文章编号：1673-3193(2025)01-0019-08　　J. North Univ. China, Nat. Sci. Ed., 2025, 46(1): 19-26.

基于自监督预训练模型和NWCE的口吃语音分类

殷志鹏，徐新洲

（南京邮电大学 物联网学院，江苏 南京 210003）

摘　要：口吃语音分类旨在利用语音信号对不同口吃类别进行分类识别，而现有相关研究没有充分考虑自监督预训练模型表示嵌入的时序特性，且只简单地表征了口吃语音数据的类别不平衡性。为此，本文提出一种基于自监督预训练模型和非线性加权交叉熵(NWCE)损失的口吃语音分类方法。该方法首先利用自监督预训练模型提取副语言表示嵌入，然后通过带自注意力机制的双向长短期记忆网络模型，捕捉嵌入中显著的时序特征和上下文信息，最后利用非线性加权交叉熵损失来关注样本较少的口吃语音类别。在口吃语音分类数据集上的实验结果表明，本文方法通过学习语音中自监督预训练模型多层表示嵌入的时序信息，并且通过NWCE充分描述了各口吃类别数据间的关系，取得了比现有方法更好的口吃语音分类性能。

关键词：计算副语言；口吃语音分类；自监督预训练模型；非线性加权交叉熵损失

中图分类号：TP183　　**文献标识码**：A　　**doi:** 10.62756/jnuc.issn.1673-3193.2023.09.0002

引用格式：殷志鹏，徐新洲．基于自监督预训练模型和NWCE的口吃语音分类[J]．中北大学学报(自然科学版)，2025，46(1)：19-26.

YIN Zhipeng，XU Xinzhou. Stuttering speech classification based on self-supervised pre-trained model and NWCE[J]. Journal of North University of China(Natural Science Edition)，2025，46(1)：19-26.

Stuttering Speech Classification Based on Self-Supervised Pre-Trained Model and NWCE

YIN Zhipeng，XU Xinzhou

（School of Internet of Things，Nanjing University of Posts and Telecommunications，Nanjing 210003，China）

Abstract：Stuttering speech classification aims to classify and recognize different categories of stuttering using spoken signals. Nevertheless，the existing related works fail to sufficiently focus on sequential characteristics for the representation embedding of self-supervised pre-trained models，and these works also simplistically address the class-imbalance issue for stuttering-speech data. In this regard，we proposed a stuttering speech classification approach based on self-supervised pre-trained models and nonlinear weighted cross-entropy （NWCE） loss. Within the proposed approach，we first employed a self-supervised pre-trained model to extract paralinguistic representation embeddings from stuttering speech. Then，we utilized a bidirectional long short-term memory network model with a self-attention mechanism to capture essential temporal features and contextual information within the embeddings. Afterwards，a nonlinear weighted cross-entropy loss was performed to focus on stuttering speech categories with fewer samples. The experimental results on stuttering speech classification dataset indicate that，the proposed

收稿日期：2023-09-05

基金项目：中国博士后科学基金面上项目(2022M711693)；国家自然科学基金面上项目(62071242，62172235)；南京邮电大学校级自然科学基金(NY222158)

作者简介：殷志鹏(1999－)，男，硕士生，主要从事语音信号处理方面的研究。

通信作者：徐新洲(1987－)，男，副教授，博士，主要从事智能信号处理方面的研究。E-mail：xinzhou.xu@njupt.edu.cn。

approach achieves better performance for classifying stuttering speech compared with state-of-the-art approaches, through learning the sequential information from self-supervised pre-trained models' multi-layer representation embedding in speech, and sufficiently describes the relationship between the data of different stuttering categories by using NWCE.

Key words: computational paralinguistics; stuttering speech classification; self-supervised pre-trained model; nonlinear weighted cross-entropy loss

0 引 言

计算副语言学[1]是计算机科学和语言学的交叉领域，旨在研究语音信号中的非言语特性及其与言语信息的关系，目前已广泛应用于语音情绪识别[2]、阿尔茨海默症检测[3]以及言语障碍诊断[4]等领域。作为一种常见的言语障碍，口吃主要涉及说话时的流畅性问题[5]，例如重复单词或音节、延长声音等。此外，口吃语段也包含一些副语言特征，例如无意义的填充词等。

早期相关研究主要基于声学特征的提取和分析，并采用了深度学习基础模型。Sheikh等[6]利用梅尔频率倒谱系数(Mel-scale Frequency Cepstral Coefficients，MFCC)和时延神经网络(Time Delay Neural Network，TDNN)描述口吃语音的时序信息。Kourkunakis等[7]则使用频谱图作为输入，结合深度残差网络(Residual Networks，ResNets)和双向长短期记忆网络(Bidirectional Long Short Term Memory Networks，Bi-LSTM)，对口吃语音进行了多分类。然而，由于口吃病例数量有限、个人隐私的保护以及获取标准化数据的难度较大等因素，口吃数据样本稀少[8]，导致无法很好地捕捉口吃语音中的副语言信息[9]。于是，Grósz等[10]将自监督预训练模型应用于口吃语音分类，得到了更通用的副语言表示嵌入。

