ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 2,Mar. 2022,pp. 288－297
DOI:10. 16337/j. 1004-9037. 2022. 02. 003

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于对比预测编码模型的多任务学习语种识别方法

赵建川[1,2]，杨浩铨[3]，徐　勇[3]，吴　恋[1,2]，崔忠伟[1,2]

(1. 贵州师范学院数学与大数据学院，贵阳 550018；2. 贵州师范学院大数据科学与智能工程研究院，贵阳 550018；3. 哈尔滨工业大学(深圳)计算机科学与技术学院，深圳 518000)

摘　要：语种识别的关键是从语音片段中提取有用的特征。通过延时神经网络(Time-delayed neural network，TDNN)可以提取包含丰富上下文信息的特征向量，有效提高系统性能。本文提出一种ECAPA(Emphasized channel attention)-TDNN＋对比预测编码(Contrastive predictive coding，CPC)模型的多任务学习语种识别网络。ECAPA-TDNN为主干网络，提取语音全局特征，改进的CPC模型为辅助网络，对ECAPA-TDNN提取的帧级特征进行对比预测学习，通过联合损失函数进行优化训练。在东方语种竞赛数据集AP17-OLR的10类语种上进行了实验。实验结果表明，本文提出的网络在1 s，3 s和全长(All)测试集测得的识别准确率相比于基础网络都有明显的提高。

关键词：语种识别；对比预测编码；多任务学习；ECAPA-TDNN；联合损失

中图分类号：TN912.34　　**文献标志码**：A

Language Identification Method for Multi-task Learning Based on Contrastive Predictive Coding Model

ZHAO Jianchuan[1,2]，YANG Haoquan[3]，XU Yong[3]，WU Lian[1,2]，CUI Zhongwei[1,2]

(1. School of Mathematics and Big Data，Guizhou Education University，Guiyang 550018，China；2. Big Data Science and Intelligent Engineering Research Institute，Guizhou Education University，Guiyang 550018，China；3. School of Computer Science and Technology，Harbin Institute of Technology(Shenzhen)，Shenzhen 518000，China)

Abstract：The key of language identification is to extract useful features from speech fragments. The time-delayed neural network (TDNN) can extract feature vectors，which contain rich context and improve system performance effectively. This paper proposes a multi-task learning method of ECAPA(Emphasized channel attention)-TDNN＋contrastive predictive coding (CPC) network for language identification. ECAPA-TDNN is the main network to extract the global features of language. The improved CPC model is the auxiliary network，and the frame level features extracted by ECAPA-TDNN are compared and predicted. Finally，the joint loss function is used to optimize the network. The proposed method is tested on the 10 language data sets provided by the AP17-OLR data set.　The result shows that the identification accuracy of the proposed network is higher than baseline on the 1 s，3 s and All test data sets of AP17-OLR.

基金项目：贵州省科技厅基础研究计划项目(黔科合基础-ZK〔2021〕一般334)；贵州省教育厅基础研究计划项目(黔科合基础〔2020〕1Y258)；贵州省省级重点学科“计算机科学与技术”项目(ZDXK〔2018〕007号)；贵州省教育厅创新群体研究项目(黔教合KY字〔2021〕022)；贵州省2018年第三批省级服务业发展引导资金项目(黔发改服务〔2018〕1181号)。
收稿日期：2022-01-17；**修订日期**：2022-02-19

Key words: language identification; contrastive predictive coding(CPC); multi-task learning; ECAPA-TDNN; joint loss

引 言

语种识别(Language identification,LID)[1]通过计算机自动判断某段音频属于哪一种语言,是智能语音处理领域的一个分支。语种识别技术在新一代信息技术中应用广泛,例如,多语种识别的语音处理技术、语音实时翻译和跨语言通信等[2]。语种识别的过程实际上是一个分类判决的过程,关键是获取分类判决有用的特征[3],其实现过程可分为3个步骤:从语音片段中获得声学特征、从声学特征中提取有用的特征和对提取的特征进行分类判决。

语种识别的声学特征是直接从音频中提取语谱特征参数,属于帧级特征。常用的声学特征包括移位分倒谱参数(Shifted delta cepstrum,SDC)[4]、感知线性预测系数(Perceptual linear predictive coefficient,PLP)[5]、梅尔倒谱参数(Mel frequency cepstral coefficient,MFCC)[6]和梅尔标度滤波器组(Filter bank,Fbank)[7]等。语种识别技术的实现主要基于底层声学特征,其发展经历了非深度学习和深度学习两个阶段。

