

基于多头注意力双向门控卷积神经网络的藏语语音情感识别

柏梦琳^{1,2} 边巴旺堆^{1,2}
BAI Menglin Bianba Wangdui

摘要

针对藏语安多、卫藏方言语音情感识别中存在的特征挖掘不充分、模型适配性弱等问题,文章提出一种基于多头注意力双向门控卷积神经网络的藏语语音情感识别方法。该方法以 CNN-BiGRU-Attention 为基础架构,通过多尺度 CNN 特征提取增强藏语语音局部情感特征表达,利用双向门控单元完成时序上下文建模,并设计藏语方言专用自注意力机制强化方言情感特征关联、弱化口音差异干扰,构建 Tib-CNN-BiGRU-MultiAtt 优化模型。在 EMODB、RAVDESS 通用数据集与藏语卫藏、安多方言数据集上对比实验表明,所提模型在藏语方言数据集上识别准确率分别提升 4.80% 和 3.82%,在通用数据集上性能稳定无衰减,同时在藏语安多方言数据集上验证了优化模块的有效性。该模型可有效提升低资源藏语方言语音情感识别精度,为民族语言智能语音交互提供技术支撑。

关键词

藏语语音;情感识别;自注意力;双向门控循环单元;卷积神经网络

doi: 10.3969/j.issn.1672-9528.2026.06.005

0 引言

语音作为人类传递信息与表达情感的重要载体,蕴含着丰富的情绪特征,使得语音情感识别成为情感计算与人机交互领域的研究热点^[1-2]。经过多年研究,语音情感识别在英语、汉语等高资源语言中已取得显著成果,在智能客服、在线教育、心理健康监测等场景得到广泛应用^[3]。然而,藏语作为我国特色少数民族低资源语言的相关研究相对较少^[4-6],由于其发音、韵律、语调与通用语种差异显著,导致传统深度学习模型在藏语场景下识别精度偏低,制约了藏语智能语音技术的发展。为此,本文提出基于多头注意力双向门控卷积神经网络的藏语语音情感识别,通过多尺度特征提取与藏语专用自注意力机制模块,有效捕捉藏语方言情感特征,提取提升模型在低资源藏语方言上的识别性能。

1 藏语语音情感识别模型

1.1 语音情感特征表示

语音情感特征的选择与提取直接决定模型识别性能。梅尔倒谱系数(mel-frequency cepstral coefficients, MFCC)^[7]是语音情感识别中最广泛应用的频谱特征,其核心原理是用 Mel 频率模拟人耳的听觉特性。人耳听到的声音高低与实际频率不呈线性关系,在 1 kHz 以下近似呈线性分布,1 kHz 以上

呈对数增长, Mel 频率与线性频率的关系用公式表示为:

$$f_{\text{mel}} = 2595 \cdot \lg \left(1 + \frac{f}{700\text{Hz}} \right) \quad (1)$$

语音短时能量即语音信号短时间内的语音能量,此处的“短”,通常是指一帧,语音短时能量就是将语音中每一帧的能量都计算出来得到短时能量数组,能够反映语音信号的强度。语音帧的能量通常显著高于静默帧,因此可以判断有声段与无声段;同一说话人同一信道下,一般清音部分能量低于浊音部分的能量,由此可以初步区分浊音和清音。短时能量用 E_n 表述:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 = \sum_{m=-N+1}^n [x(m)w(n-m)]^2 \quad (2)$$

式中: $x(m)$ 为离散语音信号的时间序列, $w(n-m)$ 为汉明窗函数, N 为窗长。

MFCC 能够准确捕捉藏语方言语音里的短时声学特征,且其计算复杂度低,抗噪性强,适合藏语方言语音情感特征提取。语音短时能量作为重要声学属性,可以展现藏语方言语音在不同情感表达时的响度起伏、强度变化。因此本文把 13 维静态 MFCC 和 1 维短时能量特征融合作为情感特征表示,既有 MFCC 的频谱精细结构信息,又引入了短时能量对情绪强度的直观呈现,从而更好地刻画藏语方言的情感特征。

