融合MFCC时空特征与胶囊网络的藏语语音情感识别

何强龙[1],黄荣兆[2],陈　雨[1],边巴旺堆[1,3]

(1. 西藏大学 信息科学技术学院,西藏 拉萨 850000;
2. 兴义市人力资源和社会保障局 信息管理中心,贵州 兴义 562400;
3. 西藏大学 信息技术国家级实验教学示范中心,西藏 拉萨 850000)

摘　要: 语音情感识别是自然语言处理领域的一个重要分支,旨在通过计算机技术自动识别与分类语音中的情感信息。在藏语这一特定语言环境中,由于相关研究较少且现有研究普遍采用多特征融合策略,存在特征提取繁琐、特征维度高、计算开销大等问题。为此,取45维的梅尔频率倒谱系数(MFCC)表征音频情绪,提出轻量级网络模型ST-CapsNet。首先,以时空特征提取器与胶囊网络为核心,通过时空特征提取器捕获MFCC的时间、空间信息;其次,将信息输入胶囊网络进行深度分析。在自建的藏语语音情感语料库TSEC-4528、公开语料库EMO-DB、RAVDESS(Speech & Song)上的实验表明,所提模型分别取得了78.59%、96.34%、87.60%和97.54%的未加权准确率,证明了该方法对藏语语音情感识别的有效性与跨语种的适应性。

关键词: 藏语;语音情感识别;梅尔频率倒谱系数;胶囊网络
DOI: 10. 11907/rjdk. 251832　　扫描二维码阅读全文:
中图分类号: TP18　　**文献标识码:** A　　**文章编号:** 1672-7800(XXXX)0XX-0001-09

Tibetan Speech Emotion Recognition by MFCC Spatiotemporal Features and Capsule Networks

HE Qianglong[1], HUANG Rongzhao[2], CHEN Yu[1], BIAN Bawangdui[1,3]

(1. *College of Information Science and Technology, Xizang University, Lhasa* 850000, *China*;
2. *Information Management Center of Xingyi Human Resources and Social Security Bureau, Xingyi* 562400, *China*;
3. *National Experimental Teaching Demonstration Center for Information Technology, Xizang University, Lhasa* 850000, *China*)

Abstract: Speech emotion recognition is an important branch of natural language processing that aims to automatically recognize and classify emotional information in speech through computer technology. In the specific language environment of Tibetan, due to the lack of relevant research and the widespread use of multi feature fusion strategies in existing studies, there are problems such as cumbersome feature extraction, high feature dimensionality, and high computational overhead. To this end, a lightweight network model ST CapsNet is proposed by using 45 dimensional Mel frequency cepstral coefficients (MFCC) to represent audio emotions. Firstly, the spatiotemporal feature extractor and capsule network are used as the core to capture the temporal and spatial information of MFCC through the spatiotemporal feature extractor; Secondly, input the information into the capsule network for deep analysis. The experiments on the self built Tibetan speech emotion corpus TSEC-4528 and the publicly available corpora EMO-DB and RAVDESS (Speech&Song) showed that the proposed model achieved unweighted accuracies of 78.59%, 96.34%, 87.60%, and 97.54%, respectively, demonstrating the effectiveness and cross lingual adaptability of the method for Tibetan speech emotion recognition.

Key Words: Tibetan; speech emotion recognition; Mel-frequency cepstral coefficients; capsule net

收稿日期: 2025-12-22
基金项目: 西藏自治区自然科学基金项目(XZ202401ZR0051);西藏自治区高原通信科研创新团队项目(XZZZQ2018003)

作者简介: 何强龙(2000-),男,CCF学生会员,西藏大学信息科学技术学院硕士研究生,研究方向为语音情感识别;陈雨(2002-),女,西藏大学信息科学技术学院硕士研究生,研究方向为语音情感识别;边巴旺堆(1970-),男,博士,西藏大学信息科学技术学院教授,研究方向为通信网络与安全、藏语语音情感识别。本文通讯作者:边巴旺堆。

0 引言

在人类沟通中，语音既承载了语义信息，又是情感的桥梁，研究者正致力于发展语音情感识别（Speech Emotion Recognition，SER）技术[1,2]。该任务包含构建语音情感语料库、提取语音情感特征、模型训练与评估等关键步骤。其中，提取语音情感特征和创建高效模型是SER任务的主要研究内容，对提升情感识别的准确性和鲁棒性具有重要意义[3]。

目前，语音的情感特征分为韵律特征、频谱特征、音质特征等[4-8]。早期，SER任务主要通过提取3类特征作为模型的输入，并通过线性判别分类器、高斯混合模型、隐马尔可夫模型、k-近邻分类器、支持向量机等常用的模型对情绪进行分类[9-11]。近年来，递归神经网络、长短期记忆神经网络等深度学习模型，通过循环结构捕捉语音信号的短期和长期依赖关系，解决了传统方法中情感动态建模难的问题[12-14]。Transformer通过自注意力机制、并行架构解决了传统模型在时序建模和计算效率上的瓶颈[15]。迁移学习通过知识迁移与参数优化缓解了SER任务中的数据稀缺问题[16]。

尽管，深度学习在语音情感识别领域取得了显著进展，但由于藏语具有复杂的韵律特征和地域文化属性，使藏语语音情感识别研究仍面临数据稀缺性、标注复杂性、情感表达存在地域性差异等挑战[17]。Guo等[18]针对卫藏方言基础声学分析层面展开探索，为后续研究提供了重要基准。面对藏语情感语料匮乏的瓶颈，次仁罗增等[19]构建了藏语文本语料库和情感语音数据集，提出的联合建模框架取得了76%的平均准确率。蔡优新[20]在自建的卫藏方言数据集上，对多种深度学习模型进行实验，并对网络结构加以优化，进一步提升了识别精度。在多模态方面，王希[21]构建了多特征融合模型，在自建的藏语数据集上情感分类准确率达到84.5%。谷泽月等[22]提出的312维卫藏方言手工特征集（TPEFS），融合了声学与语言学特征，在5类情感识别任务中的准确率为88.4%，在多语言迁移测试（德语EMODB、汉语CASIA）中保持了84%以上的泛化性能。彭毛扎西[23]提出的动态加权融合策略，进一步拓展了多特征融合的维度，并在跨方言测试中验证了模型鲁棒性。