非深度学习阶段主要又分为基于高斯混合模型(Gaussian mixed model,GMM)和基于身份向量(Identity vector,i-vector)特征的语种识别方法。文献[8]提出了高斯混合模型-通用背景模型(Gaussian mixed model-universal background model,GMM-UBM)的方法,该方法需要庞大的数据来估计协方差矩阵。数据量不足容易导致模型参数估计不准确,且跨信道使用时性能不佳。文献[9]提出了高斯混合模型-支持向量机(Gaussian mixed model-support vector machine,GMM-SVM)的均值超向量分类算法,该方法相对于GMM- UBM方法的识别性能有一定改善。i-vector特征是将每条音频的GMM超向量映射为含有音频显著特征的低维向量,这个低维向量即为i-vector。文献[10-11]使用从音频中提取的i-vector特征进行语种识别,有效地提高了识别效果,成为当时语种识别的主要方法之一。

基于深度学习的语种识别主要有i-vector语种识别方法和x-vector语种识别方法。文献[12]将增加了瓶颈层的神经网络(Bottleneck deep neural network,BN-DNN)作为i-vector的特征提取模型,对声学特征进行多层非线性映射和降维压缩,以得到鲁棒性更强的高层抽象特征。该方法有效改善了基于GMM模型的i-vector语种识别系统性能,对长时语音效果好,对短时语音则效果不佳。文献[13]提出了x-vector方法,通过延时神经网络(Time delay neural network,TDNN)将不定长的语音片段映射到固定维度的embedding,这个embedding就是x-vector。使用x-vector特征进行语种识别相比于i-vector特征具有更好的系统性能[14]。

研究者在x-vector特征提取TDNN网络的基础上进行了多种改进,以获得更有用的特征。文献[15]对TDNN网络进行改进提出了Extended-TDNN网络。Extended-TDNN网络拓展了时间上下文,并加入了Dense层,增加了网络深度。Extended-TDNN提取的x-vector相比于基础TDNN提取的x-vector性能有所提升。文献[16]提出了ECAPA(Emphasized channel attention)-TDNN网络,采用自注意力机制和多层聚合等增强方法,进一步拓展了时间上下文,并关注到全局属性,提取出的x-vector特征在语种识别中表现出更优异的识别性能。

ECAPA-TDNN网络是当前x-vector特征提取最先进的网络架构[17]。因此,本文在ECAPA-TDNN网络的基础上结合对比预测编码(Contrastive predictive coding,CPC)模型的思想,提出一种ECAPA-TDNN+CPC的多任务学习网络模型。以ECAPA-TDNN为主干网络,提取语音的全局特征;改进的CPC模型为辅助网络,对ECAPA-TDNN提取的帧级特征进行对比预测学习。最后,通过

联合损失函数进行优化训练。实验结果表明,本文提出的网络相比于基础网络ECAPA-TDNN具有更好的语种识别性能。

1 语种识别模型

1.1 标准TDNN的x-vector特征提取网络

语音信号是有时序性的数据,对于语音信号的时序相关性TDNN网络具有很好的描述能力,它能够获取语音的上下文信息,体现语音的动态特性。标准的TDNN网络由帧级别层、统计池化层和段级别层组成[18]。帧级别层为5层的时延网络结构,处理语音的帧级别特征。语音片段的声学特征序列$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$作为该层的输入,其中$n$表示声学特征的帧数。统计池化层对每一条语句的帧级别特征计算均值μ和标准差δ,表达式为

$$\mu=\frac{1}{N}\sum_{m=1}^{N}g_m \quad m=1,2,\cdots,N \tag{1}$$

$$\delta=\sqrt{\frac{1}{N}\sum_{m}^{N}(g_m-\mu)^2} \quad m=1,2,\cdots,N \tag{2}$$

式中:g_m表示帧级别特征;N表示语句的长度。

统计池化处理后得到整条语句的全局特征,但这个过程容易丢失部分语句的时序结构信息[19]。段级别层处理代表整个语音片段的全局性特征,由两层全连接层组成,靠近统计池化层的层称为Near层,远离统计池化层的层称为Far层,分别提取不同的x-vector特征,输入到全连接层后面Softmax层。

1.2 ECAPA-TDNN的x-vector特征提取网络

ECAPA-TDNN网络基于标准的TDNN网络结构设计,引入了多项增强功能以获取更强大的嵌入功能,网络结构如图1所示。首先,池化层依赖于通道和上下文注意力机制,使得网络可以关注每个通道的不同帧,赋予每一帧不同的权重,通过自注意力机制观察语句的全局属性,扩展池化层的时间上下文信息。其次,ECAPA-TDNN网络加入了SE-Res2Block模块。如图2所示,网络通过SE块与残差块Res2net[20]结合,重新调整帧级别层的通道数,在局部操作的卷积块中插入全局上下文信息,通过构建内部分层残差连接来处理多尺度特征,从而减少模型参数的数量。最后使用多层特征聚合将所有SE-Res2Block的输出特征映射相连,在池化之前合并补充信息,获取更细粒度语种特征以增强系统的鲁棒性。