1.2 多尺度 CNN 特征提取模块

在 CNN-BiGRU-Attention 为基础架构^[8]中采用的是一维卷积核为 3 的卷积层,情感特征捕捉不够全面,无法适应藏语语音长短音交替、多频段能量分布复杂的特性,很难完整

1. 西藏大学信息科学技术学院 西藏拉萨 850000

2. 西藏大学国家级实验教学示范中心 西藏拉萨 850000

提取藏语方言的局部情感声学特征。因此采用多尺度卷积层结构进行优化,选用尺寸分别为3、5、7的卷积核对输入特征进行并行处理,引入ReLU激活函数、 L_2 正则化($\lambda=0.001$),以提高特征表达能力,避免出现过拟合。特征融合部分,采用特征拼接方式把三种不同尺度的特征图融合,保留各个尺度的原始特征信息,再利用 1×1 卷积层对融合特征进行降维和特征增强操作。最后在特征输出之前增添批归一化层,规范化特征分布,加快模型收敛速度,提高训练稳定性。

1.3 双向门控循环单元

门控循环单元(gated recurrent unit, GRU)的核心是通过门控机制来控制信息的流动,其内部结构包含两个关键门:更新门(Update Gate)和重置门(Reset Gate),结构如图1所示,这种设计使网络能够自适应地保留或遗忘之前的信息状态^[9-10]。

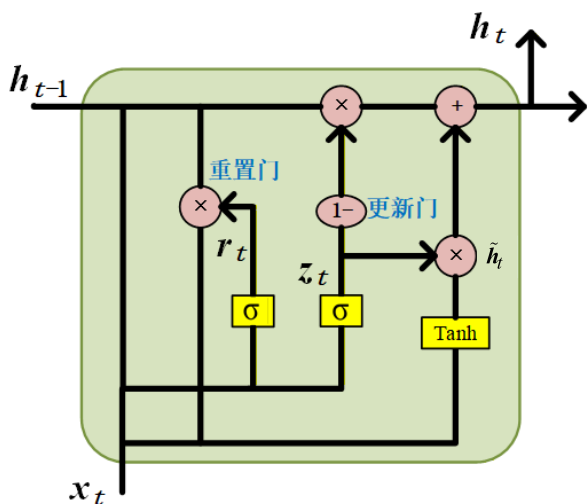


图1 门控循环单元

给定当前时刻的输入 x_t 和前一时刻的隐藏状态为 h_{t-1} 。更新门 z_t 决定了前一时刻的隐藏状态有多少信息需要传递到当前时刻,计算过程为:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (3)$$

式中: σ 为Sigmoid激活函数, W_z 、 b_z 分别为更新门的权重矩阵与偏置向量, $[h_{t-1}, x_t]$ 是把两个向量进行拼接, z_t 越趋近于1,这意味着记忆下来的数据越多,若是越靠近0,即表明遗忘得越多。

重置门 r_t 决定了需要忽略多少之前的隐藏状态信息,计算过程为:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (4)$$

r_t 的值越大,说明上一时刻需要记住的越多,新的输入信息 x_t 与前面的记忆相结合的越多;当 r_t 的值接近0时,值也接近0,说明上一时刻的内容需要全部丢弃,只保留当前

时刻的输入,所以可以用来丢弃与预测无关的历史信息;当 r_t 的值接近1时,表示保留上一时刻的隐藏状态。因此重置门越小,意味着之前的信息被遗忘得越多。

候选隐藏状态 $\tilde{h}_t$ 是当前时刻的新信息表示,计算方式为:

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \odot h_{t-1}, x_t] + b_h) \quad (5)$$

式中: $\odot$ 表示逐元素乘法, W_h 和 b_h 表示对应的权重参数。

最终隐藏状态 h_t 是前一时刻状态和当前候选状态的线性组合,数学表达式如式(6)所示, h_t 忘记传递下来的 h_{t-1} 中的某些信息,并加入当前节点输入的某些信息,即最终的记忆。