尽管多特征融合技术在藏语语音情感识别领域取得了一定进展，但仍面临多维特征融合复杂度高、模型结构冗余、手工特征拼接引发的维度爆炸等挑战。例如，InterSp09特征集的维度达到384维[21]；TPEFS包含10种低水平特征及其统计特征，融合特征维度达到312维[22]；彭毛扎西等的模型输入特征维度达到325维[23]。虽然，这些高维特征集提升了识别率，但会导致计算量和存储需求呈指数级增长，尤其在大规模数据处理时对硬件资源和算法效率的要求更高[24]。

为此，本文提出轻量化语音情感识别模型ST-CapsNet（Spatio-Temporal Capsule Network）。该模型融合了梅尔频率倒谱系数（Mel-Frequency Cepstral Coefficients，MFCC）特征的时空信息与胶囊网络（Capsule Network）的动态路由机制，能精准识别语音情感。本文所用的藏语语音特征值下载链接为https://gitee.com/Utibet__TSIP_Laboratory/tsec_4528上。

1 ST-CapsNet

图1为本文模型的网络结构，包括输入特征梅尔频率倒谱系数、时空特征提取器（Spatiotemporal feature extractor，SFE）和胶囊网络。时空特征提取器由逐时间步残差连接的长短期记忆神经网络（Long Short-Term Memory，LSTM）网络与4层卷积模块组成，以OpenSMILE工具提取的45维MFCC特征为输入（帧长40 ms，帧移10 ms）。首先，通过残差LSTM增强时间细粒度表征，利用卷积层提取关键局部特征；其次，通过胶囊网络动态路由算法迭代优化特征空间分布，生成具有情绪语义的胶囊向量；最后，通过全局平均池化层与Softmax分类器实现高效、高精度地语音情感识别。

1.1 梅尔频率倒谱系数

MFCC的求解包含分帧、加窗、快速傅里叶变换、计算功率谱、梅尔滤波、取对数、离散余弦变换等步骤。为了从连续语音信号中提取短时稳定的特征，需按预定“帧长度”和“帧偏移量”切割语音信号，再对每“帧”应用汉明窗以减少边缘突变。在实际的数字信号处理流程中，需先将连续的语音信号离散化为一系列在离散时间点上的采样值，再进行“分帧加窗”（一般称为预处理）。

假设离散化语音信号得到的离散信号为$x[n]$，n表示采样点的索引，S为总样本点数。在离散化框架下，每帧含N个采样点，帧偏移量为M个样本点，$\omega[n]$表示汉明窗，则第i帧信号为：

$$\begin{aligned} x_i[m] &= x[m+(i-1)M]\cdot\omega[m] \\ \omega[m] &= \alpha-\beta\cos(2\pi m/N-1) \end{aligned} \tag{1}$$

式中：α、β为汉明窗的参数。

预处理后，每帧信号经过快速傅里叶变化以提取频域特征$X_i[k]$。

$$X_i[\mathrm{k}]=\sum_{n=0}^{N-1}x_i[n]\cdot \mathrm{e}^{-j\frac{2\pi}{N}kn} \tag{2}$$

随后，对快速傅里叶变化结果求模并取平方，即对每一个k，计算$\left|X_i[k]\right|^2$，即可得到功率谱$P_i[k]$。梅尔频率倒谱系数在SER领域之所以得到广泛应用，原因在于其采用的梅尔刻度能模拟人耳的听觉特性。通过梅尔刻度后的频率会转换为梅尔频率：