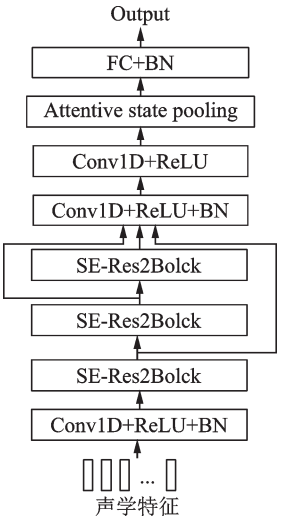

图1 ECAPA-TDNN网络结构图

Fig.1 Structure of ECAPA-TDNN network

SE-Res2Block模块在训练过程中为特征图分配权重,与目标关联大的分配较大权重,关联小的分配较小权重。

SE模块首先进行压缩操作,为每一个通道生成一个描述符,得到一个帧级特征的均值向量$\boldsymbol{z}$,表达式为

$$\boldsymbol{z}=\frac{1}{T}\sum_{t}^{T}\boldsymbol{h}_t \tag{3}$$

式中$\boldsymbol{h}_t$表示每个特征的embedding向量。

然后进行激励操作,使用$\boldsymbol{z}$中的描述符来计算每个通道的权重,即

$$s=\sigma(\boldsymbol{W}_2 f(\boldsymbol{W}_1\boldsymbol{z}+b_1)+b_2) \tag{4}$$

式中:$\sigma(\cdot)$为sigmoid函数;$f(\cdot)$为非线性函数;$\boldsymbol{W}_1\in\mathbf{R}^{R\times C}$,$\boldsymbol{W}_2\in\mathbf{R}^{C\times R}$,$C$为通道数,$R$为降维数;$b_1$,$b_2$表示偏移量。向量$\boldsymbol{s}$包含介于0和1之间的权值$\boldsymbol{s}_c$。这些权重通过乘法作用于原始输入,即

$$\tilde{\boldsymbol{h}}_c=\boldsymbol{s}_c\boldsymbol{h}_c \tag{5}$$

式中$\boldsymbol{h}_c$表示每个通道上的原始输入。

标准的TDNN网络在帧级层使用了较短的时间上下文信息,忽略了语音片段的全局信息。ECAPA-TDNN网络充分考虑了语音片段的全局属性,扩展了上下文信息,在信道估计过程中关注不同帧子集,性能更好,参数更少。

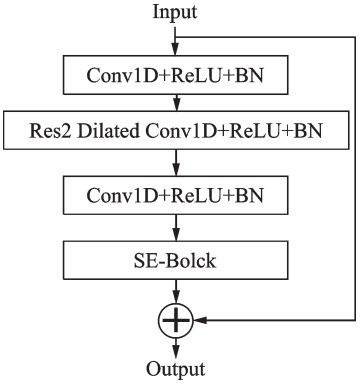

图2 SE-Res2Block模块

Fig.2 SE-Res2Block module

1.3 CPC模型方法

与预测编码模型相比,CPC模型[21]是一种无监督的特征提取模型,可以从高维数据学习到对预测最有用的表征,其依赖噪声对比估计训练模型,在图像、语音、自然语言处理和强化学习等多个领域都可以学习到高层信息。CPC模型结构如图3所示。

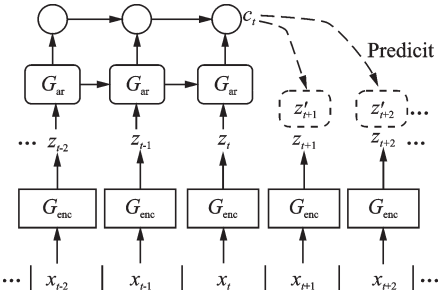

图3 CPC模型结构

Fig.3 Structure of CPC model

CPC模型以原始语音信号作为输入,采用一个非线性编码器将分割到时间窗口上的每个特征向量$\boldsymbol{x}_t$进行编码,得到一系列的表征向量$\boldsymbol{z}_t$,表达式为

$$\boldsymbol{z}_t=G_{enc}(\boldsymbol{x}_t) \tag{6}$$

然后再将$\boldsymbol{z}_t$以及潜空间中之前所有时刻的相关信息输入到一个自回归模型G_{ar}中,生成当前时刻的上下文表示为$\boldsymbol{c}_t$,即

$$\boldsymbol{c}_t=G_{ar}(\boldsymbol{z}_{\leqslant t}) \tag{7}$$

最后用$\boldsymbol{c}_t$去预测k个时刻后的$\boldsymbol{z}'_{t+k}$,通过最大化$\boldsymbol{x}_{t+k}$和$\boldsymbol{c}_t$之间的互信息,使得预测值$\boldsymbol{z}'_{t+k}$与真实值$\boldsymbol{z}_{t+k}$尽可能相似。