目前的口吃语音分类研究仍存在一些不足。首先，现有的相关研究仅关注自监督预训练模型所提取特征嵌入的统计特性，而忽略其时序特征和上下文信息。其次，口吃数据集通常存在类别不平衡的情况，这会导致模型的性能欠佳，而Sheikh等[11]提出的加权交叉熵损失(Weighted Cross-Entropy Loss，WCE Loss)缺乏对口吃语音类别间关系的充分描述。

本文将自监督预训练模型和非线性加权交叉熵损失(Nonlinear WCE Loss，NWCE Loss)相结合用于口吃语音分类。该方法利用自监督预训练模型提取副语言表示嵌入，通过带自注意力机制的双向长短期记忆(Bi-LSTM with Self-Attention，BLSA)网络模型学习该嵌入中有效的时序和上下文信息，并应用非线性加权交叉熵损失关注样本较少的口吃类别，对类平衡权重进行非线性的尺度变换，以解决数据的类别不平衡问题。

1 相关工作

1.1 口吃语音分类

口吃语音蕴含丰富的副语言信息，大多数相关研究使用MFCC、频谱图或其变体作为特征提取方法[6]，其中Jouaiti等[12]结合了MFCC特征和音素概率，使用Bi-LSTM网络实现了口吃检测。Kourkunakis等[7]使用频谱图作为输入特征，结合ResNets和Bi-LSTM进行特征提取和时序处理；FluentNet模型[13]则采用相同的输入特征，但使用了SE(Squeeze-and-Excitation)ResNets来学习更有效的特征表示，并在Bi-LSTM处理之后添加了全局注意力机制。

1.2 自监督预训练模型

自监督预训练模型通过自监督学习方法，在大量未标记的数据集上进行训练[14]，从而无需完全依赖人工标注的数据集，有效解决了数据稀少的问题。在语音信号处理领域中，wav2vec 2.0是目前应用最广泛的自监督预训练模型之一，通过对大量未标注的语音数据进行预训练，广泛应用于下游语音处理任务，如情绪识别[15]、说话人识别[16]、认知障碍检测[17]等。

自监督预训练模型在口吃语音分类中有两种应用方法。1）在目标任务的数据集上对自监督预训练模型进行微调[18-19]。Grósz等[10]针对口吃任务对自监督预训练模型进行微调，但微调大型模型需要更多的存储及更长的训练时间[20]。2）将自监督预训练模型作为静态特征提取器，对提取的嵌入，在时间维度上使用各种统计[21]进行池化，送入下游分类器。例如Bayerl等[22]将提取的

嵌入在时间维度进行平均池化和降维处理；Sheikh等[11]计算了嵌入的平均值和标准差，并对不同层的嵌入进行求和，更好地利用了自监督预训练模型的多层特征表示能力；Montacié等[23]进一步探究了对口吃语音最具区分性的特征处理策略，对嵌入信息在时间维度上使用了多个统计泛函，得到新特征集并进行了组合筛选。

2　本文方法

本文方法的结构如图1所示，包含自监督预训练模型的嵌入提取、带自注意力机制的双向长短期记忆网络，以及非线性加权交叉熵损失等三部分。

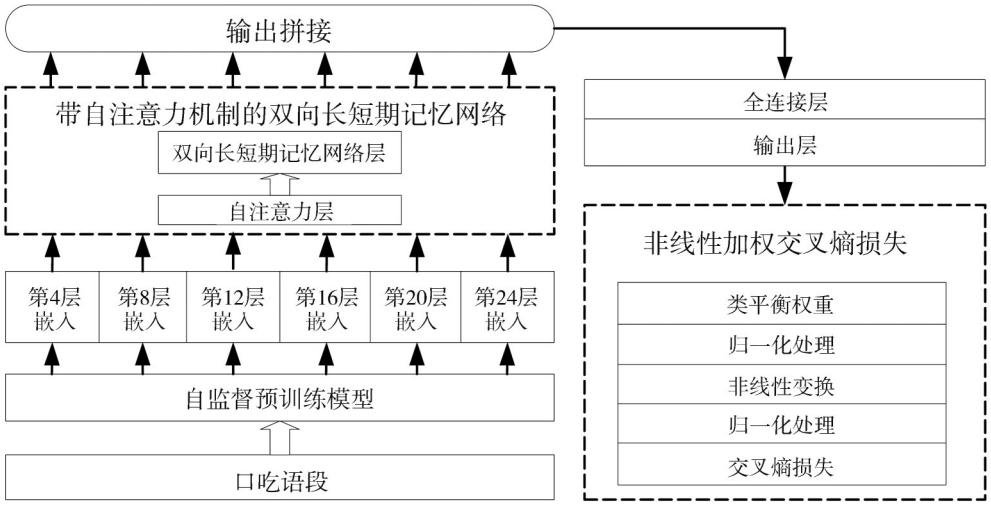


图1　本文基于自监督预训练模型和非线性加权交叉熵损失的口吃语音分类系统结构图

Fig. 1　A diagrammatic overview of the methodology in this work for stuttering speech classification system using self-supervised pre-trained models and nonlinear weighted cross-entropy loss

