基于双向长短时记忆网络的藏语语音情感识别

李珊珊[1] 边巴旺堆[1,2]
LI Shanshan BIANBA Wangdui

摘 要 为提高藏语拉萨方言的语音情感识别准确度，构建了一个包含 6000 条语音样本的语料库，采用了改进的 MFCC 特征提取方法和双向长短时记忆网络（BiLSTM）模型。改进的 MFCC 特征能更有效地表征藏语中的情感信息，而 BiLSTM 模型则能有效捕捉语音序列中的长期依赖关系，这对于情感识别任务尤为重要。研究结果显示，所设计的方法达到了 81% 的准确率，相较于传统方法有显著提升，在处理藏语情感识别方面具有很高的效果和潜力。未来的研究方向包括进一步优化模型结构，探索更多的深度学习架构，改进语音特征提取技术，以进一步提高模型的准确率和泛化能力，为语音情感识别技术在藏语等少数民族语言中的应用奠定重要的基础。

关键词 藏语情感识别；MFCC 特征；长短时记忆网络；语音情感分析；深度学习

doi：10.3969/j.issn.1672-9528.2024.10.003

0 引言

语音情感识别在人机交互领域扮演着重要角色。它的核心目标是通过分析语音中的语调、语速、音调、声音强度等特征，来识别说话者的情感和情绪状态 [1]。对藏语进行 SER 研究，可以促进藏语在人机交互中的应用，并为藏族群体提供更智能、个性化的语音交互体验，推动少数民族语言在现代科技中的应用。

对于藏语语音情感识别的研究较多。2017 年，WU 等学者提出了一种基于隐马尔科夫模型（hidden Markov model，HMM）的汉藏双语情感语音合成方法，实现了从普通话情感训练语料库到藏语中性话语和普通话中性话语的情感移植 [2]。2018 年，HU 等人开展了藏语拉萨话语音情感节奏的认知研究 [3]。2021 年，边巴旺堆等人申请了一种基于 CNN 和 LSTM 的藏语语音情感识别方法专利 [4]。

在传统的 MFCC 特征上，改进的 MFCC 特征如 AMFCC（averaged MFCC），在计算 MFCC 之后，对 MFCC 系数进行归一化处理。在所有帧中计算每个 MFCC 系数的平均值，并从每个系数中减去平均值。这种归一化处理可以减小不同说话人之间的声音差异，使得语音特征更加稳定和可比较。

1. 西藏大学信息科学技术学院 西藏拉萨 850000
2. 西藏大学信息技术国家级实验教学示范中心 西藏拉萨 850000

[基金项目] 00061250/004/ 藏财预指〔2024〕1 号中央支持 - 重点科研平台建设 - 信息技术国家级实验教学示范中心支持

模型结构部分结合了多个深度学习架构，包括三个双向长短时记忆网络（Bi-LSTM）和卷积神经网络（CNN），并且添加了注意力机制。这种结合可以有效地捕捉语音序列中的长期依赖关系（Bi-LSTM）、局部特征（CNN）以及重要信息的加权分布（注意力机制），能够显著提高模型的表现。

1 模型 MFCC 特征的改进

1.1 MFCC 特征提取

梅尔频率倒谱系数（MFCC）是一种常用的语音特征提取方法，它模拟了人耳听觉机理，可以很好地描述语音的静态和动态信息。

单纯的 MFCC 可能无法充分反映语音中的动态情感变化，Gupta 等人提出的加权梅尔频率倒谱系数（weighted MFCC，WMFCC）是对传统 MFCC 的一种改进。WMFCC 基于加权思想，结合了 MFCC 的一阶和二阶差分。这种加权的方式可以提高对语音信号中动态变化的响应能力，从而更好地捕捉和表征说话者情感状态的变化 [5]。

