基于dVAE-BERT模型的阿尔茨海默症检测方法

陈旭初,蒲 钰,张卫强*
(清华大学电子工程系,北京 100084)

摘 要: 阿尔茨海默症(Alzheimer's Disease,AD)是一种神经退行性疾病,患者会出现失语症、语言流畅性降低等症状.目前已经有研究者使用发音特征,流畅性、停顿等副语言学特征,或者从转录的文本中提取特征检测阿尔茨海默症.但是,传统声学特征检测方法难以获取语义信息,而将语音转录成文本又费时费力,并且由于老年人口音、患病等影响,转录质量下降明显.本文使用离散变分自编码器(discrete Variational Autoencoders,dVAE)将语音转换为伪音素序列后,利用BERT(Bidirectional Encoder Representations from Transformers)模型对伪音素序列的连接关系进行建模,提出一种dVAE-BERT模型,从而提取音频在语言维度的表征.该模型在ADReSSo(Alzheimer's Dementia Recognition through Spontaneous Speech only)数据集上,模型的准确率为70.42%,比基线系统提高5.63%,其与Wav2vec2.0、HuBERT(Hidden-unit BERT)模型融合后,准确率分别为76.06%、71.83%.

关键词: 阿尔茨海默症;语音检测;dVAE;BERT
基金项目: 国家自然科学基金(No.62276153)
中图分类号: TP391.5 **文献标识码:** A **文章编号:** 0372-2112(2024)09-2971-08
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230050

Detection of Alzheimer's Disease Based on dVAE-BERT Model

CHEN Xu-chu, PU Yu, ZHANG Wei-qiang*
(*Department of Electronic Engineering*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: Alzheimer's disease (AD) is a neurodegenerative disease that causes symptoms such as aphasia and decreased speech fluency. Researchers have used articulatory features, paralinguistic features such as fluency and pauses, or features extracted from transcribed text to detect Alzheimer's disease. However, traditional acoustic feature detection methods are difficult to obtain semantic information, while transcribing speech into text is time-consuming and laborious, and the quality of transcription is significantly degraded due to the effects of accent and disease in the elderly. In this paper, we propose a dVAE-BERT (discrete Variational Autoencoders-Bidirectional Encoder Representations from Transformers) model, which uses discrete Variational Autoencoders (dVAE) to convert speech into pseudo-phoneme sequences, and then uses the Bidirectional Encoder Representations from Transformers (BERT) model to model the connection relations of the pseudo-phoneme sequences to extract the representation of audio in the language dimension. The accuracy of the model on the ADReSSo (Alzheimer's Dementia Recognition through Spontaneous Speech only) dataset is 70.42%, which is 5.63% better than the baseline system, and its accuracy is 76.06% and 71.83% after fusion with Wav2vec2.0 and Hidden-unit BERT (HuBERT) models, respectively.

Key words: Alzheimer's disease; speech detection; dVAE; BERT
Foundation Item(s): National Natural Science Foundation of China (No.62276153)

1 引言

阿尔茨海默症(Alzheimer's Disease,AD)是一种神经退行性疾病,随着症状加重,患者会出现语言障碍、记忆丧失等症状[1,2].随着老龄化加速,被诊断为阿尔茨海默症的人数正在迅速增加,但是传统的阿尔茨海默症检测方法耗费时间长,成本高昂[3].研究表明,阿尔茨海默症患者由于脑部发生病变,会出现失语症,在讲话时难以找到合适的词汇,言语中实词减少、语言空洞[4,5],且语言流畅性变差,话语中会出现多次犹豫和

收稿日期:2023-01-13;修回日期:2023-07-19;责任编辑:梅志强
*通讯作者:张卫强

停顿[6]，这些症状会随着病症的加重愈发明显．由于使用语音检测阿尔茨海默症具有成本低、无创、便于大规模检测等优势，因此，利用语音数据筛查阿尔茨海默症已经成为研究热点．

文献[7~10]中的研究表明，可以使用从自发语言中提取的发音特征，流畅性、停顿等副语言学特征检测阿尔茨海默症．比如，文献[8]使用ComParE(Computational Paralinguistics Challenge)、eGeMAPS(extended Geneva Minimalistic Acoustic Parameter Set)[11]等声学特征集和不同的分类器检测阿尔茨海默症，文献[12]使用OpenSMILE工具提取IS09(INTERSPEECH 2009 Emotion Challenge Features)、IS10(INTERSPEECH 2010 Paralinguistic Challenge Features)等声学特征集，而后使用门控卷积神经网络提取高维特征，通过多数投票的方法确定最终检测结果．这些使用发音和韵律等声学特征的传统检测方法，虽然取得有一定效果，但是难以挖掘语音的深层表征，并且使用全局统计特征，会丢失具体的韵律学细节，难以分析患者语音特征的变化规律，在建模过程中也会忽略患者语义信息的变化．随着机器学习技术的发展，部分研究者尝试使用神经网络提取高维表征．文献[13]中，使用卷积循环神经网络从语音的对数滤波器组特征中提取出特征序列，然后使用循环神经网络进行阿尔茨海默症检测．文献[14]在提取声学特征后，使用长短时记忆网络进行检测．但是传统的神经网络难以获取长距离、语言维度的特征．