首先，使用自监督预训练模型对口吃语音进行嵌入提取；接着，使用带自注意力机制的双向长短期记忆网络模型，通过对自监督预训练模型嵌入进行上下文信息建模，实现对口吃语音的分类；最后，结合非线性加权交叉熵损失，进一步关注较少样本类别的损失。

2.1　带自注意力机制的双向长短期记忆网络模型

本文使用自监督预训练模型对任一样本进行嵌入提取，得到了更加通用的副语言表示嵌入$\boldsymbol{x}^{(i)}\in\mathrm{R}^{T\times n_d}$，其中，$\boldsymbol{x}^{(i)}$为自监督预训练模型第$i$层嵌入，$T$为所提取嵌入的时间步长，$n_d$为自监督预训练模型的嵌入维度。本文选择第$i$（$i$取值为4,8,12,16,20,24）层嵌入，分别送入带自注意力机制的双向长短期记忆网络模型。

本文采用一个使用缩放点积注意力的自注意力层，作为带自注意力机制的双向长短期记忆网络模型的第一层，以减少冗余和无关的输入信息。将任一样本的第i层嵌入$\boldsymbol{x}^{(i)}$馈送到自注意力层，从而得到对应输出

$$\boldsymbol{O}^{(i)}=\mathrm{softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{m}}\right)\boldsymbol{V}\in\mathrm{R}^{T\times m},\tag{1}$$

式中：softmax(•)表示Softmax函数。自注意力层将$\boldsymbol{x}^{(i)}$线性映射为查询（Query）向量$\boldsymbol{Q}$、键（Key）向量$\boldsymbol{K}$、值（Value）向量$\boldsymbol{V}$。

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{x}^{(i)}\boldsymbol{W}^{(Q)}\in\mathrm{R}^{T\times m}\\ \boldsymbol{K}=\boldsymbol{x}^{(i)}\boldsymbol{W}^{(K)}\in\mathrm{R}^{T\times m},\\ \boldsymbol{V}=\boldsymbol{x}^{(i)}\boldsymbol{W}^{(V)}\in\mathrm{R}^{T\times m}\end{cases}\tag{2}$$

式中：$\boldsymbol{W}^{(Q)}$，$\boldsymbol{W}^{(K)}$，$\boldsymbol{W}^{(V)}\in\mathrm{R}^{n_d\times m}$分别为映射参数矩阵；$m$为映射维度。

将任一样本的第i层嵌入对应的自注意力层输出$\boldsymbol{O}^{(i)}$馈送到双向长短期记忆网络层，以对表示嵌入中的时序信息和上下文信息进行建模。故设$\boldsymbol{O}^{(i)}$在时间t对应的输出$\boldsymbol{O}_t^{(i)}\in\mathrm{R}^{m\times 1}$。

长短期记忆网络（LSTM）单元的遗忘门$\boldsymbol{f}_t$、输入门$\boldsymbol{z}_t$、输出门$\boldsymbol{u}_t$和候选状态$\tilde{\boldsymbol{c}}_t$根据当前时间步t的输入$\boldsymbol{O}_t^{(i)}$和前一时间步$t-1$的隐藏状态$\boldsymbol{h}_{t-1}\in\mathrm{R}^{n_c\times 1}$生成相应的输出，分别可表示为

$$\boldsymbol{f}_t=\sigma\left(\boldsymbol{W}^{(f)}\left[(\boldsymbol{O}_t^{(i)})^{\mathrm{T}},(\boldsymbol{h}_{t-1})^{\mathrm{T}}\right]^{\mathrm{T}}+\boldsymbol{b}^{(f)}\right),\tag{3}$$

$$\boldsymbol{z}_t=\sigma\left(\boldsymbol{W}^{(z)}\left[(\boldsymbol{O}_t^{(i)})^{\mathrm{T}},(\boldsymbol{h}_{t-1})^{\mathrm{T}}\right]^{\mathrm{T}}+\boldsymbol{b}^{(z)}\right),\tag{4}$$

$$\boldsymbol{u}_t=\sigma\left(\boldsymbol{W}^{(u)}\left[(\boldsymbol{O}_t^{(i)})^{\mathrm{T}},(\boldsymbol{h}_{t-1})^{\mathrm{T}}\right]^{\mathrm{T}}+\boldsymbol{b}^{(u)}\right),\tag{5}$$

$$\tilde{\boldsymbol{c}}_t=\tanh\left(\boldsymbol{W}^{(\tilde{c})}\left[(\boldsymbol{O}_t^{(i)})^{\mathrm{T}},(\boldsymbol{h}_{t-1})^{\mathrm{T}}\right]^{\mathrm{T}}+\boldsymbol{b}^{(\tilde{c})}\right),\tag{6}$$