$$f=2596\cdot\log_{10}\left(1+\frac{f_{HZ}}{700}\right) \tag{3}$$

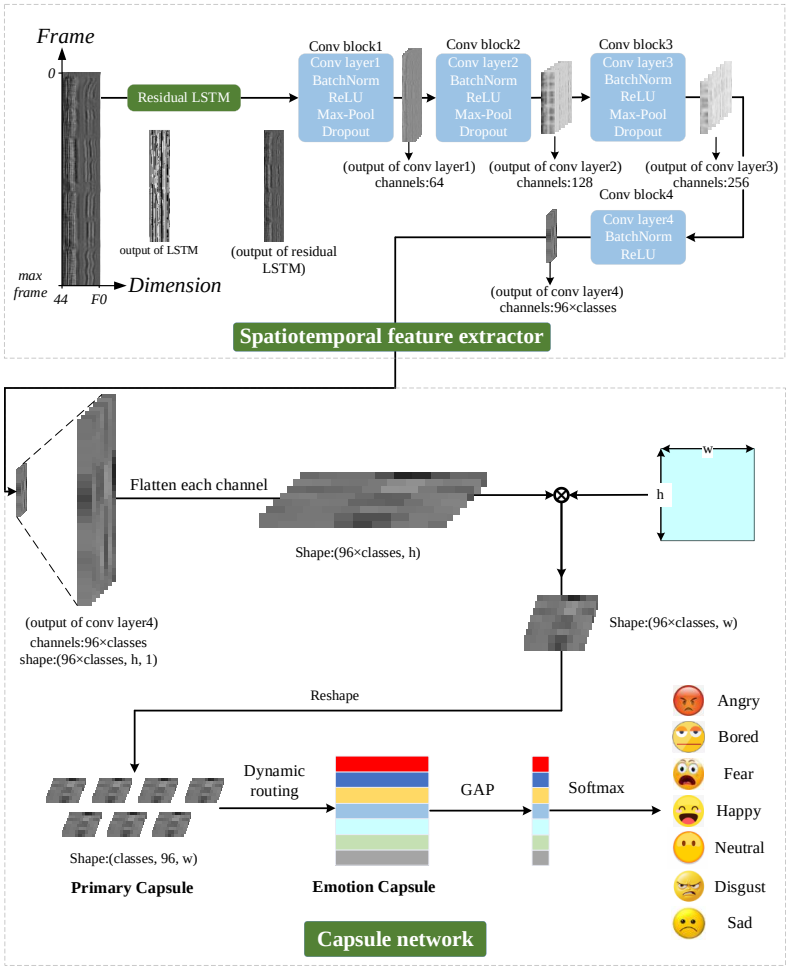


Fig. 1　Overall structure of the model

图 1　模型整体结构

相较于高频信号，人耳对低频语音信号的差异具有更高的感知敏感度。在实际应用中，将f_{HZ}转换成f需要使用梅尔滤波器进行滤波。事实上，梅尔滤波器是一系列的三角形滤波器，记Q为梅尔滤波器（1≤q<Q）的数量，第q个三角滤波器的中心频率为$f(q)$，对应的频率响应为：

$$H_q(p)=\begin{cases}0 & p<f(q-1),f>f(q+1)\\ \dfrac{p-f(q-1)}{f(q)-f(q-1)} & f(q-1)\leqslant p\leqslant f(q)\\ \dfrac{p-f(q-1)}{f(q)-f(q-1)} & f(q)\leqslant p\leqslant f(q+1)\end{cases}\tag{4}$$

对第q个滤波器输出的功率谱取对数为：

$$s_i(q)=\ln\left(\sum_{k=0}^{N-1}P_i[k]H_q(k)\right)\tag{5}$$

最后，经过离散余弦变换得到第i帧的MFCC为：

$$C_i(n)=\sum_{m=1}^{Q}s_i(m)\cos\left(\frac{\pi n(m-0.5)}{Q}\right)\tag{6}$$

式中：$n\in\{1,2,.....,L\}$，L为MFCC特征的阶数，本文以45维MFCC特征作为模型输入；Q为梅尔滤波器的个数。

该特征提取方法操作简便，无需复杂的计算流程，对硬件配置要求较低，能有效降低数据处理成本，提升处理效率。

1.2　长短时记忆网络

本文通过LSTM为MFCC增加时间颗粒表征。LSTM解决长期依赖问题的关键在于细胞状态，可利用遗忘门、记忆门和输出门协同控制信息的流动与存储、调整信息量，不直接参与输出计算，能在多个时间步中保留关键信息，避免梯度消失或爆炸。

遗忘门让网络决定遗忘的内容，原因为上一时刻单元状态部分信息会“过时”，因此无需输入到下一个时刻，故可遗忘部分分量。具体公式为：

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f)\tag{7}$$

式中：σ表示激活函数，常为Sigmoid函数；$[h_{t-1},x_t]$将上一时刻的隐藏状态h_{t-1}与输入x_t进行拼接，并将拼接结果与权重W_f相乘并加偏置，再通过激活函数得到当前门的输出f_t；Sigmoid激活函数将输入压缩到(0,1)区间内，若输出为0就会被遗忘丢弃，输出为1时当前的单元状态就会被完整地保留；LSTM网络通过Sigmoid函数实现长期记忆重要信息，且可随输入动态调整。

随后记忆门接管，从当前输入中提取关键新信息生成候选记忆值，将其与与遗忘门过滤后的历史记忆融合，形

成更新后的细胞状态。

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{8}$$

式(8)中,输入x_t与上一时刻隐藏状态h_{t-1}拼接后与权重相乘,再与偏置相加后输入激活函数,即可得到当前的输入。i_t与f_t的概念不同,f_t是决定要遗忘或保留细胞中的何种信息,i_t则是更新细胞状态的输入数据,也被称为输入门。除了确定输入信息,记忆门还需要利用tanh函数“候选”出当前输入x_t的有效信息,即$\tilde{C}_t$。在计算得到$\tilde{C}_t$、i_t后,将二者相乘,再与通过遗忘门的细胞相加得到更新后的细胞。

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{9}$$

综上,输出门用于决定当前时刻的输出及修改当前时刻的隐藏状态,以便在下一个时刻影响遗忘门、记忆门。

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_0)$$
$$h_t = o_t \cdot \tanh(C_t) \tag{10}$$

式中:o_t为当前时刻的输出,通过Sigmoid函数筛选输出内容,并与tanh函数处理后的细胞相乘,生成隐藏状态h_t,隐藏状态可视为短期记忆,将传递到下一个时间步长中。

总之,LSTM网络间各门控相互协作:遗忘旧信息、添加新信息并生成输出以实现模型预测。

1.3 时空特征提取器

为了提取MFCC特征的时空信息,设计了时空特征提取器。首先,通过逐时间步残差连接的LSTM网络,增加输入特征的时间细粒度;其次,利用多层卷积模块的放缩不变性捕捉关键信息;最后,多层卷积模块使用多个卷积核学习、提取不同空间特征,其数量随着卷积层叠加而依次递增,并组合时空特征更全面地理解语音信号中情绪的时空信息,提高识别性能。

此外,多层卷积模块中引入了批次归一化(Batch Normalization,BN)减轻内部变量偏移;采用ReLU(Rectified Linear Unit)函数增加非线性表达能力;使用最大池化层(Max-Pooling)降维保留特征图信息;采用丢弃(Dropout)防止过拟合。表1为多层卷积模块的具体参数配置。