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (6)$$

单向门控单元只能顺着语音正向的时间顺序来进行建模,没办法借助后文的信息去关联前文的特征,基于此缺陷,就衍生出了双向门控循环单元(bidirectional gated recurrent unit, BiGRU)。BiGRU通过堆叠两个方向相反的GRU层构成,分别从正向、反向去捕捉时间顺序上的关联,进而获取每个时刻更丰富的上下文信息,其计算公式为:

$$\begin{aligned} \vec{h}_t &= \text{GRU}(x_t, \vec{h}_{t-1}) (\text{前向}) \\ \overleftarrow{h}_t &= \text{GRU}(x_t, \overleftarrow{h}_{t+1}) (\text{后向}) \\ h_t &= [\vec{h}_t; \overleftarrow{h}_t] (\text{拼接}) \end{aligned} \quad (7)$$

图2为BiGRU网络结构示意图,整体有上下两条并行支路:下方是前向GRU支路,沿着语音时序从左至右即从帧1到帧k正向传递特征,以此提取从前文到后文的时序关联。上方是反向GRU支路,其沿着语音时序从右至左也就是从帧k到帧1逆向传递特征,从而提取从后文到前文的时序关联。在两条支路的末端设有特征拼接层,该层会把同一时刻的前向和反向输出特征进行拼接融合,进而形成最终的双向上下文特征输出,呈现出双向时序建模的数据流。

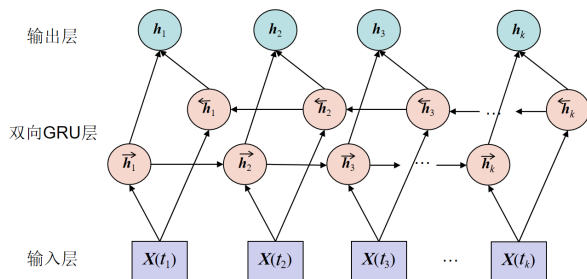


图2 BiGRU网络结构示意图

1.4 藏语方言专用自注意力机制

藏语拥有多种方言,不同方言在发音、语调、韵律等方面存在显著差异。基础自注意力机制仅能全局加权语音特征帧,未针对藏语方言特性做定制化设计,难以有效捕捉这些方言特性。为此,专门针对藏语方言的语音特点设计了藏语

方言专用注意力机制，该机制基于多头注意力架构专项改造，核心用于捕捉藏语安多、卫藏方言间的情感特征关联，弱化方言内部口音差异对识别结果的干扰。主要包括多头注意力机制、方言特征提取、方言门控机制和特征融合四个部分。

多头注意力机制^[11]是一种并行的注意力计算方法，通过多个独立的注意力头同时从不同维度捕捉输入序列的特征信息。每一个注意力头都能够自主地学习不一样类型的特征模式，进而强化模型对于藏语语音特征的表达能力。每一个注意力头用独立的权重矩阵来计算注意力分布。注意力数学表达为：

$$u_i = \tanh(W_i \cdot x_t + b_i) \quad (8)$$

式中： x_t 为BiGRU在时间步 t 的输出， W_i 和 b_i 为第 i 个注意力头的可学习参数。

注意力权重归一化计算方法为：

$$a_i = \text{softmax}(U_i \cdot u_i) \quad (9)$$

式中： U_i 为第 i 个注意力头的权重矩阵。

随后对特征进行加权聚合：

$$h_i = \sum_{t=1}^T a_i \cdot x_t \quad (10)$$

式中： T 为输入序列长度， h_i 为第 i 个注意力头的输出特征。将4个注意力头的输出特征拼接，形成多头注意力的最终输出 $H=[h_1, h_2, h_3, h_4]$ 。

方言门控机制是一种自适应权重调整机制，可对不同藏语方言特征贡献程度进行动态平衡。该机制对输入语音的特征研究分析，自动计算出每种方言的门控权重，实现对不同方言特征的动态调节。