式中：$\sigma(\cdot)$表示Sigmoid函数；$\tanh(\cdot)$表示双曲正切函数；网络线性权重$\boldsymbol{W}^{(f)}$，$\boldsymbol{W}^{(z)}$，$\boldsymbol{W}^{(u)}$，$\boldsymbol{W}^{(\tilde{c})}\in\mathbf{R}^{n_e\times(m+n_e)}$和偏置$\boldsymbol{b}^{(f)},\boldsymbol{b}^{(z)},\boldsymbol{b}^{(u)},\boldsymbol{b}^{(\tilde{c})}\in\mathbf{R}^{n_e\times1}$是可学习的参数，$n_e$为LSTM的映射维度。进而LSTM单元根据$\boldsymbol{f}_t$和$\boldsymbol{z}_t$的输出生成状态

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{z}_t\odot\tilde{\boldsymbol{c}}_t,\tag{7}$$

式中：$\odot$为向量元素乘积。Bi-LSTM结合了前向隐藏输出$\vec{\boldsymbol{h}}_t$和后向隐藏输出$\overleftarrow{\boldsymbol{h}}_t$，得到当前时间$t$的输出为

$$\boldsymbol{H}_t=\left[(\vec{\boldsymbol{h}}_t)^{\mathrm{T}},(\overleftarrow{\boldsymbol{h}}_t)^{\mathrm{T}}\right]^{\mathrm{T}}\in\mathbf{R}^{2n_e\times1},\tag{8}$$

其中任一方向隐藏输出

$$\boldsymbol{h}_t=\boldsymbol{u}_t\odot\tanh(\boldsymbol{c}_t)\in\mathbf{R}^{n_e\times1}。\tag{9}$$

本文将所有时刻的隐藏输出$\boldsymbol{H}_t$构成Bi-LSTM的输出，则得到任一样本的第i层嵌入所对应的长短期记忆网络层输出$\boldsymbol{O}_{\mathrm{lstm}}^{(i)}=\left[\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_T\right]^{\mathrm{T}}\in\mathbf{R}^{T\times2n_e}$。

2.2 非线性加权交叉熵损失

分别将任一样本的第i(i取值为4,8,12,16,20,24)层嵌入输入到带自注意力机制的双向长短期记忆模型，并将所有输出在空间维度上进行拼接，得到任一样本的融合输出$\boldsymbol{O}_{\mathrm{con}}=\left[\boldsymbol{O}_{\mathrm{lstm}}^{(4)},\boldsymbol{O}_{\mathrm{lstm}}^{(8)},\cdots,\boldsymbol{O}_{\mathrm{lstm}}^{(24)}\right]\in\mathbf{R}^{T\times12n_e}$。将$\boldsymbol{O}_{\mathrm{con}}$送入以高斯误差线性单元(Gaussian Error Linear Units，GELU)为激活函数的全连接层，进一步降低到所需的维度，输出

$$\boldsymbol{O}_{\mathrm{fc}}=\boldsymbol{O}_{\mathrm{con}}\boldsymbol{W}_{\mathrm{fc}}+\boldsymbol{b}_{\mathrm{fc}}\in\mathbf{R}^{T\times d},\tag{10}$$

式中：$\boldsymbol{W}_{\mathrm{fc}}\in\mathbf{R}^{12n_e\times d}$和$\boldsymbol{b}_{\mathrm{fc}}\in\mathbf{R}^{T\times d}$分别为全连接层的映射矩阵和偏置；$d$为全连接层的节点数。

将$\boldsymbol{O}_{\mathrm{fc}}$在时间维度上进行平均池化得$\boldsymbol{O}_{\mathrm{avg}}\in\mathbf{R}^{1\times d}$。最后将$\boldsymbol{O}_{\mathrm{avg}}$送入输出层，得到任一样本对应的每个口吃类别的预测输出$\boldsymbol{y}=\left[y_1,y_2,\cdots,y_c\right]\in\mathbf{R}^{1\times c}$表示为

$$\boldsymbol{y}=\boldsymbol{O}_{\mathrm{avg}}\boldsymbol{W}+\boldsymbol{b},\tag{11}$$

式中：$\boldsymbol{W}\in\mathbf{R}^{d\times c}$和$\boldsymbol{b}\in\mathbf{R}^{1\times c}$分别为输出层的映射矩阵和偏置；$c$为口吃语音类别数。

针对口吃语音数据集不平衡的情况，为了关注到样本较少的口吃类别和平衡各个口吃类别的损失，本文对不同类别的损失赋予不同的权重，根据各个类别的训练样本数量N_j获得各个类别的类平衡权重w_j，并进行归一化处理使得任一类别的类平衡权重$w_j\in[0,1]$。类平衡权重

$$w_j=\frac{\dfrac{N}{cN_j}}{\displaystyle\sum_{j=1}^{c}\frac{N}{cN_j}},\tag{12}$$