Table 1 Parameter configuration of multi-layer convolution module

表1 多层卷积模块参数配置

Layer Name	Component	Kernel	Stride	Filters	说明
Conv1	Convolution layer	(3, 3)	(1, 1)	64	
	BN + ReLU	–	–	–	
	Max-Pooling	(2, 2)	(2, 2)	–	
	Dropout	–	–	–	参数:0.5
Conv2	Convolution layer	(3, 3)	(1, 1)	128	
	BN + ReLU	–	–	–	
	Max-Pooling	(2, 2)	(2, 2)	–	
	Dropout	–	–	–	参数:0.15
Conv3	Convolution layer	(3, 3)	(1, 1)	256	
	BN + ReLU	–	–	–	
	Max-Pooling	(2, 2)	(2, 2)	–	
	Dropout	–	–	–	
Conv4	Convolution layer	(3, 3)	(1, 1)	96×num_emotions	num_emotions为语料情绪种类数
	ReLU+BN				

1.4 胶囊网络

胶囊网络由主胶囊层(Primary Capsule)与情绪胶囊层(Emotion Capsule)组成,显著特点是将计算得到的神经元从单一数值转为向量。该向量兼具神经元的存在性、方向性和空间性[25]。原始输入经时空特征提取器进行局部上下文建模后,将输出的三维特征张量(通道×时间步×频率单元)沿时间和频率维度展平,在保持通道不变的同时得到保留了原始时间步顺序的特征向量序列,如图2所示。

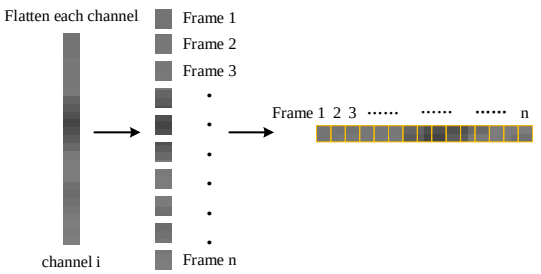


Fig. 2 Feature flattening

图2 特征展平

接下来,将展平数据与权重矩阵相乘以重塑主胶囊层,使胶囊经动态路由算法输出至情绪胶囊层。动态路由更新算法是胶囊网络核心内容,需对主胶囊层向量执行多次迭代计算(本文使用3次迭代)。假设每个胶囊为u_i,算法开始前令路由系数$b_{ij} \leftarrow 0$,先对路由系数使用Softmax函数得到系数c_{ij},再将u_i与权重矩阵$\boldsymbol{W}_{ij}$相乘得到预测向量$\hat{\boldsymbol{u}}_{j|i}$,最后将其与系数$c_{ij}$相乘累加得到$s_j$并进行squash归一化得到$v_j$。

$$\hat{\boldsymbol{u}}_{j|i} = \boldsymbol{W}_{ij} \cdot u_i \tag{11}$$
$$c_{ij} = \mathrm{softmax}(b_{ij}) \tag{12}$$
$$s_j = \sum_i c_{ij} \hat{\boldsymbol{u}}_{j|i} \tag{13}$$
$$s_j = \sum_i c_{ij} \hat{\boldsymbol{u}}_{j|i} \tag{14}$$

得到v_j后，将其与预测向量$\hat{\boldsymbol{u}}_{j|i}$相乘得到路由系数$b_{ij}$。

$$b_{ij} = b_{ij} + \hat{\boldsymbol{u}}_{j|i} \cdot v_j \tag{15}$$

由此可知，更新后的路由系数b_{ij}会影响s_j，进而决定v_j的值（见式（13）、式（14））。在满足迭代次数后，v_j被传入情绪胶囊层。

上述公式可简单地通过向量及其内积进行说明，如图3所示。本文仅取1个胶囊作为示例，权重矩阵$\boldsymbol{W}_{ij}$的j=3，利用式（11）即可得到$\hat{\boldsymbol{u}}_{1|1}$、$\hat{\boldsymbol{u}}_{2|1}$、$\hat{\boldsymbol{u}}_{3|1}$。在每次路由迭代前，需令路由系数$b_{ij}$=0，由Softmax函数可知$c_{ij}$的每个元素值相等，那么$c_{ij}$与$\hat{\boldsymbol{u}}_{1,2,3|1}$相乘不会改变其方向，仅改变值的大小。为了方便画图，先忽略c_{ij}对值大小的影响，则式（13）可用图3的向量相加表示。接下来，对和向量结果取squash归一化即可得到v_1。最后，利用式（15）更新路由系数，由向量的内积可知，两个向量方向一致时，二者内积得到的值最大。

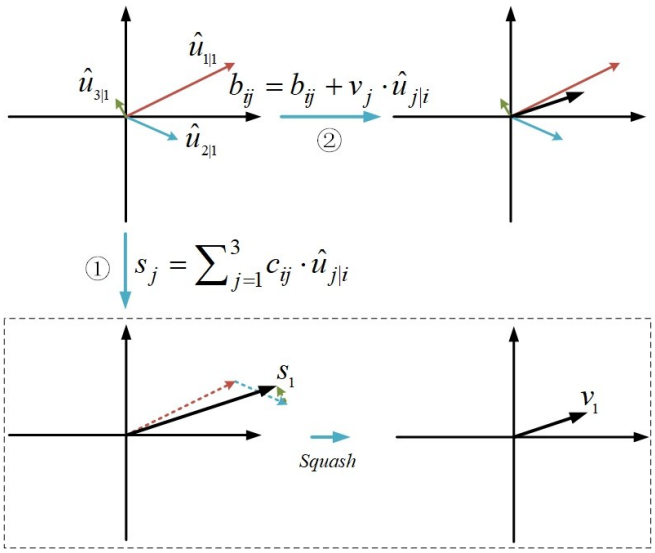


Fig. 3　Example of dynamic routing update algorithm

图3　动态路由更新算法示例

从动态路由更新算法机制而言，更新时会根据“和向量”与各个预测向量之间的相似度来调整向量长度：即相似度高的预测向量会被拉长，相似度低的则变短。这种机制确保了与“和向量”更为相似的预测向量更多地输入情绪胶囊层，有效实现了分类效果。此外，动态路由机制通过向量化表征显著优化特征空间分布，能更好地保留特征的空间关系，对SER这种需要理解特征空间关系的任务至关重要。