方言门控机制的实现进程能够归纳为三个关键步骤。第一步，把多头注意力模块输出的特征 (h_1, h_2, h_3, h_4) 同方言特征提取模块所生成的原始方言特征 (f_1, f_2) 予以拼接，进而形成门控网络的输入特征 $X_{\text{concat}} = \text{concat}(h_1, h_2, h_3, h_4, f_1, f_2)$ ；第二步，借助门控网络处理拼接后的特征，并运用softmax激活函数生成方言门控权重向量 $g = [g_1, g_2]$ ；第三步，将生成的门控权重逐个元素应用到原始方言特征上，依据式(11)计算得出门控后的方言特征 $[\hat{f}_1, \hat{f}_2]$ 。

$$\hat{f}_d = g_d \cdot f_d \quad (11)$$

式中： g_d 为第 d 个方言的门控权重， $\hat{f}_d$ 为门控后的方言特征。

特征融合模块负责将多头注意力输出的多维度特征与门控后的方言特征进行有效整合，生成最终的综合特征，用公式表示为：

$$h_{\text{final}} = F(\text{concat}(h_1, h_2, h_3, h_4, \hat{f}_1, \hat{f}_2)) \quad (12)$$

1.5 搭建 Tib-CNN-BiGRU-MultiAtt 藏语语音情感识别网络

该模型整体保留了基础 CNN-BiGRU 的核心架构，依次

包括输入层、多尺度 CNN 特征提取层、BiGRU 时序建模层、藏语方言专用注意力层、全连接层和输出层。从数据集里提取 14 维融合特征作为特征输入，多尺度 CNN 层并行提取不同尺度的频谱特征并进行融合，以此来增强对藏语复杂语音结构的表征能力；BiGRU 层负责完成双向长时序情感上下文建模，进而捕捉语音前后的依赖关系；藏语专用注意力层结合多头注意力机制与方言门控，有效捕捉藏语方言情感特征，最终通过两层全连接网络完成对 5 类情感分类输出。

2 实验与结果分析

2.1 语音情感数据集

实验采用两组通用语音情感数据集与两组藏语方言数据集开展对比验证。包括 RAVDESS 英语多模态情感数据集^[12]，涵盖中性、快乐、悲伤、愤怒、恐惧等 8 类情感共 7 356 条语音样本；EMODB 德语情感语音基准数据集^[13]，包含愤怒、悲伤、中性等 7 种情感，共 535 条语音样本；TSEC-4528 藏语卫藏方言情感语音集和 Anduo-3500 藏语安多方言情感语音数据集，均涵盖 5 类基本情感，有效样本分别有 4 528 条、3 500 条。本实验只采用各个数据集中的中性、快乐、悲伤、愤怒、恐惧 5 类情感。实验以准确率作为模型性能评价指标。对数据集采用随机划分的方式来处理，依据 8:2 的比例，划分为训练集、测试集。

2.2 结果对比

表 1 为四个数据集在模型 Tib-CNN-BiGRU-MultiAtt 藏语语音情感识别网络上的识别准确率情况。

表 1 优化模型与基础模型在四个数据集上的性能对比

	单位：%		
	CNN-BiGRU-Attention	Tib-CNN-BiGRU-MultiAtt	提升幅度
EMODB	90.39	91.72	+1.33
RAVDESS	87.21	88.18	+0.97
TSEC-4528	78.35	83.15	+4.80
Anduo-3500	80.43	84.25	+3.82

通过表 1 可见，优化模型在两组通用英文数据集上准确率小幅提升，涨幅分别为 1.33% 和 0.97%，无性能衰减情况，说明针对藏语方言的定制化优化，并未破坏模型对通用语音的识别能力，优化方案具备良好的适配普适性；在两组藏语方言数据集上，优化模型性能提升显著，卫藏方言数据集准确率提升 4.80%，安多方言数据集提升 3.82%，均突破 3.5 个百分点，充分证明藏语方言专用注意力机制与多尺度 CNN 模块的组合优化，能适配藏语安多、卫藏方言的语音与情感表达特性，提升藏语情感识别精度。

2.3 优化模块有效性分析

为进一步验证优化效果,选取藏语安多方言数据集 Anduo-3500,通过替换优化模型中的核心组件,量化多尺度 CNN 特征提取和藏语方言专用注意力机制对模型性能的贡献,验证各组件的有效性及其协同效应。