式中：N为训练集的样本数量。

由于各类别样本数差距过大，线性的WCE损失缺乏对各类间关系的充分描述。因此，为更准确地衡量口吃数据集中类别间的权重，本文提出NWCE损失。该方法通过对类平衡权重进行非线性的尺度变换，将较高的权重映射到较低尺度，缩小权重之间差距，得到归一化的非线性加权权重

$$v_j=\frac{k^{w_j}-1}{\displaystyle\sum_{j=1}^{c}(k^{w_j}-1)},\tag{13}$$

式中：k为非线性映射系数。

非线性加权交叉熵损失函数为

$$L=\frac{1}{N}\sum_{i=1}^{N}\frac{v_{d_i}\log\left(\dfrac{\mathrm{e}^{y_{d_i}}}{\displaystyle\sum_{j=1}^{c}\mathrm{e}^{y_j}}\right)}{\displaystyle\sum_{j=1}^{c}v_j},\tag{14}$$

式中：d_i为第i个样本的真实类别标签索引；v_{d_i}为第i个样本所属类别的非线性加权权重；y_{d_i}为第i个样本所属类别的预测输出。

3 实 验

3.1 实验准备

3.1.1 数据集KSF-C

实验使用Kassel流利度挑战数据集(Kassel State of Fluency Corpus，KSF-C)[24-25]，该数据集包含37名德语使用者的5 597个语音片段，各语音片段长度为3 s，总时长4.6 h。3位标注者将所有语音片段分为6个口吃相关类别，包括语言阻滞(Block)、音节延长(Prolongation，简称Prolong)、声音重复(Sound Repetition)、单词重复(Word Repetition)、修改语音技术(Modified Speech Technique，简称Modified)、填充词(Interjection，又称Fillers)。为确保标注的准确性，删除了所有多标签片段，最终得到4 601个片段。最终数据集类别包括6个口吃相关类别和无口吃表现(No Disfluency)类以及垃圾类(Garbage)，其

中，垃圾类表示无法理解、不包含语音或受背景噪声影响的片段，详细信息如表 1 所示。

表 1　KSF-C 数据集类别信息及其他信息详表

Tab. 1　Category information and other details of KSF-C dataset

类别信息			其他信息	
类别	训练集样本数	开发集样本数	样本数量	音频采集
Block	310	102	训练集样本2 489个，开发集样本982个	单声道，时长3 s，采样率16 kHz
Fillers	205	104		
Garbage	52	33		
Modified Prolong	687	185		
Sound Repetition	183	53		
Word Repetition	169	38		
No Disfluencies	53	23		
	830	444		

3.1.2　音频特征表示

wav2vec 2.0 是一种从原始音频数据中提取特征表示的自监督学习模型。在预训练阶段，wav2vec 2.0使用自监督学习方法在大量未标记的音频数据集上学习语音信号的特征表示。为了适应不同的语音识别任务，可以通过连接时序分类（Connectionist Temporal Classification，CTC）损失，在目标任务数据集上进行微调。

本文选择在 Common Voice[26]数据集（多语言，9 283 h）上进行预训练的 wav2vec 2.0大型模型[10]用于提取音频表示，并将该模型针对德语语音识别任务进行微调。本文选择该模型的原因是其可以针对德语音频语段进行识别，且与KSF-C数据集的语言一致，能更好地适应目标任务。

3.1.3　超参数设置

基于深度学习现有的经验，本文方法的实验模型参数设置为：自监督预训练模型嵌入维度 n_d 为1 024，自注意力层映射维度 m 为1 024，双向长短期记忆层的映射维度 n_e 为256，全连接层节点数 d 取值为 256。训练时选用 Adam（Adaptive moment estimation）优化器，批大小（Batch Size）为 64，最大训练轮数（Epoch）为 100，初始学习率（Initial Learning Rate）的取值范围为{0.000 5, 0.000 1, 0.000 05, 0.000 01, 0.000 005}，最后记录最佳结果。

3.2　实验结果

3.2.1　本文方法

使用本文BLSA方法，考虑焦点损失函数和加权交叉熵损失以及非线性加权交叉熵损失，进行5次重复实验，记录本文方法在不同损失函数下，开发集的最大未加权平均召回率（Unweighted Average Recall，UAR）和其对应的宏 $F1$ 分数（macro F1-Score）结果，如表 2 所示。其中，对于Focal损失的参数设置，可调焦点参数取值为2.0，类别权重采用WCE中的类平衡权重；非线性加权交叉熵损失的参数 k 取值范围为{1.5, 2.0, 2.5, 3.0}。Focal 损失下的 UAR 最大值高于 WCE 损失下的最大值。所有 NWCE 损失下的 UAR 最大值（$R_{\text{UAR-max}}$）均高于 WCE 损失下的最大值，这表明在口吃识别模型中，使用NWCE损失可以取得更好的性能表现。当 k 取值为 2.0 时，口吃语音分类模型的性能最好。进一步给出不同 k 值的 NWCE 损失的 UAR 平均值（$R_{\text{UAR-avg}}$）及其标准差，结果如图 2 所示，可以看出，各 k 值对应的 UAR 平均值在大部分情况下都高于 WCE 的 UAR 最大值。

