$$\Delta x_t = \frac{\sum_{n=1}^{N} n(x_{t+n} - x_{t-1})}{2\sum_{n=1}^{N} n^2}, \quad \Delta\Delta x_t = \Delta(\Delta x_t) \quad (24)$$

其中，N为差分窗口宽度，本文取2。

2) 拼接与投影

将 3 类特征拼接为 200 维向量，输入可学习的线性投影层，映射至 512 维统一表征

$$\boldsymbol{z} = \boldsymbol{W} \cdot [\text{MFCC}_{\text{final}}; \text{Log-Mel}_{\text{final}}] + \boldsymbol{b}, \quad \boldsymbol{W} \in \mathbb{R}^{512 \times 200}, \boldsymbol{b} \in \mathbb{R}^{512} \quad (25)$$

其中，$\boldsymbol{z}$为投影后的特征向量。该方法通过 3 个子空间互补，覆盖临高方言声学特性的长尾分布。MFCC 子空间通过梅尔滤波器组与离散余弦变换压缩频域冗余信息，突出声调差异；FBANK 子空间保留原始滤波器组能量分布，适配非线性较强的方言音素；Log-Mel 子空间结合时域一阶与二阶动态差分特征，抑制环境噪声干扰。3 类特征经拼接后形成 200 维向量，通过可学习的线性投影层映射为 512 维统一表征，作为编码器输入。特征融合后可同时捕捉全局包络、局部频谱结构及感知优化能量分布，减少单一特征对噪声或数据偏差的敏感性，最大化声学信息利用率。

2.3.2　模型训练

模型训练采用分层迁移学习与动态优化策略：首先加载预训练参数，冻结底层 Conformer 模块（前 6 层）以保留跨语言声学特征，微调顶层参数及中间 CTC 分支（第 6 层编码器后）以适配临高方言特有音素。解码器由 2 层双向 LSTM 构成，隐藏层维度 512，随机初始化后通过 LAS 注意力机制动态对齐编码器输出。

训练基于 Fairseq 框架实现数据并行，全局批次大小固定为 64 以确保计算效率与显存占用的平衡，启用混合精度训练（FP16）以加速计算。优化器选用 Adam（初始学习率为 1×10^{-3}），结合余弦退火学习率调度，前 25 000 步实施线性热身稳定参数初始化。在正则化策略方面：编码器与解码器分别施加 0.1 和 0.3 的 Dropout 比率。训练共进行 50 轮次，每 500 步记录训练损失曲线，每轮次在验证集上评估 CER 指标。

3　实验

3.1　实验数据

3.1.1　数据来源

1) 语料文本。为适应不同应用需求并契合临高语发音特征，本研究设计了多元化的方言语料。鉴于临高方言仅有口头形式而无书写系统，且缺乏其他权威方言词典，语料文本选取主要依据刘剑三的《临高语话语材料集》与《临高汉词典》。常用字词语料主要源于此两部著作。日常用语则参考梁德慧《HSK 汉语水平考试（初、中等）》，并依据临高方言词汇特点，增补了高频生活用语及称谓等内容。最终语料涵盖 7 个领域：临高地名、常用词汇、常用句式、日常会话、本地新闻、小说故事、特色俚语，具体比例如表 1 所示。

表 1　　文本语料占比

领域	占比
临高地名	5%
常用词语	20%
常用句子	20%
日常对话	20%
临高新闻	20%
小说故事	10%
特色俚语	5%

2) 语音采集。语音录制在无噪声、无混响的室内环境中进行，使用麦克风采集，发音人口距麦克风约 10 cm。录音参数设置为：采样率 16 kHz，量化位数 16 bit，语音幅值范围控制在 3 000~20 000，信噪比大于 20 dB，保存格式为单声道 WAV 格式。录制方式包含发音人按正常语速朗读文本，或多人自由交谈。发音人覆盖临高县主要方言点（临城、新盈、调楼、和舍、波莲、南宝），年龄划分为 A（16~25 岁）、B（26~40 岁）、C（41 岁以上）3 组，共 18 人（男女各 9 名）。录音完成后，使用 Cool Edit 专业软件进行降噪处理。为便于管理和检索，语音文件按发音人分类存储，相关的发音人信息、文本内容及标注等元数据存入关系型数据库。针对时长超过 30 s 的音频，通过切分处理生成多个独立语句片段。