2 本文方法

多任务学习[22]是把多个相关的任务放在一起并行学习,通过多个梯度同时反向传播、多个任务参数共享来补充学习的一种机器学习方法,其参数共享方式分为硬参数共享和软参数共享两种模式。本文采用硬参数共享的多任务学习模型,把语种识别的训练任务分为主任务——语音特征提取和辅助任务——对比预测学习。主任务采用ECAPA-TDNN网络模型,首先提取语音片段的帧级特征,然后经过注意力池化层和全连接层进行语种的分类判决。辅助任务采用改进的CPC网络模型,以帧级特征作为输入进行对比预测学习。网络架构如图4所示,其中:J表示卷积核大小;d表示空洞卷积率,$d=1$表示正常卷积;C表示通道维度;T表示时间维度;S表示语种的类别数;GRU为门控循环单元;Z为经过Conv1D+ReLU层处理后得到的帧级特征;k为时间步长,一般取偶数。

2.1 主任务模块

主任务模块以ECAPA-TDNN作为主干网络,ECAPA-TDNN网络的帧级别层首先从声学特征中提取帧级特征向量$\boldsymbol{z}$。然后将网络进行分支:一个分支为辅助任务模块,另一个分支为主任务模块。两个分支均以帧级别特征向量$\boldsymbol{z}$作为输入,最终网络由这两个分支共同优化训练。

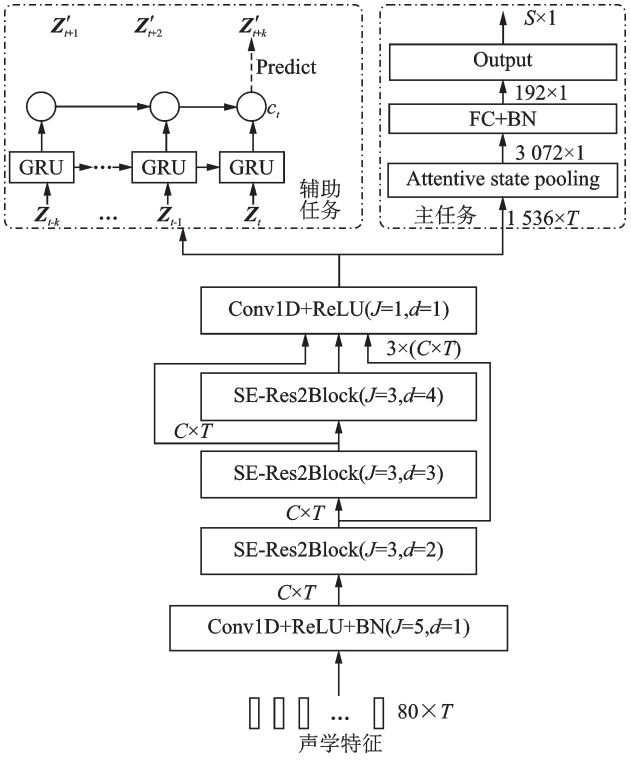

图4 本文方法网络整体架构图

Fig.4 Network architecture of the proposed method

2.2 辅助任务模块

多任务学习网络的辅助任务模块为改进的CPC模型。改进的CPC模型以ECAPA-TDNN网络的帧级网络取代CPC模型的非线性编码器,ECAPA-TDNN网络处理得到的帧级特征输入到改进的CPC模型自回归模块中,然后通过自回归模块进行对比预测学习构造正负样本对。

辅助任务模块中$Z=\{z_{t-k},\cdots,z_{t-2},z_{t-1},z_t\}$作为输入特征,自回归模型选用网络。GRU网络可以通过调节被提取特征的语音序列长度,得到丰富的上下文信息c_t,即

$$c_t=\mathrm{GRU}(Z) \tag{8}$$

用线性变换预测出z'_{t+k}。z'_{t+k}与$z_{l(t+k)}$构成正样本对,z'_{t+k}与$z_{j(t+k)}$构成负样本对。脚标l表示l类语种,j表示j类语种。CPC模型中,引入互信息的概念来优化网络,通过充分学习当前的上下文信息c_t最大程度减小未来z的不确定度,最大化正样本对之间的互信息,最小化负样本对之间的互信息,从而达到优化网络的效果。

2.3 联合损失函数

在语种识别任务中,语种识别特征训练模型的优化由多任务学习网络的损失函数共同完成。因此,为了提高正样本对的相似度和负样本的区分度,本文使用交叉熵损失函数L_{ce}和改进的噪声对比估计损失函数L_{infoNCE}对训练网络进行联合监督学习。交叉熵损失函数L_{ce}表达式为

$$L_{\mathrm{ce}}=-\sum_{i=1}^{B}\ln\frac{\exp(\boldsymbol{W}_{y_i}^{\mathrm{T}}\boldsymbol{x}_i+b_{y_i})}{\sum_{j=1}^{n}\exp(\boldsymbol{W}_j^{\mathrm{T}}\boldsymbol{x}_i+b_j)} \tag{9}$$