表 2　本文方法使用不同损失函数时的性能比较

Tab. 2　The performance comparison for the proposed approach in using different loss functions

损失函数类型	$R_{\text{UAR-max}}$/%	$F1$/%
Focal	50.4	45.4
WCE	50.1	47.9
NWCE（k=1.5）	51.5	46.9
NWCE（k=2.0）	**53.2**	**49.6**
NWCE（k=2.5）	51.7	48.1
NWCE（k=3.0）	53.1	47.4

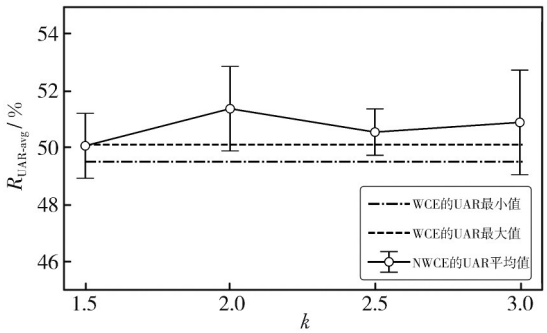


图 2　不同 k 值的非线性加权交叉熵损失的UAR平均值及其标准差

Fig. 2　The average UAR and its standard deviation for the nonlinear weighted cross-entropy loss in different k values

为了探讨本文BLSA模块中自注意力机制对口吃语音分类的影响，构建了仅包含Bi-LSTM层的模型，该模型与本文方法的不同是去除了自注意力层。模型使用多层wav2vec 2.0嵌入作为输入，并采用 k 取值为2.0的非线性加权交叉熵损失。进行5次重复实验，记录5次实验的开发集的UAR最大值（$R_{\text{UAR-max}}$）和平均值（$R_{\text{UAR-avg}}$），实验结果如表 3 所示。由表 3 可知，本文方法中的自注意力层可减少自监督预训练模型嵌入中的冗余信息。

表3 本文方法有无自注意力层时的性能比较

Tab. 3 The performance comparison of the proposed approach with and without a self-attention layer

本文模型类型	$R_{\text{UAR-max}}$/%	$R_{\text{UAR-avg}}$/%
含注意力层(NWCE,k=2.0)	**53.2**	**51.3**
不含注意力层(NWCE,k=2.0)	49.2	47.3

为了研究BLSA模块中Bi-LSTM层的输出维度对模型性能的影响，本文进行了多组超参数敏感性实验。将Bi-LSTM层的输出维度分别设置为{128,256,512}，并进行5次重复实验，记录5次实验的开发集的UAR最大值($R_{\text{UAR-max}}$)和平均值($R_{\text{UAR-avg}}$)，如表4所示。由表4可知，当Bi-LSTM层的输出维度取值为256，本文模型的性能最好。

表4 本文方法使用不同Bi-LSTM层的输出维度数时的性能比较

Tab. 4 The performance comparison of the proposed approach for different output dimensionality of Bi-LSTM layer

输出维度(NWCE,k=2.0)	$R_{\text{UAR-max}}$/%	$R_{\text{UAR-avg}}$/%
128	51.8	49.9
256	**53.2**	**51.3**
512	52.4	50.7

3.2.2 对比分析

将本文方法与现有口吃语音分类方法进行UAR性能对比，结果如表5所示。

表5 本文方法与其他方法在与KSF-C数据集上的性能对比

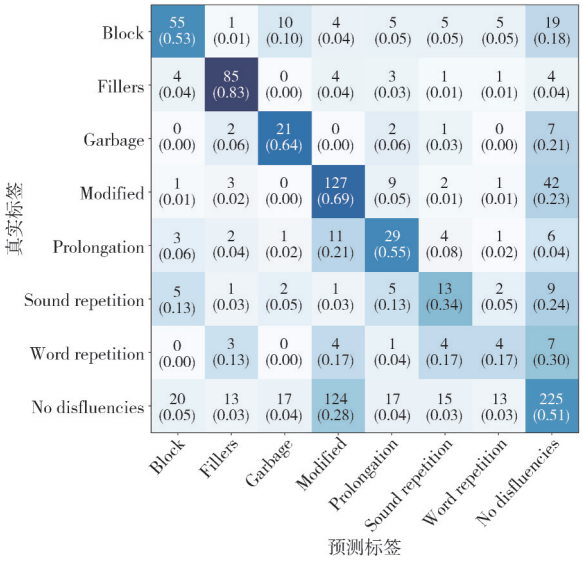
Tab. 5 The performance comparison between this paper approach and other approach on the KSF-C dataset