3) 数据融合。通过社区调查与实地录音采样，

系统性地收集高质量临高话语音数据达 200 h。由于普通话与临高话同属汉藏语系，共享部分音素（如爆破音/b/、/p/）和声调特征，为提升方言与通用语音特征的兼容性，同步引入公开普通话数据集 AIShell-1（80 h），数据总规模达 280 h，涵盖语音样本约 300 000 条。该数据集的设计兼顾方言稀缺性与语音普适性：临高话数据覆盖生活化场景（如市集对话、家庭交流），而普通话新闻语料补充了标准发音与快语速样本，为多任务训练的声学建模奠定基础。通过严格的质量控制（剔除信噪比<15 dB 样本）与文本标准化（保留高频方言词汇），构建了适配低资源方言识别的基准语料库。

3.1.2 数据清洗

针对原始语音与文本数据，设计多级清洗流程以提升数据质量。针对音频数据清洗，采用动态时间规整（DTW，dynamic time warping）算法等自动化工具对齐音频与标注文本，剔除时长偏差超过 5% 的无效样本；通过能量阈值检测（平均能量< –30 dBFS）与信噪比（SNR<15 dB）筛选，移除低质量音频；统一标准化音频格式至 16 kHz 采样率、单声道 WAV，消除硬件采集差异；基于静音检测（VAD 算法）切分长音频，静音段持续≥0.5 s 时分割为独立片段，并限制每段时长在 1~15 s，避免训练时的显存溢出。

针对文本数据清洗，清除文本中的换行符、冗余空格；剔除长度超过 80 字符或存在乱码的样本；移除<、>、[、]、~、/、\、=等特殊符号以及非中英文内容；基于临高方言词典（含 5 000 个高频词）规范标注文本，保留特有词汇（如“侬”表示“孩子”）并建立标准对照表；统计词频分布，对低频词（出现次数<5）进行替换或删除，减少模型过拟合风险。清洗后数据集包含 20 万条有效样本，错误标注率低于 0.8%，为模型训练提供高质量实验数据。

3.1.3 数据增强

为提升模型对复杂声学场景的泛化能力，设计多层次数据增强方案，相关增强操作均基于公开工具库实现，关键参数如下。

1) 语音变异增强：采用基音同步叠加相加（PSOLA，pitch synchronous overlap and add）算法对音频进行时长缩放（语速调整为 0.9 倍与 1.1 倍，时长变化±15%），以模拟方言发音的节奏与情绪特征，共生成 3.2 万条增强样本。

2) 噪声鲁棒性增强：引入 NOISEV2 语料库噪声干扰，以信噪比 5~20 dB 随机混合至原始语音，采用的 NOISEV2 语料库包含风声、街道噪声等多种环境噪声，风声与街道噪声按 1:1 比例混合，以模拟真实户外环境。

3) 频谱掩蔽增强：基于 SpecAugment 策略，在时频域对语音特征进行随机遮蔽。即在时间轴上随机掩蔽一段长度为 20~50 帧的连续语音块，以模拟时域上的信息中断；在频率维度上随机屏蔽 2~5 个连续的 Mel 频段，以模拟频域上的信息丢失。通过以上方法组合增强模型对语音信号中各种变化的适应性。

4) 静音优化：利用语音活动检测（VAD，voice activity detection）算法识别并修剪首尾静音段（阈值为–40 dB），保留有效语音内容，同时限制片段时长为 1~15 s，提升训练效率。经增强后，数据集规模扩展至 25 万条。

经上述处理后的数据，方言与普通话比例为 7:3。训练集占比为 80%（224 h）；验证集占比 10%（28 h）；测试集占比 10%（28 h）。

3.2 实验环境

实验硬件环境：CPU 为 Intel Xeon Gold 5218 @ 2.30 GHz ×64 核心，支持多线程并行计算；GPU 为 NVIDIA Tesla T4×4（16 GB 显存/卡），通过 NV-Link 互联实现数据并行加速；内存为 128 GB DDR4，保障大规模数据加载与特征处理。

软件环境：操作系统为 CentOS 7（64 位），内核版本 3.10.0；深度学习框架为 PyTorch 1.10.0 + CUDA 11.3，适配 GPU 算力；工具链为 WeNet 1.0（语音识别框架）、Kaldi（声学特征提取）、Fairseq（分布式训练支持）。