MFCC 特征提取步骤包括几下几方面。

预处理：对音频信号进行预处理，如去除噪声、降低采样率等。

快速傅里叶变换（FFT）：将时域信号转换为频域信号，以获得每个时间点的频域谱。通过 FFT 算法，可以快速计算出信号的频谱信息，包括频率成分、振幅和相位等。

Mel 滤波器组：使用一种人耳听觉系统感知声音频率的方式，它模拟了人耳听觉系统在不同频率下的响应。使用一

组 Mel 尺度的三角形滤波器对能量谱进行转换，以更好地模拟人耳的听觉机制。对频域谱进行滤波，将高频能量区域划分成若干个区间，并将每个区间内的能量值相加，得到该区间的 Mel 系数，图 1 为滤波器组系数。

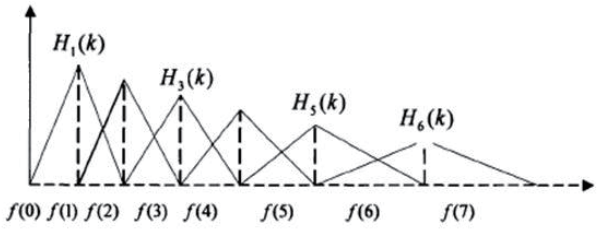

图 1 Mel 滤波器组系数

对数运算：对 Mel 系数取对数。人耳听觉系统对频率的感知是呈现对数关系的，因此在语音信号处理中，通常会对能量谱进行一次对数运算，以更好地模拟人耳的感知机制。这样可以更好地描述语音信号中不同频率成分的能量分布情况，其公式为：

$$\mathrm{Mel}(f) = 2595 \times \log(1 + \frac{f}{7}) \tag{1}$$

离散余弦变换（DCT）：对取了对数的 Mel 系数进行离散余弦变换，得到一组 MFCC 系数。

1.2 一阶差分及 AMFCC 特征提取

改进的 MFCC 特征提取方法在 MFCC 系数上进行二次计算，使用差分算法，计算相邻帧之间的 MFCC 系数差异，并将其添加到原始 MFCC 系数中，以增强特征的动态性能。通过计算 MFCC 系数的一阶差分特征（ΔMFCC）来实现。

利用连续相邻两个系数计算其一阶差分特征：

$$\Delta\mathrm{MFCC}_K = \frac{\mathrm{MFCC}_{K+1}}{2} - \frac{\mathrm{MFCC}_{K-1}}{2} \tag{2}$$

不同说话人的声音特性（如音高、音色等）会导致 MFCC 特征的分布有所不同。通过对 MFCC 系数进行均值归一化，可以有效减少这些差异，使得特征更加稳定和可比较。对 MFCC 系数进行归一化的一种常见方法是对每个系数减去其在 Mel 滤波器组上的平均值，然后除以一个标准差。计算公式如下：

$$\mathrm{AMFCC} = \mathrm{MFCC} - \overline{\mathrm{MFCC}}|_{\mathrm{Mel}} \tag{3}$$

在提取特征时的具体实现：首先，对语音情感数据集的语音信号进行分帧，将其分成长度为 265 的帧，并使用适当的重叠来保留时间信息。接下来，计算出 MFCC 系数，然后通过 MFCC 参数计算 MFCC 的一阶差分以及 AMFCC 参数，各 70 维，以捕获信号的动态特性。最后，将两种特征拼接在一起，形成 (265, 140) 维的特征向量，用于本文的藏语语音情感识别的特征。

2 模型构建部分

Bi-LSTM 中具有判断信息是否有用的记忆单元处理模块 [6]，CNN 则进一步提取并保留重要的特征频谱，从而提高识别准确率。因此，结合 Bi-LSTM 和 CNN 可以更有效地识别语音情感。