特征/表示	方法	$R_{\text{UAR-max}}$/%
ComParE	SVM[24]	30.2
wav2vec 2.0	MB StutterNet[11]	41.0
wav2vec 2.0	SVM[23]	49.0
wav2vec 2.0	MTL+Fine-Tune[22]	49.1
wav2vec 2.0	Fine-Tune[10]	50.1
wav2vec 2.0	BLSA-NWCE (本文方法)	**53.2**

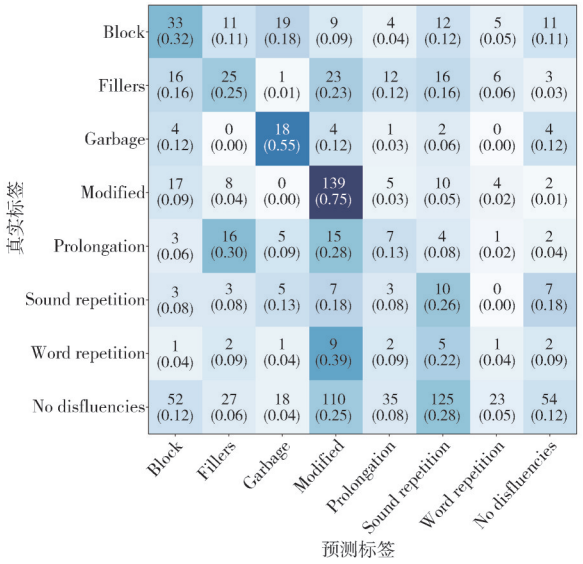
文献[24]给出了在KSF-C数据集上的基线实验结果；文献[11,22-23]则关注于wav2vec 2.0嵌入的统计特性，仅在时间维度上使用不同的统计泛函进行分析；文献[10]采用对wav2vec 2.0模型进行微调的方法，分别微调了小型、大型、超大型wav2vec 2.0模型。虽然超大型wav2vec 2.0模型取得了最佳结果，但仍与其他规模的模型结果处在同一量级且参数量大，因此本文选择大型模型结果作为对比对象。此外，本文还尝试将常用的传统音频特征MFCC作为输入特征来进行比较，即提取40维MFCC特征，通过二维卷积将其扩展为1 024维，输入BLSA-NWCE模型(k=2.0，单通道)，记录5次重复实验的开发集UAR的结果，最大值为25.8%。

所有对比实验均取KSF-C开发集上的UAR的最大值。由表5可知，所有基于wav2vec 2.0模型嵌入作为特征的方法均远高于基线结果。同时，本文方法性能优于其他3种只考虑wav2vec 2.0模型嵌入统计特性的方法。此外，与微调的wav2vec 2.0模型方法相比，本文方法具有较好的分类性能。

为进一步展示本文方法在识别不同口吃类别的性能，给出了实验结果中性能最佳模型生成的混淆矩阵和KSF-C数据集基线的混淆矩阵[24]，分别如图3(a)和图3(b)所示。



(a) 本文方法的开发集混淆矩阵



(b) 基线方法的开发集混淆矩阵

图3 本文方法和基线方法的开发集混淆矩阵对比

Fig. 3 Comparison of confusion matrices on the development set between this paper approach and the baseline approach

通过与基线混淆矩阵的比较可知，本文方法对大多数口吃类别识别的性能都优于基线方法。

4 总结与展望

本文提出了一种基于自监督预训练模型和非线性加权交叉熵损失的口吃语音分类方法。基于自监督预训练模型来提取口吃语音的副语言表示嵌入，使用带自注意力机制的双向长短期记忆网络对口吃语音进行分类，进一步使用非线性加权交叉熵损失关注样本较少的口吃类别。未来可以将语音信号与文本信息、图像信息等多模态信息相融合，以提升口吃语音分类的准确度和稳健性。

参考文献：

[1] SCHULLER B W，BATLINER A，AMIRIPARIAN S，et al. The ACM multimedia 2023 computational paralinguistics challenge：Emotion share & requests[C]//ACM International Conference on Multimedia，2023：9635-9639.

[2] XU X，DENG J，ZHANG Z，et al. Zero-shot speech emotion recognition using generative learning with reconstructed prototypes[C]//IEEE International Conference on Acoustics，Speech and Signal Processing（ICASSP）. IEEE，2023：1-5.

[3] HAIDER F，DE LA FUENTE S，LUZ S. An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech [J]. IEEE Journal of Selected Topics in Signal Processing，2019，14(2)：272-281.

[4] SHAHIN M，ZAFAR U，AHMED B. The automatic detection of speech disorders in children：Challenges，opportunities，and preliminary results[J]. IEEE Journal of Selected Topics in Signal Processing，2019，14(2)：400-412.