表 2 不同优化模块在 Anduo-3500 数据集上的识别性能比较

模型名称	优化方向		识别准 确率 /%	较基础 模型 /%
	藏语专用 注意力	多尺度 CNN		
CNN-BiGRU-Attention	—	—	80.43	—
CNN-BiGRU-MultiAtt	√	—	83.24	+2.81
Tib-CNN-BiGRU-Attention	—	√	81.07	+0.64
Tib-CNN-BiGRU-MultiAtt	√	√	84.25	+3.82

基于表 2 比较结果显示,藏语方言专用注意力机制是模型性能提升的核心来源,单独应用可将准确率从 80.43% 提升至 83.24%,提升了 2.81%,而多尺度 CNN 作为辅助优化组件,单独应用仅提升 0.64%;当两者结合时,完整优化模型(Tib-CNN-BiGRU-MultiAtt)准确率达到 84.25%,提升了 3.82 个百分点,两个优化模块的组合产生了 1+1>2 的效果,体现出明显的协同效应。验证了针对藏语特性的专门优化策略的有效性,为藏语语音情感识别提供了更高效的技术路径。

3 结论

本文提出一种基于多头注意力双向门控卷积神经网络的藏语语音情感识别方法,通过多尺度 CNN 与藏语方言专用自注意力机制,显著提升模型对安多、卫藏方言情感特征的提取与建模能力。实验结果表明,该模型在藏语方言数据集上识别精度提升明显,且在通用数据集保持稳定,可为低资源民族语言语音情感识别提供有效方案。

参考文献:

- [1] 李海峰,陈婧,马琳,等.维度语音情感识别研究综述[J].软件学报,2020,31(8):2465-2491.
- [2] 陶建华,陈俊杰,李永伟.语音情感识别综述[J].信号处理,2023,39(4):571-587.
- [3] Akçay M B, Oğuz K. Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers[J]. Speech Communication, 2020, 120: 102744.
- [4] 彭毛扎西,才智杰,才让卓玛.藏语情感语音数据库构建[J].北京大学学报(自然科学版),2023,59(5):773-781.
- [5] 蔡优新,边巴旺堆.基于双向 GRU 模型的藏语语音情感识别[J].信息技术与信息化,2023(10):209-213.

- [6] 王希,王君堡,边巴旺堆.基于卷积神经网络的藏语语音情感识别[J].信息技术与信息化,2022(11):202-206.
- [7] 韩一,王国胤,杨勇.基于 MFCC 的语音情感识别[J].重庆邮电大学学报(自然科学版),2008(5):597-602.
- [8] Zhang Liyan, Wang Yetong, Du Jiaxin, et al. CNN-BiGRU speech emotion recognition based on attention mechanism[C]//2023 2nd International Conference on Artificial Intelligence and Intelligent Information Processing (AIHIP). Piscataway: IEEE, 2023: 85-89.
- [9] 孙韩玉,黄丽霞,张雪英,等.基于双通道卷积门控循环网络的语音情感识别[J].计算机工程与应用,2023,59(2):170-177.
- [10] 卢官明,袁亮,杨文娟,等.基于长短期记忆和卷积神经网络的语音情感识别[J].南京邮电大学学报(自然科学版),2018,38(5):63-69.
- [11] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.
- [12] Livingstone S R, Russo F A. The ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in north american english[J]. PLoS ONE, 2018, 13(5): e0196391.
- [13] Burkhardt F, Paeschke A, Rolfes M, et al. A database of German emotional speech[C]//Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005).Lisbon, Portugal: International Speech Communication Association (ISCA), 2005: 1517-1520.

【作者简介】

柏梦琳(2001—),女,重庆荣昌人,硕士研究生,研究方向:藏语语音情感识别。

边巴旺堆(1970—),男,西藏日喀则人,硕士,教授,研究方向:藏文信息处理、通信网络与安全。

(收稿日期:2026-04-27 修回日期:2026-06-11)