本文针对胶囊网络特性，深入研究了动态路由算法中权重矩阵$\boldsymbol{W}_{ij}$的分配机制，如图4所示（左图为每个胶囊独享权重矩阵，右图为共享权重矩阵）。由于为每个胶囊单独分配权重矩阵训练，会导致模型参数剧增，引发过拟合、梯度爆炸/消失等问题。为此，采用共享权重策略，在降低模型复杂度的同时，让多个胶囊连接共享同一组权重，使模型能更准确地把握特征间的内在联系，提升其泛化能力。

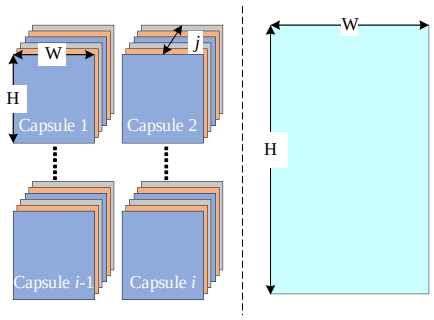


Fig. 4　Weight matrix allocation mechanism

图4　权重矩阵分配机制

1.5　损失函数

本文使用的损失函数为Margin-Loss函数，相较于交叉熵、均方误差等传统损失函数而言，更侧重于样本之间的相对距离与间隔。由于胶囊层输出情绪胶囊层中的v_j都是向量，通过Margin-Loss可最大化不同类别的样本间隔来优化模型。

$$\begin{aligned} L_T &= T_c \cdot \max(0, m^+ - \|\boldsymbol{v}_c\|)^2 \\ L_F &= \lambda(1 - T_c)\max(0, \|\boldsymbol{v}_c\| - m^-)^2 \\ L_c &= L_T + L_F \end{aligned} \tag{16}$$

式中：T_c为将情绪标签经过One-Hot编码后得到的序列；$\boldsymbol{v}_j$表示每个情绪实体的向量表示；$\lambda \in (0, 1)$，本文λ=0.8；m^+、m^-分别为0.9、0.1。

2　实验结果与分析

2.1　实验语料

本文使用藏语语音情感语料库TSEC_4528验证所提模型的性能，同时选用柏林情感语音语料库（Berlin Emotional Speech Database，EMO-DB）、RAVDESS（Speech & Song）两个公开语料库进行补充。TSEC_4528由12名西藏大学卫藏方言母语者（6男6女）的演讲构成，含愤怒、恐惧、开心、中性和悲伤等5类情感。录制时，12名录音人员互相检查音频情感表达情况，以剔除不合格音频，保证语料库的整体质量，具体信息如表2所示。实验表明，TSEC_4528的Kappa一致性评估为0.63，属于高度一致，能为研究提供坚实的数据支撑。

EMO-DB由德国柏林工业大学录制，含5男5女的表

Table 2　TSEC-4528 speech emotion corpus

表2　TSEC_4528语音情感语料库

属性	特点
语言	藏语
方言	卫藏方言
类型	表演型语料库
情绪种类	愤怒、恐惧、开心、中性、伤心
录音人数	6男6女
声调	4个
辅音发音特点	浊辅音清化，复辅音简化
元音发音特点	鼻化元音与非鼻化元音
时长	0.8~4秒不等
采样与量化	16KHz，16bit

演型语料，涵盖生气、无聊、厌恶、恐惧、开心、中性、伤心等7种情绪类型。本文除无聊外，其余情绪均被用以评估模型的性能。RAVDESS（Speech & Song）语料库是语音情感识别领域常用的语料库，由24名专业演员（12男12女）录制而言，含语音、歌曲，歌曲部分全部采纳，语音部分选取了除伤心之外的所有情绪。

2.2 实验环境与评估方法

本文采用NVIDIA GeForce RTX 4070super作为硬件平台，选用TensorFlow作为深度学习框架。所有语料文件均用OpenSMILE提取MFCC特征，数据集按照4∶1的比例划分训练集与验证集，整个训练过程设定为300个轮次，优化器为Adam，学习率为0.000 1。表3为语料的详细数据，R-Speech、R-Song表示RAVDESS的两个不同语料。

Table 3 Corpus emotions and quantity

表3 语料库情绪及数量

	TSEC_4528		EMO-DB		R-Speech		R-Song	
	Total	Val	Total	Val	Total	Val	Total	Val
Angry	713	143	79	16	192	39	184	37
Calm	–	–	–	–	192	38	184	37
Disgust	–	–	46	9	192	39	–	–
Fear	772	154	69	14	192	39	184	37
Happy	989	198	71	14	192	38	184	37
Neutral	1062	213	79	16	96	19	92	18
Sad	992	198	62	13	–	–	184	37
Surprise	–	–	–	–	192	38	–	–

在一般SER任务中，评估模型的指标包含未加权准确率、加权准确率（Weighted Accuracy，UA）

$$\mathrm{UA} = \frac{T}{N} \tag{17}$$

$$\mathrm{WA} = \frac{\sum \lambda_i T_i}{N} \tag{18}$$

式中：N表示总样本数量；T表示预测正确的总样本数量；λ_i表示第i类情绪在总样本中所占的比例；T_i表示第i类情绪中预测正确的样本数量。

为了更详细说明，本文还使用精确度（Precision）、召回率（Recall）和F1-score评估模型性能。

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{19}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{20}$$

$$F1 = \frac{2PR}{P + R} \tag{21}$$

式中：TP表示真正例；FP表示假正例；FN表示假负例。

2.3 实验结果

本文除了模型评价指标外，还与近年来的先进模型进行性能比较，并绘制混淆矩阵（见图5）统计情绪分类情况。此外，使用t-SNE（t-Distributed Stochastic Neighbor Embedding）验证模型的分类能力，如表4、表5所示。其中，P、R表示精准度和召回率；Avg为未加权的平均值。由此可知，在TSEC_4528语料库中，模型的未加权平均召回率为77.91%，F1-score平均值为78.70%，恐惧情绪F1-score表现最佳，为80.00%；在EMO-DB语料库中，模型的未加权平均召回率为95.92%，F1-score平均值为96.10%，Neutral情绪F1-score达100.00%；在RAVDESS的Speech语料中，模型的未加权平均召回率为87.75%，F1-score平均值为87.87%，得分最高的情绪为Calm，为92.68%；在R-Song语料中，所提模型的表现最佳，未加权平均召回率为97.75%，F1-score的平均值为97.73%，Calm、Happy、Neutral情绪的得分均为100%。