式中:B表示批次的大小;$\boldsymbol{x}_i$表示第y_i类中第i个样本的特征;$\boldsymbol{W}_j$为$\boldsymbol{W}$的第j行的参数;b为偏置量。

改进的噪声对比估计损失函数可以实现互信息最大化,损失值越小说明正样本对的相似度越高,表达式为

$$L_{\text{infoNCE}}=-E\left[\log\frac{f_k(\boldsymbol{x}_{t+k},\boldsymbol{c}_t)}{\sum\limits_{x_j\in X}f_k(\boldsymbol{x}_j,\boldsymbol{c}_t)}\right] \tag{10}$$

式中:$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$为一组样本;$(\boldsymbol{x}_{t+k},\boldsymbol{c}_t)$为正样本对;$(\boldsymbol{x}_j,\boldsymbol{c}_t)$为负样本对,正样本对取自与时间上下文$\boldsymbol{c}_t$间隔$k$个时间步长的样本,负样本为序列中随机选取的样本。$f_k(\boldsymbol{x}_{t+k},\boldsymbol{c}_t)$为密度比函数,表示信息上下文$\boldsymbol{c}_t$的预测值和未来真实值$\boldsymbol{x}_{t+k}$之间相似程度,正比于未来真实值与随机采样值的概率之比,即

$$f_k(\boldsymbol{x}_{t+k},\boldsymbol{c}_t)\propto\frac{p(\boldsymbol{x}_{t+k}|\boldsymbol{c}_t)}{p(\boldsymbol{x}_{t+k})} \tag{11}$$

用一个线性矩阵$\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_k$乘以$\boldsymbol{c}_t$,得到预测值$\boldsymbol{z}'^{\mathrm{T}}_{t+k}$,即

$$\boldsymbol{z}'^{\mathrm{T}}_{t+k}=\boldsymbol{W}_k\boldsymbol{c}_t \tag{12}$$

令$\boldsymbol{z}^{\mathrm{T}}_{t+k}$为真实值,则式(11)可简化为

$$f_k(\boldsymbol{x}_{t+k},\boldsymbol{c}_t)=\exp(\boldsymbol{z}^{\mathrm{T}}_{t+k}\boldsymbol{z}'^{\mathrm{T}}_{t+k}) \tag{13}$$

最后即可得到联合损失为

$$L_{\text{total}}=(1-\beta)L_{\text{ce}}+\beta L_{\text{infoNCE}} \tag{14}$$

联合损失L_{total}等于交叉熵损失L_{ce}和改进的噪声估计损失L_{infoNCE}之和,其中β为改进的噪声对比估计损失的权重系数,取值范围为0到1之间。

3 实验配置及效果分析

3.1 实验设置

实验使用东方语种识别竞赛提供的10种不同语言数据集AP17-OLR[23],10种语言分别为日语、韩语和哈萨克语(时长分别为5.8 h、5.9 h和5.4 h);粤语、普通话、印度尼西亚语(时长分别为7.7 h、7.6 h和7.5 h);越南语和俄语(时长分别为8.4 h和9.9 h),藏语和维吾尔语(时长均为10 h)。每个语种的语音采样频率为16 kHz。实验中随机抽取80%为训练集,20%为验证集。测试集包含1 s,3 s和全长(All)三个不同持续时间的子集。

本文实验中训练模型选用Adam优化器,epochs设置为50,batch_size设置为128。在多任务学习的辅助任务模型CPC模型的自回归选用GRU网络,损失函数权重系数β设置为0.001。实验选用准确率Acc作为评价指标[24-25]。

3.2 实验效果分析

3.2.1 多任务学习模型的性能分析

本节对多任务学习ECAPA-TDNN+CPC网络模型进行性能分析。将每类语种的MFCC声学特征输入到网络中,以3 s时长的语音作为测试集,改进的CPC网络模型中时间步长k取12,分别记录每一次迭代训练的损失、准确率和学习率,得到如图5、6所示周期性训练时系统参数变化曲线。由图5可知,在模型的训练过程中,学习率调整的机制为先增加后减小。由图6可知,第1次迭代训练的损失为1.926左右,准确率为92.75%,说明模型刚开始训练时,损失较大,准确率较低。随着迭代周期增加,损