实验具体软硬件环境如表 2 所示。

表 2 实验环境

设备	配置
操作系统	Cent OS 7
内存	128 GB
CPU	Intel(R) Xeon(R) Gold 5218 CPU @ 2.30 GHz×64
GPU	NVIDIA Corporation TU104GL [Tesla T4]×4
Python	3.8.18
Torch	1.13.0

3.3 评价指标

实验采用 CER 作为模型评价指标，其定义为

$$\mathrm{CER}=\frac{S+D+I}{N}\times 100\% \tag{26}$$

其中，S为被替换的字数，D为被删除的字数，I为被插入的字符数，N表示总字数。

4　实验结果分析

4.1　模型收敛性对比分析

图3为不同语音识别模型在训练过程中的损失变化趋势。从图3可以看出，BiLSTM-LAS、Transformer-LAS、Conformer-LAS及TeleSpeech-ASR1.0-large这4种模型的损失值均随训练步数的增加呈下降趋势，但收敛速度与最终性能存在显著差异。TeleSpeech-ASR1.0-large模型损失下降最快且始终处于最低水平，表明其优化效率与模型拟合能力最优；Conformer-LAS次之，损失下降速率稳定且优于Transformer-LAS；Transformer-LAS损失曲线收敛速度中等；BiLSTM-LAS模型损失下降最缓慢且最终损失值最高，反映其参数学习效率较低。整体曲线对比显示，TeleSpeech-ASR1.0-large凭借分层迁移学习与多任务优化策略（如中间CTC分支辅助训练），在训练动态性上显著优于其他基线模型，验证了其架构设计对复杂方言声学特征的高效适配能力。

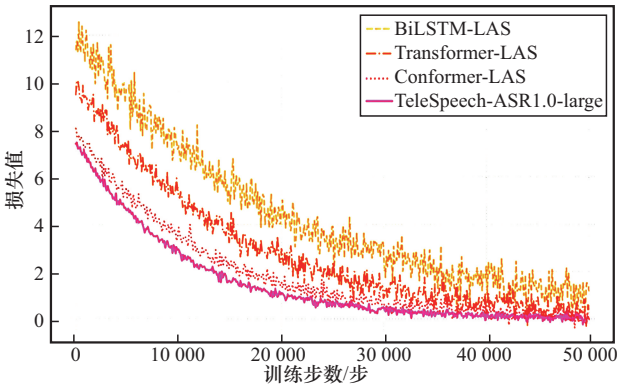

图3　训练过程损失

4.2　识别性能对比分析

为全面评估所提方法的有效性，表3对比了本文模型与多种基线模型在临高方言测试集上的性能。结果表明，本文提出的基于TeleSpeech-ASR1.0-large的融合模型的核心评价指标字错误率与词错误率（WER）均显著优于所有基线模型。具体而言，相较于BiLSTM-LAS，本文模型、Transformer-LAS以及未引入中间CTC损失的Conformer-LAS的CER相对降幅分别达到15.44%、6.58%和12.94%。这一结果充分证明了本文所采用的多特征融合、联合优化架构及迁移学习策略对于提升低资源方言语音识别性能的有效性与必要性。

表3　不同模型实验结果对比分析

模型	CER	WER	CER相对降幅
BiLSTM-LAS	22.34%	35.01%	—
Transformer-LAS	20.87%	31.78%	6.58%
Conformer-LAS（无中间CTC）	19.45%	28.92%	12.94%
TeleSpeech-ASR1.0-large	18.89%	25.62%	15.44%

4.3　消融实验分析

为验证各模块对模型性能的贡献，设计消融实验系统评估多特征融合、联合损失函数及迁移学习策略的有效性。实验结果如表4所示。

表4　消融实验分析

模型配置	CER	WER	CER相对增幅
完整模型	18.89%	25.62%	—
仅MFCC特征	20.45%	28.13%	8.26%
仅FBANK特征	20.12%	27.84%	6.51%
仅Log-Mel特征	19.87%	27.45%	5.19%
无中间CTC损失	19.45%	28.92%	2.96%
无迁移学习（从零训练）	23.67%	32.45%	25.30%

消融实验结果表明：多特征融合策略对性能提升贡献显著，单一特征输入时CER增加5.19%~8.26%，验证了3类特征的互补性；中间CTC损失的引入使CER增加2.96%，说明中间层监督有效提升了底层声学特征的建模能力；迁移学习策略贡献最大，从零训练模型CER达23.67%，相对完整模型增加25.30%，凸显了预训练知识迁移在低资源场景中的关键作用。