Table 4 Performance of models in different datasets

表4 模型在不同数据集中的性能 （%）

Dataset	indicator	Angry	Calm	Disgust	Fear	Happy	Neutral	Sad	Surprise	Avg
TSEC_4528	P	85.37	–	–	90.91	72.69	76.29	76.17	–	80.26
	R	73.43	–	–	71.43	79.29	83.10	82.32	–	77.91
	F1	78.95	–	–	80.00	75.85	79.55	79.13	–	78.70
EMO-DB	P	100.00	–	100.00	93.33	92.86	100.00	92.86	–	96.50
	R	93.75	–	88.89	100.00	92.86	100.00	100.00	–	95.92
	F1	96.77	–	94.12	96.55	92.86	100.00	96.30	–	96.10
R-Speech	P	91.43	86.36	87.18	76.09	96.55	94.44	–	89.74	88.83
	R	82.05	100.00	87.18	89.74	73.68	89.47		92.11	87.75
	F1	86.49	92.68	87.18	82.35	83.58	91.89		90.91	87.87
R-Song	P	94.87	100.00	–	97.14	100.00	100.00	94.59	–	97.77
	R	100.00	100.00	–	91.89	100.00	100.00	94.59	–	97.75
	F1	97.37	100.00	–	94.44	100.00	100.00	94.59	–	97.73

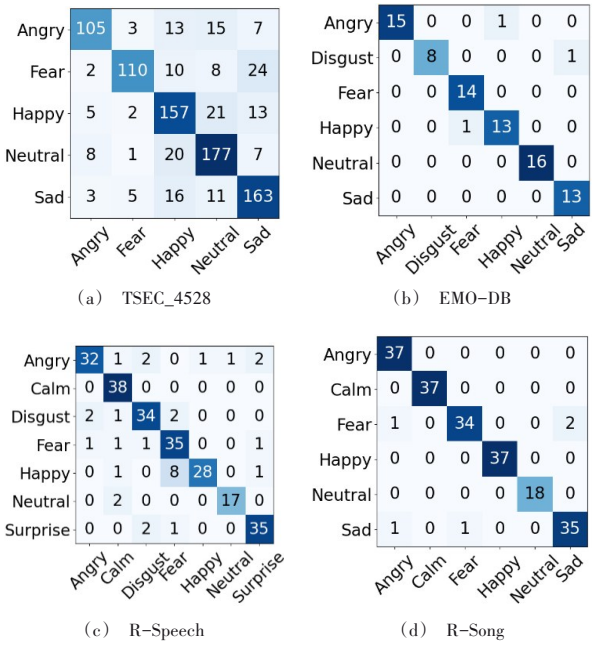
Table 5 UA and WA of the model in different datasets

表5 模型在不同数据集中的UA、WA （%）

Dataset	UA	WA
TSEC_4528	78.58	79.39
EMO-DB	96.34	96.51
R-Speech	87.60	88.35
R-Song	97.54	97.56

由图5可见，在TSEC_4528语料库上分类正确率最高的情绪为neutral，213个测试语料中177个被正确预测，结合召回率而言，该情绪的预料识别表现也为最佳；存在24个fear被预测为sad、20个neutral被预测为happy、21个happy被预测为neutral。从整体性能而言，906个测试样本中712个被模型正确预测，整体UA达78.58%。

在EMO-DB语料库中，模型UA达96.34%，仅3个语料



（a） TSEC_4528　（b） EMO-DB　（c） R-Speech　（d） R-Song

Fig. 5　Confusion matrix statistics

图5　混淆矩阵统计情况

被错误分类，1个angry被预测为happy，1个disgust被当作sad，1个happy误判为fear。

在R-Speech语料中，模型在情绪识别任务中性能优异，模型UA达87.60%，calm情绪的所有样本均被正确预测，happy表现最差，8个被预测为fear。在R-Song语料中，模型表现最佳，UA达97.54%，203个测试语料中198个被正确分类，2个fear被预测为sad，1个fear语料被误判为angry，2个sad情绪被分别认为是fear、angry。

为了分析模型的泛化能力，将所提模型与近年的先进模型进行比较，如表6所示。由此可知，在不同语言的数据集上，所提模型均取得了较高的WA、UA，展现出一定的跨语言适应性。

Table 6　Performance of the proposed model on different datasets compared to existing models

表6　所提模型与现有模型在不同数据集上的性能

Dataset	Model	Year	WA/%	UA/%
EMO-DB	Spatiotemporal and frequential cascaded attention Networks[26]	2021	83.30	82.10
	TIM-Net[27]	2022	89.30	91.09
	CNN-BiLSTM[28]	2023	80.60	82.60
	Newman-Watts-Strogatz[29]	2024	87.14	87.89
	BiLSTM-AlexNet-WavLM[30]	2025	92.61	91.68
	ST-CapsNet（Ours）	2025	**96.51**	**96.34**
R-Speech	ESN[31]	2021	74.54	74.23
	CNN-BiLSTM[28]	2023	64.80	64.20
	Newman-Watts-Strogatz[29]	2024	76.45	75.79
	IGRFXG[32]	2025	–	79.28
	ST-CapsNet（Ours）	2025	**88.35**	**87.60**
R-Song	L^3-NET[33]	2021	–	71.00
	VQ-MAE-S[34]	2023	–	84.20
	MERT[35]	2025	–	84.73
	ST-CapsNet（Ours）	2025	**97.56**	**97.54**