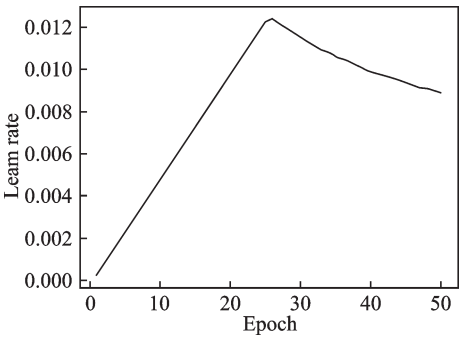

图5 周期性训练时学习率变化曲线图

Fig.5 Change curve of learning rate during periodic training

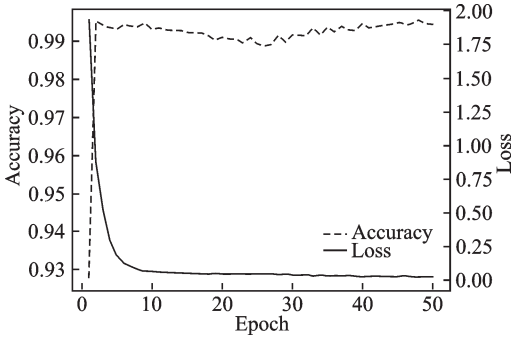

图6 周期性训练时准确率与损失变化曲线图

Fig.6 Change curve of accuracy and loss during periodic training

失开始下降,准确率逐渐增加。第10次迭代训练时,损失降为0.063左右,此时的准确率大约为99.31%,后面训练过程中损失逐渐减小,准确率会有小幅波动,说明模型收敛速度快。第40次迭代训练时模型已经基本趋于稳定。第48次迭代训练时准确率最高,达到99.54%,损失为0.020 4,此时得到的网络参数就是最终优化的网络参数指标。

图7为单图形处理器(Graphic processing unit, GPU)下ECAPA-TDNN网络和ECAPA-TDNN+CPC网络周期性训练时运行时间变化曲线图。由图7可知,ECAPA-TDNN网络每次训练时间在331~333 s之间,平均运行时间为331.12 s。ECAPA-TDNN+CPC网络每次训练的时间在336~339 s之间,平均运行时间为337.79 s,相对于基础网络相差了6.67 s。ECAPA-TDNN网络参数量为4.57 MB,ECAPA-TDNN+CPC网络参数量为7.47 MB,相对于基础网络增加了63.46%。虽然改进多任务学习网络的参数量增加了,但是与基础网络的系统运行时间并没有太大区别。

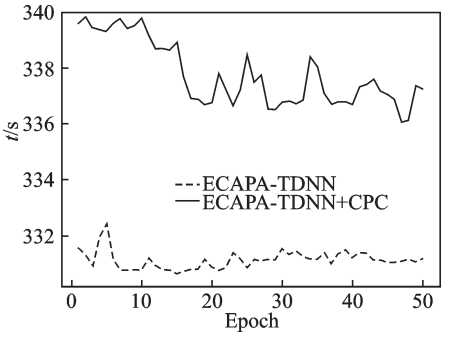

图7 周期性训练时运行时间变化曲线图

Fig.7 Change curve of running time during periodic training

3.2.2 多任务学习模型的实验效果分析

本节以MFCC和FBank声学特征作为输入,时间步长k取12,在1 s、3 s和All测试集上分别验证多任务学习ECAPA-TDNN+CPC网络和基础网络ECAPA-TDNN和CPC的语种识别准确率。实验分析结果如表1、2所示。

由表1可见,1 s、3 s和All三个测试集的实验中,多任务学习网络的识别准确率相比于ECAPA-TDNN网络分别提高了1.92%、3.69%和2.80%,相比于CPC网络分别提高了49.42%、36.15%和40.86%。

由表2可见,1 s、3 s和All三个测试集的实验中,多任务学习网络的识别准确率相比于ECAPA-TDNN网络分别提高了6.01%、4.11%和3.12%,相比于CPC网络分别提高了51.73%、25%和41.31%。

表1 多任务学习模型中输入为MFCC声学特征的准确率

Table 1 Accuracy of multi-task learning model when inputting MFCC acoustic characteristics

网络模型	Acc/%		
	1 s	3 s	All
CPC	13.19	42.10	38.63
ECAPA-TDNN	60.69	74.56	76.69
ECAPA-TDNN+CPC	62.61	78.25	79.49

实验结果表明,当输入为MFCC和FBank特征时,改进的多任务学习网络充分利用了ECAPA-TDNN网络和CPC模型的优势,获取了全局属性和丰富的上下文信息,同时通过CPC模型中的GRU自回归模块进行对比预测学习,进一步增强了特征提取的一致性,有效提高了语种识别准确率,能够更好地进行语种识别。

同时,由表1、2的实验数据对比可知,对于同一个网络FBank特征作为输入时比MFCC特征作为输入时的语种识别准确率更高,说明FBank声学特征提取了更有用的语种特征。

表2 多任务学习模型中输入为FBank声学特征的准确率

Table 2 Accuracy of multi-task learning model when inputting FBank acoustic characteristics

网络模型	Acc/%		
	1 s	3 s	All
CPC	15.48	56.64	43.37
ECAPA-TDNN	61.20	77.53	81.65
ECAPA-TDNN+CPC	67.21	81.64	84.68