4.4　与现有自监督方法性能对比分析

为全面评估所提方法的有效性与先进性，将所提方法与近年来代表性的自监督低资源语音识别方法在相同测试集上进行性能对比。代表性方法包括在低资源任务中广泛使用的基于自监督预训练的通用语音识别模型Wav2Vec 2.0-Base和具备较强的跨语言泛化能力的隐单元预测自监督模型HuBERT-Large。对比结果如表5所示。

表5　与现有低资源语音识别方法的性能对比

方法	模型类型	是否预训练	CER	WER
Wav2Vec 2.0-Base	自监督	是	21.45%	29.83%
HuBERT-Large	自监督	是	20.67%	28.91%
所提方法	多特征+迁移学习	是	18.89%	25.62%

从表5可以看出，所提方法在CER与WER两项指标上均优于所有对比方法。与同为预训练模型的Wav2Vec 2.0和HuBERT相比，所提方法的CER分别下降11.93%和8.61%，表明基于多特征融合与迁移学习的LAS-CTC-InterCTC联合优化机制在方言声学建模中有一定优势。

4.5　噪声鲁棒性验证

表6为噪声环境下模型性能对比。为验证模型在复杂环境下的鲁棒性，针对低信噪比场景（10 dB）进行噪声测试，该噪声为风声与街道噪声按1:1比例混合的复合噪声。未添加噪声时，模型在临高话测试集上的CER为18.89%；引入风声、街道声等真实背景噪声后，CER显著上升至22.76%，增幅达20.49%。为提升抗噪能力，训练阶段采用多维度数据增强策略：通过卷积噪声扰动技术模拟声学环境失真，并结合SpecAugment中的时间遮蔽（随机掩蔽25%时间帧）与频率遮蔽（掩蔽2~5个频率段）强化特征鲁棒性。优化后，模型在相同10 dB噪声环境下的CER降至20.32%，相对未增强方案降幅达10.72%。测试结果表明，噪声增强技术有效抑制了环境干扰对低资源方言识别的影响，为实际应用场景提供可靠保障。

表6　噪声环境下模型性能对比

测试条件	CER
无噪声（原始模型）	18.89%
10 dB噪声（未增强）	22.76%
10 dB噪声（增强后）	20.32%

4.6　数据增强效果验证

表7为数据增强技术性能贡献分析，其中，无增强指未采用增强策略的TeleSpeech-ASR1.0-large模型。数据增强通过多技术协同显著提升模型泛化能力。实验对比显示：基线模型CER为25.41%；经语速调整（0.9倍、1.1倍动态变速）、噪声添加及SpecAugment（时间遮蔽25%时间帧+频率遮蔽2~5段）组合优化后，全增强策略使CER降至18.89%，相对基线降幅达25.66%；SpecAugment贡献最为突出，单独使用可使CER降低17.24%，其通过随机遮蔽频谱特征迫使模型学习局部不变性，有效缓解过拟合；语速调整与噪声添加分别带来6.41%和12.83%的降幅，二者结合可模拟真实场景中的语速变化与环境干扰，增强模型对动态输入的适应性。全增强策略的协同效应使模型在低资源方言识别中的鲁棒性提升25.66%，验证了数据增强在解决实际应用问题中的核心价值。

表7　数据增强技术性能贡献分析

增强策略	CER	CER相对降幅	技术说明
无增强	25.41%	—	—
语速调整	23.78%	6.41%	动态时间扭曲（0.9倍、1.1倍变速）
噪声添加	22.15%	12.83%	多类型环境噪声混合
SpecAugment	21.03%	17.24%	时间遮蔽（25%时间帧）+频率遮蔽（2~5个Mel频段）
全增强组合	18.89%	25.66%	上述技术协同优化