本文模型的输入采用改进的 MFCC 特征，使用的特征形状为一个高度为 265、宽度为 140、通道数为 1 的三维数组。

LSTM 能够捕捉语音信号中的长期依赖，例如对语音信号的语调、音频序列中的上下文信息进行有效建模。LSTM 主要由 3 个单元构成：输入门、遗忘门、输出门。LSTM 通过遗忘门来决定哪些信息应该被保留或丢弃，通过输入门来决定如何更新细胞状态，通过输出门来决定哪些信息会作为当前时间步的输出，如式（4）～（8）。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4}$$

$$i_t = \sigma(W_t \cdot [h_{t-1}, x_t] + b_i) \tag{5}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{6}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \tag{7}$$

$$h_t = o_t \circ \tanh(c_t) \tag{8}$$

式中：f_t 表示遗忘门，i_t 表示输入门，o_t 表示输出门，c_t 表示当前状态，h_t 表示当前时间步的隐藏状态，h_{t-1} 表示前一时间步的隐藏状态，σ 表示激活函数。通过三个门控制信息进出单元的传输，最终得到与输入序列长度相同的隐层状态序列。

在 BiLSTM 中，前向 LSTM 负责从序列的起始位置开始学习信息，而后向 LSTM 则从序列的末尾位置开始学习。通过这种双向学习的方式，模型可以同时考虑到序列数据中前后位置的信息，从而更全面地理解输入数据。此层网络有三个并行的双向 LSTM 层，每个 LSTM 层的输出都被发送到后续的卷积层。图 2 为 LSTM 的结构图。

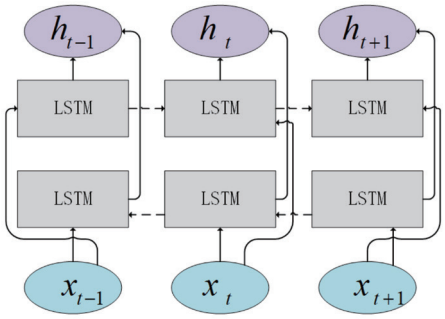

图 2 LSTM 网络结构

然后通过一维卷积层：每个 LSTM 层的输出都经过一个一维卷积层处理。最大池化层：对每个卷积层的输出进行池化。

使用混合层次特征融合和注意力模块可以帮助模型更有效地选择和整合与任务相关的信息，尤其是在情感分析等需

要从多个角度理解文本或语音数据的应用中。自注意力机制用于捕捉输入序列的全局依赖关系，计算序列中的每个位置对应的表示向量与其他位置向量的相似度以得到权重向量，利用权重向量对表示向量加权求和[7]。

利用 concatenate，其作用是将多个张量在某个维度上进行拼接，可以处理具有不同特征的数据或者在模型中合并多个层的输出。混合多尺度卷积充分利用了多尺度信息的丰富性，通过合理设计卷积核的尺寸和结构，能够在减少网络深度和参数数量的同时，实现更高效的特征提取和模型训练[8]。

使用全连接层结合 Softmax 激活函数输出 5 种情感类别的概率归一化到 [0,1]。使用激活函数输出 5 类分类结果，分别是生气、恐惧、快乐、中性、悲伤[9]。

连接三个 LSTM 和 CNN，并结合最大池化层及连接注意力机制，能显著增强深度学习模型在序列数据处理中的性能。LSTM 层适用于捕捉长期依赖关系，CNN 则优于提取局部特征，二者结合利用最大池化层进一步精炼特征表示。连接多层特征能够丰富输入表达，提高模型对复杂数据结构的理解能力，同时最大池化层有助于减少参数数量，降低过拟合风险。加入注意力机制使模型能够动态关注输入中的关键信息，从而提升任务性能和泛化能力。网络结构图如图 3。