3.2.3 不同时间步长的实验效果分析

在ECAPA-TDNN+CPC网络中,时间步长k取不同值对网络的识别准确率也有一定的影响,本节分别对k取8、12、16、20进行实验,分析不同时间步长时网络的性能。网络的输入选用MFCC声学特征。时间步长k取不同值时的实验结果如表3所示。由表3的实验数据可见,当测试的音频时长为1 s、k取16时,测得的识别准确率最高,相对于k取8、12、20分别增加了4.06%、2.38%和0.62%。测试的音频时长为3 s、k取12时,测得的识别准确率最高,相对于k取8、16、20分别增加了1.92%、0.12%和0.90%。测试的音频为All、k取20时,测得的识别准确率最高,相对于k取8、12、16分别增加了0.73%、0.66%和1.02%。

表3 k取不同值时的准确率

Table 3 Accuracy when k taking different values

k	Acc/%		
	1 s	3 s	All
8	62.34	76.33	79.42
12	64.02	**78.25**	79.49
16	**66.40**	78.13	79.31
20	65.78	77.35	**80.15**

3.2.4 不同网络上的实验效果分析

本节以MFCC和FBank声学特征作为网络输入,时间步长k取12,在不同网络上进行实验效果对比,结果如表4、5所示。由表4、5的实验数据可见,ECAPA-TDNN+CPC网络的实验效果相对于TNDD+CPC网络和EX-TDNN+CPC网络的语种识别正确率均有提高。当输入特征为MFCC声学特征时,在1 s、3 s和All数据集的实验效果相比于TDNN+CPC网络准确率分别提高了10.57%、18.52%和16.58%,相比于EX-TDNN+CPC网络准确率分别提高了4.9%、11.53%和9.44%。

当输入特征为FBank声学特征时,在1 s、3 s和All数据集的实验效果相比于TDNN+CPC网络准确率分别提高了16.98%、22.99%和24.84%,相比于EX-TDNN+CPC网络准确率分别提高了10.78%、20.68%和20.92%。

4 结束语

本文提出一种融合CPC模型的多任务学习语种识别网络,ECAPA-TDNN+CPC模型。该模型在主干网络ECAPA-TDNN中加入一个自回归模块,对ECAPA-TDNN网络提取的帧级特征进行对比预测学习,构造正负样本对,通过最大化正样本对之间的相似度和最小化负样本对之间的相似度来优化网络,增强所提特征的一致性。最后在东方语种竞赛数据集AP17-OLR上进行验证。实验结果表明,

表4 不同网络中输入为MFCC声学特征时的准确率

Table 4 Accuracy of different networks when inputting MFCC acoustic characteristics

网络模型	Acc/%		
	1 s	3 s	All
TDNN+CPC	52.04	59.73	62.91
EX-TDNN+CPC	57.71	66.72	70.05
ECAPA-TDNN+CPC	**62.61**	**78.25**	**79.49**

表5 不同网络中输入为FBank声学特征时的准确率

Table 5 Accuracy of different networks when inputting FBank acoustic characteristics

网络模型	Acc/%		
	1 s	3 s	All
TDNN+CPC	50.23	58.65	59.84
EX-TDNN+CPC	56.43	60.96	63.76
ECAPA-TDNN+CPC	**67.21**	**81.64**	**84.68**

提出的ECAPA-TDNN+CPC网络可以快速收敛,识别准确率明显提高,能够更好地对语种进行分类。

参考文献:

[1] CAI Weicheng, CHEN Jinkun, LI Ming. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system[EB/OL]. (2018-04-14)[2022-02-10]. https://arxiv.org/abs/1804.05160.

[2] 柏财通,崔翛龙,李爱.基于蒸馏后联邦学习的鲁棒性语音识别技术[J/OL].计算机工程.[2022-01-20]. DOI:10.19678/j.issn.1000-3428.0062812.
BAI Caitong, CUI Xiaolong, LI Ai. Robust speech recognition technology based on federal learning after distillation[J/OL]. Computer Engineering. [2022-01-20]. DOI: 10.19678/j.issn.1000-3428.0062812.

[3] WADHAWAN A, KUMAr P. Deep learning-based sign language recognition system for static signs[J]. Neural Computing and Applications, 2020, 32(2): 1-12.

[4] ALLEN F, AMBIKAIRAJAH E, EPPS J. Warped magnitude and phase-based features for language identification[C]// Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. Toulouse, France: IEEE, 2006: 201-204.

[5] WONG K Y E. Automatic spoken language identification utilizing acoustic and phonetic speech information[D]. Australia: Queensland University of Technology, 2004.

[6] MUDA L, BEGAM M, ELAMVAZUTHI I. Voice recognition algorithms using MEL frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques[J]. OALib Journal, 2010, 2(3): 138-143.