5　结束语

本文针对低资源海南临高方言语音识别中因数据匮乏导致的高字错误率问题，提出了一种融合多特征迁移学习的端到端语音识别方法。首先，构建了一个以TeleSpeech-ASR1.0-large多方言预训练模型为基座的分层迁移学习框架，有效利用了大规模无标注数据中学习到的通用声学知识；其次，设计了一种Conformer-LAS-CTC联合优化架构，通过CTC、InterCTC与LAS注意力多任务损失的协同作用，提升了模型在低资源条件下的收敛速度与识别精度；最后，提出了一种面向临高方言发音特点的多特征融合机制，综合MFCC、FBANK与Log-Mel这3类特征的互补优势，并结合多层次数据增强策略，显著增强了模型的鲁棒性与泛化能力。实验结果表明，所提方法显著降低了临高方言的识别错误率。未来，笔者将探索引入方言特定的音素建模单元，并尝试将本框架扩展至其他濒危汉语方言及少数民族语言，以进一步验证其普适性，为语言多样性保护寻找更加有效的解决方案。

参考文献:

[1] 符昌忠. 临高语塞音韵尾弱化的声学表现[J]. 民族语文, 2019(4): 55-62.
FU C Z. Acoustic manifestations of the weakening of plosive endings in the Lingao language[J]. Minority Languages of China, 2019(4): 55-62.

[2] 王莉宁, 康健侨. 中国方言文化保护的现状与思考[J]. 语言战略研究, 2022, 7(4): 76-85.
WANG L N, KANG J Q. Thoughts on current situation of Chinese dialect cultural protection[J]. Chinese Journal of Language Policy and Planning, 2022, 7(4): 76-85.

[3] CHEN L. Computational modeling of lexical variation in Lingao[D]. Singapore: National University of Singapore, 2020.

[4] BABU A R, WANG C H, TJANDRA A, et al. XLS-R: self-supervised cross-lingual speech representation learning at scale[C]//Proceedings of the Interspeech. Piscataway: IEEE Press, 2022: 2278-2282.

[5] KERMANSHAHI M A, AKBARI A, NASERSHARIF B. Transfer learning for end-to-end ASR to deal with low-resource problem in Persian language[C]//Proceedings of the 2021 26th International Computer Conference, Computer Society of Iran (CSICC). Piscataway: IEEE Press, 2021: 1-5.

[6] 张明, 李华, 王伟, 等. 基于细粒度建模单元元学习的低资源语音识别方法[J]. 自动化学报, 2023, 49(7): 1450-1464.
ZHANG M, LI H, WANG W, et al. Meta-learning based low-resource speech recognition method with fine-grained modeling units[J]. Acta Automatica Sinica, 2023, 49(7): 1450-1464.

[7] 王昊, 张卫强, 索华哲, 等. 基于自监督预训练模型的多语言零资源语音识别研究[J]. 清华大学学报(自然科学版), 2022, 62(10): 1721-1730.
WANG H, ZHANG W Q, SUO H Z, et al. Research on multilingual zero-resource speech recognition based on self-supervised pre-trained models[J]. Journal of Tsinghua University (Science and Technology), 2022, 62(10): 1721-1730.

[8] 林佳燕, 黄胡恺, 卢胜辉, 等. 面向闽南方言的自监督模型迁移学习[J]. 厦门大学学报(自然科学版), 2024, 63(4): 687-693.
LIN J Y, HUANG H K, LU S H, et al. Transfer learning of self-supervised models for Minana dialect[J]. Journal of Xiamen University (Natural Science), 2024, 63(4): 687-693.

[9] 陈晓雨, 李晨阳, 王海涛. 非标准化方言语音识别的表征迁移瓶颈分析[J]. 西北工业大学学报, 2024, 42(1): 45-54.
CHEN X Y, LI C Y, WANG H T. Analysis of representation transfer bottlenecks in non-standard dialect speech recognition[J]. Journal of Northwestern Polytechnical University, 2024, 42(1): 45-54.

[10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2017: 5998-6008.

[11] GULATI A, QIN J, CHIU C C, et al. Conformer: convolution-augmented transformer for speech recognition[C]//Proceedings of the Interspeech. Piscataway: IEEE Press, 2020: 5036-5040.

[12] PARK D S, CHAN W, ZHANG Y, et al. SpecAugment: a simple data augmentation method for automatic speech recognition[C]//Proceedings of the Interspeech. Piscataway: IEEE Press, 2019: 2613-2617.

[13] BAEVSKI A, ZHOU Y, MOHAMED A, et al. wav2vec 2.0: a framework for self-supervised learning of speech representations[C]//Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2020: 12449-12460.

[14] HSU W N, BOLTE B, TSAI Y H, et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[15] LI J, YEOH W, TONG R, et al. Adapting Wav2Vec2 for speech recognition on low-resource Tibetan[C]//Proceedings of the Annual Conference of the International Speech Communication Association. Piscataway: IEEE Press, 2022: 2118-2122.