同时，为了进一步说明模型的分类能力，采用t-SNE技术对测试集特征空间进行可视化分析。将测试语料的输入特征进行降维分析，与其输出值进行比较，如图6、图7所示。其中，不同颜色的样本点对应不同情绪。由此可见，原始语音输入经t-SNE降维后，各情绪样本点呈弥散分布，缺乏明显的聚类结构，而将模型前向传播后的输出值进行t-SNE降维后，可使不同情绪形成高密度聚类簇。

（a） TSEC_4528　（b） EMO-DB　（c） R-Speech　（d） R-Song

Fig. 6　T-SNE results of input features

图6　输入特征的t-SNE结果

（a） TSEC_4528　（b） EMO-DB　（c） R-Speech　（d） R-Song

Fig. 7　t-SNE result graph output by the model

图7　模型输出的t-SNE结果

2.4 消融实验配置与模型复杂度

为了研究各模块对模型的有效性和复杂度，在数据集TSEC_4528、EMO-DB、RAVDESS(Speech & Song)上对ST-CapsNet进行消融实验(见表7)，并在EMO-DB数据集上统计模型参数量(Params)、浮点运算数(Floating Point Operations，FLOPs)，以评估模型复杂度(见表8)。其中，Exp 1，Exp 2、Exp3分别表示移除时空特征提取器的残差LSTM模块、4层卷积模块、Capsule Network模块后的模型。

Table 7 Results of ablation experiment

表7 消融实验结果 (%)

Model	TSEC_4528		EMO-DB		R-Speech		R-Song	
ST-CapsNet	78.58		96.34		87.60		97.54	
Exp 1	76.49	↓2.09	89.02	↓7.32	85.60	↓2.00	96.06	↓1.48
Exp 2	65.34	↓13.24	75.61	↓20.73	53.60	↓34.00	82.86	↓14.68
Exp 3	71.85	↓6.73	74.39	↓21.95	72.00	↓15.60	92.12	↓5.42

Table 8 Model complexity

表8 模型复杂度

Model	SFE		Capsule Network	Params/M	FLOPs/G
	LSTM	4-Layer CNN			
ST-CapsNet	√	√	√	1.771	0.415
Exp 1	–	√	√	1.755	0.414
Exp 2	√	–	√	0.043	0.008
Exp 3	√	√	–	3.732	0.372

由表7可知，缺失残差LSTM使模型的UA平均下降3.22%，验证了残差LSTM可有效捕获语音信号的动态变化和时序的依赖关系；CNN模块的缺失导致模型UA平均下降20.66%，印证其作为空间特征提取核心组件的不可替代性，尤其在提取关键局部特征中具有决定性作用；移除胶囊网络会引发UA平均下降12.43%，证明了动态路由机制通过向量化表征可显著优化特征空间分布，提升项目在跨语言场景下的相似情感区分能力。

需要注意的是，胶囊网络动态路由机制的多次迭代计算虽导致FLOPs略微增加，但能更有效地利用特征空间，使模型参数量减少52.55%。ST-CapsNet以单一45维MFCC特征实现轻量化设计，模型参数量仅1.771 M，浮点运算量为0.415 G，在跨语言任务中性能优势明显，验证了时空特征提取器与胶囊网络的协同优化，可有效平衡模型性能与泛化能力。

3 结语

本文聚焦藏语语音情感识别领域，针对当前特征提取方法复杂、特征维度过高的问题，构建了一个轻量级的语音情感识别模型ST-CapsNet。首先，模型仅以45维MFCC作为输入特征，通过时空特征提取器捕获MFCC特征的时空信息；其次，结合胶囊网络的动态路由更新算法进行深度分析；最后，以Margin-Loss指导模型更新参数。

在藏语语音情感语料库TSEC_4528与公开语料库EMO-DB、RAVDESS的实验表明，所提模型的性能优势明显，结合各指标与t-SNE可视化分析发现，所提模型取得了令人满意效果，相较于现有模型优势突出。未来，将结合深度学习与机器学习提出新的算法与模型，探索降低胶囊网络训练复杂度的新方法，以更好地处理语音情感识别任务。

参考文献：

[1] WANI T M, GUNAWAN T S, QADRI S A A, et al. A comprehensive review of speech emotion recognition systems[J]. IEEE Access, 2021, 9: 47795-47814.

[2] AL-DUJAILI M J, EBRAHIMI-MOGHADAM A. Speech emotion recognition: a comprehensive survey[J]. Wireless Personal Communications, 2023, 129(4): 2525-2561.

[3] SINGH Y B, GOEL S. A systematic literature review of speech emotion recognition approaches[J]. Neurocomputing, 2022, 492: 245-263.

[4] MADANIAN S, CHEN T, ADELEYE O, et al. Speech emotion recognition using machine learning —— a systematic review[J]. Intelligent Systems with Applications, 2023, 20: 200266.

[5] AKÇAY M B, OĞUZ K. Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers[J]. Speech Communication, 2020, 116: 56-76.

[6] KUMAR S, HAQ M A, JAIN A, et al. Multilayer neural network based speech emotion recognition for smart assistance[J]. Computers, Materials and Continua, 2022, 74(1): 1523-1540.

[7] LATIF S, RANA R, KHALIFA S, et al. Survey of deep representation learning for speech emotion recognition[J]. IEEE Transactions on Affective Computing, 2021, 14(2): 1634-1654.

[8] AOUANI H, AYED Y B. Speech emotion recognition with deep learning[J]. Procedia Computer Science, 2020, 176: 251-260.

[9] GARCIA-CUESTA E, SALVADOR A B, PÃEZ D G. EmoMatchSpanishDB: study of speech emotion recognition machine learning models in a new Spanish elicited database[J]. Multimedia Tools and Applications, 2024, 83: 13093-13112.