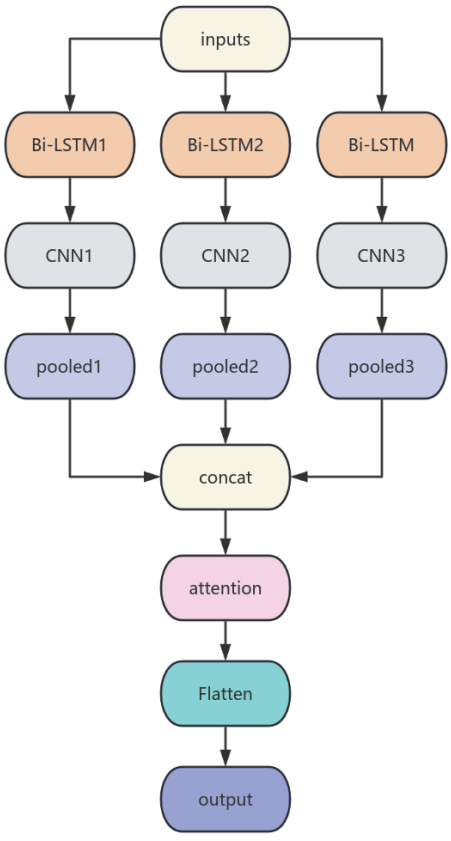

图 3 网络结构图

具体参数：3 个双向 LSTM 层，每个 LSTM 层有 64 个隐藏单元，每个 LSTM 的输出通过具有 32 个过滤器和 3 个卷积核大小的 1D 卷积层处理。每个卷积层后面接一个最大池化层，使用池化大小为 2，有助于减少特征的数量并保留重要信息。

池化后的特征被连接成一个更大的特征向量，然后应用了自注意力机制。展平层将注意力机制的输出展平成一维向量，最后利用 Softmax 函数进行分类得出结果。

3 模型的比较与评估

本文的 MFCC 一阶差分以及 AMFCC 特征应用到不同的网络模型的对比如表 1。

表 1 模型准确率比较

模型	准确率
CNN	0.64
CNN+LSTM	0.66
LSTM+Attention	0.77
本文模型	0.81

从表 1 中可以看出，LSTM+Attention 和本文模型在这个文本分类任务中表现出了较好的性能。结果表明在处理文本数据时，使用 LSTM 结合注意力机制可能会带来更好的性能。

对于每个类别（Angry, Fear, Happy, Neutral, Sad），报告了精确率（precision）、召回率（recall）和 F_1 分数，见表 2。

表 2 评价指标

情绪	精确率	召回率	F_1 分数
Angry	0.90	0.80	0.85
Fear	0.74	0.90	0.81
Happy	0.77	0.82	0.79
Neutral	0.76	0.75	0.75
Sad	0.86	0.73	0.79
平均值	0.81	0.80	0.80

评价指标包括准确率（accuracy）、召回率（recall）、精确率（precision）以及 F_1-score。这些指标通常用于评估分类模型在多类别分类任务中的表现[10]。