[5] SHEIKH S A，SAHIDULLAH M，HIRSCH F，et al. Machine learning for stuttering identification：Review，challenges and future directions[J]. Neurocomputing，2022，514：385-402.

[6] SHEIKH S A，SAHIDULLAH M，HIRSCH F，et al. Stutternet：Stuttering detection using time delay neural network[C]//European Signal Processing Conference（EUSIPCO）. IEEE，2021：426-430.

[7] KOURKOUNAKIS T，HAJAVI A，ETEMAD A. Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory [C]//IEEE International Conference on Acoustics，Speech and Signal Processing（ICASSP）. IEEE，2020：6089-6093.

[8] SEBASTIAN P B，DOMINIK W，ELMAR N，et al. Detecting dysfluencies in stuttering therapy using wav2vec 2.0[C]//Annual Conference of the International Speech Communication Association （INTERSPEECH），2022: 347.

[9] BAYERL S P，WAGNER D，NÖTH E，et al. The influence of dataset partitioning on dysfluency detection systems [C]//International Conference on Text，Speech，and Dialogue（ICTSD）. Cham：Springer International Publishing，2022：423-436.

[10] GRÓSZ T，PORJAZOVSKI D，GETMAN Y，et al. wav2vec2-based paralinguistic systems to recognise vocalised emotions and stuttering [C]//ACM International Conference on Multimedia，2022：7026-7029.

[11] SHEIKH S A，SAHIDULLAH M，OUNI S，et al. End-to-end and self-supervised learning for ComParE 2022 stuttering sub-challenge[C]//ACM International Conference on Multimedia，2022：7104-7108.

[12] JOUAITI M，DAUTENHAHN K. Dysfluency classification in stuttered speech using deep learning for real-time applications[C]//IEEE International Conference on Acoustics，Speech and Signal Processing（ICASSP）. IEEE，2022：6482-6486.

[13] KOURKOUNAKIS T，HAJAVI A，ETEMAD A. FluentNet：End-to-end detection of stuttered speech disfluencies with deep learning[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2021，29：2986-2999.

[14] BAEVSKI A，ZHOU Y，MOHAMED A，et al. wav2vec 2.0：A framework for self-supervised learning of speech representations[J]. Advances in Neural Information Processing Systems，2020，33：12449-12460.

[15] SUN H，LIAN Z，LIU B，et al. EmotionNAS：Two-stream architecture search for speech emotion recognition [DB/OL]. （2022-03-25）[2023-09-05]. https://arxiv.org/abs/2203.13617.

[16] VAESSEN N，VAN LEEUWEN D A. Fine-tuning wav2vec2 for speaker recognition[C]//IEEE International Conference on Acoustics，Speech and Signal Processing（ICASSP）. IEEE，2022：7967-7971.

[17] BRAUN F，ERZIGKEIT A，LEHFELD H，et al. Going beyond the cookie theft picture test：Detecting cognitive impairments using acoustic features [C]//International Conference on Text，Speech，and Dia-

logue (ICTSD). Cham: Springer International Publishing, 2022: 437-448.

[18] GHOSH S, TYAGI U, KUMAR S, et al. A novel multimodal dynamic fusion network for disfluency detection in spoken utterances [DB/OL]. (2022-11-27)[2023-09-05]. https://arxiv.org/abs/2211.14700.

[19] SHARMA M. Multi-lingual multi-task speech emotion recognition using wav2vec 2.0[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 6907-6911.

[20] REN Z, NGUYEN T T, CHANG Y, et al. Fast yet effective speech emotion recognition with self-distillation [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[21] BAYERL S P, WAGNER D, BAUMANN I, et al. Detecting vocal fatigue with neural embeddings[DB/OL]. (2022-04-07) [2023-09-05]. https://arxiv.org/abs/2204.03428.

[22] BAYERL S P, GERCZUK M, BATLINER A, et al. Classification of stuttering—The ComParE challenge and beyond[J]. Computer Speech & Language, 2023, 81: 101519.

[23] MONTACIÉ C, CARATY M J, LACKOVIC N. Audio features from the wav2vec 2.0 embeddings for the ACM multimedia 2022 stuttering challenge[C]//ACM International Conference on Multimedia, 2022: 7195-7199.

[24] SCHULLER B, BATLINER A, AMIRIPARIAN S, et al. The ACM multimedia 2022 computational paralinguistics challenge: Vocalisations, stuttering, activity, & mosquitoes[C]//ACM International Conference on Multimedia, 2022: 7120-7124.

[25] BAYERL S, VON GUDENBERG A W, HÖNIG F, et al. KSoF: The Kassel State of Fluency dataset—a therapy centered dataset of stuttering[C]//Language Resources and Evaluation Conference (LREC), 2022: 1780-1787.

[26] ARDILA R, BRANSON M, DAVIS K, et al. Common Voice: A massively-multilingual speech corpus[C]//Language Resources and Evaluation Conference (LREC), 2020: 4218-4222.