[10] LOPE J D, GRAÑA M. An ongoing review of speech emotion recognition[J]. Neurocomputing, 2023, 528: 1-11.

[11] ZHAO H J, YE N, WANG R C. Improved cross-corpus speech emotion recognition using deep local domain adaptation[J]. Chinese Journal of Electronics, 2023, 32(3): 640-646.

[12] NING J, ZHANG W. Speech-based emotion recognition using a hybrid RNN-CNN network[J]. Signal, Image and Video Processing, 2025, 19: 124.

[13] BARHOUMI C, BENAYED Y. Real-time speech emotion recognition using deep learning and data augmentation[J]. Artificial Intelligence Review, 2025, 58(2): 49.

[14] AL-SELWI S M, HASSAN M F, ABDULKADIR S J, et al. RNN-LSTM: from applications to modeling techniques and beyond—systematic review[J]. Journal of King Saud University - Computer and Information Sciences, 2024, 36(5): 102068.

[15] CHEN W D, XING X F, XU X M, et al. DST: deformable speech transformer for emotion recognition[C]// 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, 2023: 1-5.

[16] WEN X C, YE J X, LUO Y, et al. CTL-MTNet: a novel capsnet and transfer learning-based mixed task net for the single-corpus and cross-

corpus speech emotion recognition[C]// International Joint Conference on Artificial Intelligence, 2022: 2305-2311.

[17] PENG M Z X, CAI Z J, CAI R Z M. Construction of a Tibetan emotional speech database [J]. Acta Scientiarum Naturalium Universitatis Pekinensis (Natural Science Journal of Peking University), 2023, 59(5): 773-781.
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.

[18] GUO D, YU H, HU A, et al. Statistical analysis of acoustic characteristics of Tibetan Lhasa dialect speech emotion[C]// 2015 3rd International Conference on Information Technology and Career Education, 2015: 106-110.

[19] CI R L Z. Research on Tibetan speech emotion recognition methods[D]. Lhasa: Tibet University, 2019.
次仁罗增. 藏语语音情感识别方法研究[D]. 拉萨:西藏大学,2019.

[20] CAI Y X. Research on Tibetan speech emotion recognition method based on deep learning[D]. Lhasa: Tibet University, 2024.
蔡优新. 基于深度学习的藏语语音情感识别方法研究[D]. 拉萨:西藏大学,2024.

[21] WANG X. Research on Tibetan speech emotion recognition methods based on multi-feature fusion[D]. Lhasa: Tibet University, 2023.
王希. 基于多特征融合的藏语语音情感识别方法研究[D]. 拉萨: 西藏大学,2023.

[22] GU Z Y, BIAN B W D, QI J D. Tibetan speech emotion recognition based on multi-feature fusion[J]. Modern Electronic Technology, 2023, 46(21): 129-133.
谷泽月,边巴旺堆,祁晋东. 基于多特征融合的藏语语音情感识别[J]. 现代电子技术,2023,46(21): 129-133.

[23] PENG M Z X. Research on Tibetan speech emotion recognition technology integrating multi-feature fusion[D]. Xining: Qinghai Normal University, 2023.
彭毛扎西. 融合多特征的藏语语音情感识别技术研究[D]. 西宁:青海师范大学,2023.

[24] QU A T. A study of tibetan vowels[M]. Xining: Qinghai Nationalities Publishing House, 1991.
瞿霭堂. 藏语韵母研究[M]. 西宁: 青海民族出版社,1991.

[25] ZHANG H, HUANG H, ZHAO P, et al. CENN: capsule-enhanced neural network with innovative metrics for robust speech emotion recognition[J]. Knowledge-Based Systems, 2024, 304: 1124

[26] LI S Z, XING X F, FAN W Q, et al. Spatiotemporal and frequential cascaded attention networks for speech emotion recognition[J]. Neurocomputing, 2021, 448: 238-248.

[27] YE J X, WEN X C, WEI Y J, et al. Temporal modeling matters: a novel temporal emotional modeling approach for speech emotion recognition [C]// 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, 2023: 1-5.

[28] BAEK J Y, LEE S P. Enhanced speech emotion recognition using DCGAN-Based data augmentation[J]. Electronics, 2023, 12(18): 3966.

[29] SOLTANI R, BENMOHAMED E, LTIFI H. Newman-Watts-Strogatz topology in deep echo state networks for speech emotion recognition[J]. Engineering Applications of Artificial Intelligence. 2024, 133:108293.

[30] ZHANG W Z, LI H Y, YANG H W. Tibetan-speech-emotion recognition under low-resource conditions [J]. Journal of Signal Processing, 2025, 41(9): 1558-1569.
张维昭,李皓渊,杨鸿武. 低资源条件下的藏语语音情感识别[J]. 信号处理,2025,41(9):1558-1569.

[31] IBRAHIM H, LOO C K, ALNAJJAR F. Speech emotion recognition by late fusion for bidirectional reservoir computing with random projection [J]. IEEE Access, 2021, 9: 122855-122871.

[32] TRIPATHI A, RANI P. Multilingual speech emotion recognition using IGRFXG-Ensemble feature selection approach[J]. Applied Acoustics, 2025, 240: 110905.

[33] KOH E, DUBNOV S. Comparison and analysis of deep audio embeddings for music emotion recognition[C]// Proceedings of the 4th Workshop on Affective Content Analysis, 2021: 15-22.

[34] SADOK S, LEGLAIVE S, SÉGUIER R. A vector quantized masked autoencoder for speech emotion recognition[C]// 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops, 2023: 1-5.

[35] SUN Y, ZHAO Z, RICHMOND Z, et al. Exploring acoustic similarity in emotional speech and music via self-supervised representations [C]// 2025 IEEE International Conference on Acoustics, Speech and Signal Processing, 2025: 1-5.

(责任编辑:刘嘉文)