宏平均（macro avg）指标，即对所有类别的精确率、召回率和 F_1 分数的平均值。

通过这些指标，可以得出以下结论：模型在识别“Angry”情绪方面表现最佳，具有较高的精确率、召回率和 F_1 分数。

“Fear”情绪的召回率最高，达到了 0.90，但精确率稍低。

整体而言，模型在各个类别上的表现相对均衡，宏平均的精确率、召回率和 F_1 分数都在 0.80 左右。

数据集的加权平均指标与宏平均指标非常接近，这表明数据集中各类别样本数量相对均衡。

评估语音情感识别模型在每个情感类别上的准确率通过构建混淆矩阵并计算每个类别的正确预测比例来完成，如图 4。

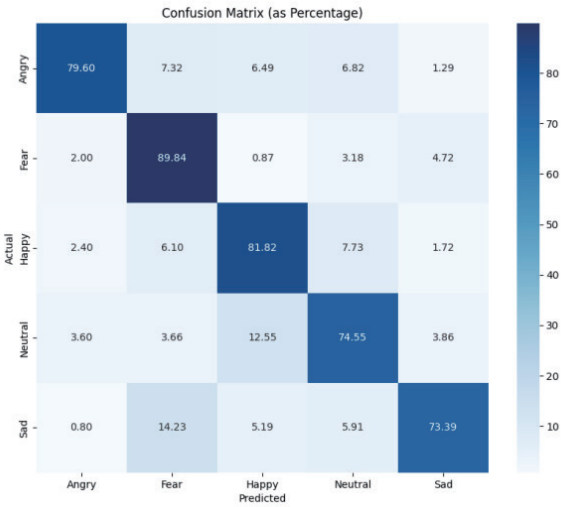

图 4 混淆矩阵

混淆矩阵的每一行代表实际类别，每一列代表预测类别。通过混淆矩阵，可以看到模型在哪些类别上表现较好，在哪些类别上容易产生混淆。在本实验中，通过混淆矩阵可以分析出模型在预测 Fear 类别上表现较好，几乎有 90% 的样本被正确分类。模型在预测 Happy 和 Angry 类别上也表现较好，分别有约 82% 和 79% 的样本被正确分类。但是在预测 Neutral 和 Sad 类别上的表现相对较差，分别有约 75% 和 73% 的样本被正确分类。模型在某些类别之间存在一定程度的混淆，如 Neutral 和 Happy 之间以及 Neutral 和 Fear 之间。

4 结论

本文通过融合基于 MFCC 特征的提取和改进，结合深度学习的应用，实验结果表明 Bi-LSTM-CNN 模型在藏语语音情感识别任务中取得了良好的识别率，达到了 81%，有效提高了语音情感特征，取得了良好的语音情感识别效果。

参考文献：

[1]LATIF S,RANA R,KHALIFA S, et al.Survey of deep representation learning for speech emotion recognition[J]. IEEE transactions on affective computing, 2023, 14(2): 1634-1654.

[2]WU P, YANG H, GAN Z.Towards realizing mand-arin-tibetan bi-lingual emotional speech synthesis with mandarin emotional training corpus[C]// International Conference of Pioneering Computer Scientists. Singapore: Engineers and Educators. Berlin:Springer, 2017: 126-137.

[3]HU A X, XU N, YU H Z. A cognitive study of tibetan lhasa speech emotional rhythm based on emotional rhythm and emotional lexical meaning[C]//2018 2nd International Conference on Data Science and Business Analytics (ICDSBA). Piscataway: IEEE, 2018: 263-267.

[4] 边巴旺堆，王希，王君堡，等 . 一种基于 CNN 和 LSTM 的藏语语音情感识别方法 :CN113808620A[P].2021-12-17.

[5]GUPTA S, SHUKLA R S, SHUKLA R K. Weighted Mel frequency cepstral coefficient based feature extraction for automatic assessment of stuttered speech using Bi-directional LSTM[J].Indian journal of science and technology, 2021, 14(5): 457-472.

[6]WANG Y, HU W P. Speech emotion recognition based on improved MFCC[C]//CSAE'18:Proceedings of the 2nd International Conference on Computer Science and Application Engineering. New York, NY: Association for Computing Machinery, 2018: 22-24.

[7] 滕思航 . 自适应独立性假设与音字特征转换的非自回归中文语音识别研究 [D]. 南宁 : 广西大学 ,2023.

[8] 梁科普，张海军，刘雅情，等 . 混合多尺度卷积结合双层 LSTM 语音情感识别 [J]. 计算机与现代化 ,2023(1):63-68.

[9] 黄喜阳，杜庆治，龙华，等 . 基于 MFCC 特征融合的语音情感识别算法 [J]. 陕西理工大学学报 (自然科学版), 2023, 39(4):17-25.

[10] 陶砚蕴，岳国旗，王凯欣，等 . 心电图信号双任务学习的时空级联神经网络及心律失常分类模型 [J]. 南京大学学报 (自然科学),2021,57(2):318-326.

【作者简介】

李珊珊（2000—），女，辽宁沈阳人，硕士研究生，研究方向：藏语语音情感识别。

边巴旺堆（1970—），通信作者（email:banwangg@163.com），男，西藏拉萨人，教授，硕士生导师，研究方向：藏语语音情感识别、通信网络与安全。

（收稿日期：2024-07-13）