[7] FARHANG-BOROUJENY B. Filter bank spectrum sensing for cognitive radios[J]. IEEE Transactions on Signal Processing, 2008, 56(5): 1801-1811.

[8] TORRES-CARRASQUILLO P A, SINGER E, KOHLER M A, et al. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features[C]//Proceedings of the 7th International Conference on Spoken Language. Denver, Colorado, USA: IEEE, 2002.

[9] CAMPBELL W M, SINGER E, TORRES-CARRASQUILLO P A, et al. Language recognition with support vector machines[C]//Proceedings of the Speaker and Language Recognition Workshop. [S.l.]: IEEE, 2004.

[10] DEHAK N, TORRES-CARRASQUILLO P A, REYNOLDS D, et al. Language recognition via i-vectors and dimensionality reduction[C]//Proceedings of Conference of the International Speech Communication Association. Florence, Italy: ISCA, 2011: 857-860.

[11] RAMOJI S, GANAPATHY S. Supervised i-vector modeling for language and accent recognition[J]. Computer Speech and Language, 2020, 60: 101030.1-101030.19.

[12] SARKAR A K, TAN Z H. Time-contrastive learning based DNN bottleneck features for text-dependent speaker verification [EB/OL]. (2019-05-11)[2022-02-10]. https://arxiv.org/abs/1704.02373v2.

[13] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-Vectors: Robust DNN embeddings for speaker recognition[C]// Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE,

2018: 5329-5333.

[14] CAI W, CAI Z, ZHANG X, et al. A novel learnable dictionary encoding layer for end-to-end language identification[C]// Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE, 2018: 5189-5193.

[15] SNYDER D, GARCIA-ROMERO D, SELL G, et al. Speaker recognition for multi-speaker conversations using X-vectors [C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). [S.l.]: IEEE, 2019: 5796-5800.

[16] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification[EB/OL]. (2020-08-10)[2022-02-10]. https://arxiv.org/abs/2005.07143.

[17] TONG F, ZHAO M, ZHOU J, et al. ASV-SUBTOOLS: Open source toolkit for automatic speaker verification[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). [S.l.]: IEEE, 2021: 6184-6188.

[18] RICHARDSON F, REYNOLDS D, DEHAK N. Deep neural network approaches to speaker and language recognition[J]. IEEE Signal Processing Letters, 2015, 22(10): 1671-1675.

[19] LIU D, XU J, ZHANG P, et al. A unified system for multilingual speech recognition and language identification[J]. Speech Communication, 2020, 127(10): 17-28.

[20] VILLALBA J, CHEN N, SNYDER D, et al. State-of-the-art speaker recognition for telephone and video speech: The JHU-MIT submission for NIST SRE18[C]//Proceedings of Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019: 1488-1492.

[21] OORD A, LI Y, VINYALS O. Representation learning with contrastive predictive coding[EB/OL]. (2018-09-25)[2022-02-10]. https://arxiv.org/abs/1807.03748.

[22] 李亚,张雨楠,彭程,等.基于多任务学习的人脸属性识别方法[J].计算机工程, 2020, 46(3): 229-236.
LI Ya, ZHANG Yu'nan, PENG Cheng, et al. Face attributes recognition method based on multi-task learning[J]. Computer Engineering, 2020, 46(3): 229-236.

[23] LI Jing, WANG Binling, ZHI Yiming, et al. Oriental language recognition (OLR) 2020: Summary and analysis[EB/OL]. (2021-07-05)[2022-02-10]. https://arxiv.org/abs/2107.05365.

[24] RAMESH G, KUMAR C S, MURTY K S R. Self-supervised phonotactic representations for language identification[C]// Proceedings of Conference of the International Speech Communication Association. Brno, Czech Republic: ISCA, 2021: 1514-1518.

[25] 秦晨光, 王海, 任杰, 等. 基于多任务学习的方言语种识别[J].计算机研究与发展, 2019, 56(12): 2632-2640.
QIN Chenguang, WANG Hai, REN Jie, et al. Dialect language recognition based on multi-task learning[J]. Journal of Computer Research and Development, 2019, 56(12): 2632-2640.

作者简介:

赵建川(1988-),女,讲师,研究方向:深度学习、语种识别,E-mail:zhaojianchuan@gznc.edu.cn。

杨浩铨(1998-),男,硕士研究生,研究方向:智能语音处理、语音转换,E-mail:20S051066@stu.hit.edu.cn。

徐勇(1972-),**通信作者**,男,博士,教授,研究方向:模式识别、特征提取,E-mail:later fall@hit.edu.cn。

吴恋(1988-),女,博士,副教授,研究方向:机器学习、计算机视觉,E-mail:373201377@qq.com。

崔忠伟(1980-),男,博士,教授,研究方向:物联网技术,E-mail:33374225@qq.com。

(编辑:张黄群)