[16] WANG H, KHAN S, CHUANG S, et al. MTLoRA: Multi-task low-rank adaptation for low-resource dialect speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 1582-1594.

[17] WANG D, ZHANG X. THCHS-30: a free Chinese speech corpus[J]. arXiv Preprint, arXiv: 1512.01882, 2015.

[18] GAO T, LI J, ZHU J, et al. Late fusion with learning-based activation function for low-resource speech recognition[C]//Proceedings of the Annual Conference of the International Speech Communication Association. Piscataway: IEEE Press, 2020: 779-783.

[19] ZHANG Y, QIN J, WANG D, et al. Integrating text and audio features for low-resource speech recognition using multimodal conformer[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6182-6186.

[20] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd International Conference on Machine Learning. New York: ACM Press, 2006: 369-376.

[21] CHOROWSKI J, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition[C]//Proceedings of the 29th International Conference on Neural Information Processing Systems. New York: ACM Press, 2015: 577-585.

[22] GRAVES A. Sequence transduction with recurrent neural networks[J]. arXiv Preprint, arXiv: 1211.3711, 2012.

[23] KIM S, HORI T, WATANABE S. Joint CTC-attention based end-to-end speech recognition using multi-task learning[C]//Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2017: 4835-4839.

[24] CHEN N, CHEN Z, WU Y, et al. Improving vietnamese speech recognition with phone-level CTC and cross-lingual transfer learning[C]//Proceedings of the Annual Conference of the International Speech Communication Association. Piscataway: IEEE Press, 2021: 2466-2470.

[25] WATANABE S, HORI T, KIM S, et al. Hybrid CTC/attention architecture for end-to-end speech recognition[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1240-1253.

[26] China Telecom Research Institute. Technical report: telespeech-ASR1.0-large: a 15-billion-parameter speech recognition model for telecom scenarios[R]. 2024.

[27] LUO Y, MESGARANI N. TaSNet: time-domain audio separation network for real-time, single-channel speech separation[C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: ACM Press, 2018: 696-700.
[28] BAEVSKI A, HSU W N, CONNEAU A, et al. data2vec: a general framework for self-supervised learning in speech, vision and language[C]// Proceedings of the 39th International Conference on Machine Learning. New York: ACM Press, 2022: 1298-1312.
[29] DONG L H, XU S, XU B. Speech-transformer: a No-recurrence sequence-to-sequence model for speech recognition[C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2018: 5884-5888.
[30] KARITA S, CHEN N X, HAYASHI T, et al. A comparative study on transformer vs RNN in speech applications[C]//Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Piscataway: IEEE Press, 2019: 449-456.
[31] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: attentive language models beyond a fixed-length context[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2019: 2978-2988.
[32] KIM J, EL-KHAMY M, LEE J. Dual-cross-attention for multimodal audio-visual speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1945-1958.
[33] TENENBAUM J B, FREEMAN W T. Separating style and content with bilinear models[J]. Neural Computation, 2000, 12(6): 1247-1283.
[34] CHAN W, JAITLY N, LE Q, et al. Listen, attend and spell: a neural network for large vocabulary conversational speech recognition[C]// Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New York: ACM Press, 2016: 4960-4964.

[作者简介]

王忠（1969– ），男，海南临高人，海南大学博士生，主要研究方向为人工智能应用、自然语言处理、语音识别。

曹春杰（1977– ），男，河北衡水人，博士，海南大学教授、博士生导师，主要研究方向为无线网络安全、区块链、人工智能安全。

谢夏（1978– ），女，湖北武汉人，海南大学教授、博士生导师，主要研究方向为大数据挖掘、知识图谱应用、人工智能应用等。

穆罕默德·艾哈迈德·拉扎（1995– ），男，巴基斯坦人，海南大学博士生，主要研究方向为网络安全、物联网、智能数据分析与传感器网络。

陈勇青（1987– ），男，海南海口人，博士，海南经贸职业技术学院教授，主要研究方向为自然语言处理、语音识别、自动驾驶环境感知、机器学习、计算机视觉。

陈昱珏（1996– ），男，湖南郴州人，博士，北京原点临近科技有限责任公司高级工程师，主要研究方向为面向风电的AI运维模型、方言语音模型等。