在语音识别领域，预训练模型的使用极大的提高了自动语音识别的准确率，因此，有研究者使用预训练模型作为特征提取器，用于检测阿尔茨海默症．比如文献[15]使用预训练模型Wav2vec2.0[16]提取高维表征后，而后使用逻辑回归、支持向量机等进行分类．但是预训练语音模型主要使用正常人的语音进行训练，难以针对性的提取患者语音的表征．

部分研究者使用语音转录后的文本进行检测，取得相比于语音更好的效果，例如，文献[17]使用Word2vec[18]和GloVe(Global Vectors for word representation)[19]模型提取文本特征，而后使用卷积神经网络和长短时记忆网络进行分类．文献[20]对患者和健康人的词性特征、词汇丰富度和句法复杂性特征进行了比较．文献[21]使用fastText[22]和卷积神经网络，提取文本中基于n-gram的语言特征．但使用文本检测阿尔茨海默症时，需要先将语音转录为文本，在此过程中，声音的波动、情绪的变化都会被删除，而且如果需要标注出语音中的停顿、犹豫等标志，会更加耗时耗力．文献[23]使用自动语音识别模型将患者的语音转录为文本，并研究了语言模型对于识别结果的影响，但是由于口音、方言、年龄、健康状况等差异，阿尔茨海默症数据集的文本转录质量并不高．

为从语音中获取语义信息，同时保留语音中的停顿、流畅性等信息，并避免阿尔茨海默症患者转录文本质量低的问题，受到文献[24]启发，我们提出dVAE-BERT模型，该模型主要由离散变分自编码器(discrete Variational Autoencoders，dVAE)[25]和BERT(Bidirectional Encoder Representations from Transformers)模型[26]组成，dVAE模型的编码器由多个一维跨步卷积层和残差块组成，能够实现语音梅尔谱到伪音素序列的转换，而后使用BERT模型对伪音素序列的连接关系进行建模，最终使用dVAE-BERT模型提取的音频表征进行阿尔茨海默症检测，这种方法在INTERSPEECH2021 ADReSSo数据集[27]上取得了较好的效果，在与Wav2vec2.0、HuBERT(Hidden-unit BERT)[28]等自监督语音预训练模型融合后，检测性能进一步提高．

2 系统结构

dVAE-BERT模型的目标是将连续的梅尔谱转换为离散的伪音素序列，并对序列的连接关系进行建模，得到语义维度的特征，而后使用该特征进行阿尔茨海默症检测．其主要由dVAE模型和BERT模型组成，图1为系统结构示意图．

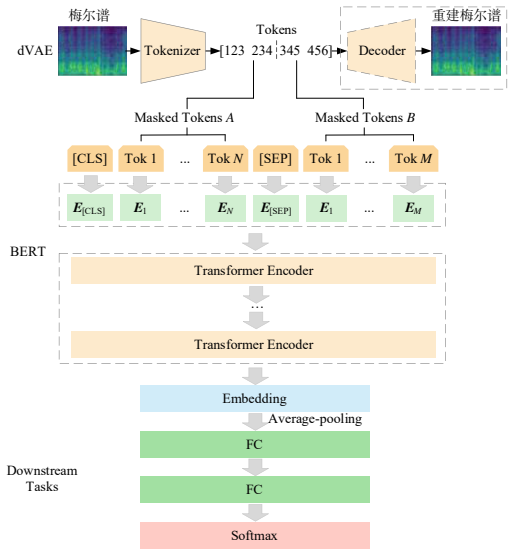


图1 dVAE-BERT系统结构图

2.1 dVAE模型

传统的变分自编码器(Variational Autoencoders，VAE)主要由编码器网络q_φ和解码器网络p_θ组成[29]．变分自编码器使用编码器将输入序列$\boldsymbol{x}=[x_0,x_1,\cdots,x_n]$编码为隐变量$\boldsymbol{z}$，并假定隐变量的先验分布为多变量高斯分布$p(\boldsymbol{z})$，而后通过解码器$p_\theta(\boldsymbol{x}|\boldsymbol{z})$生成序列$\boldsymbol{x}$．解码器的联合分布被定义为$p(\boldsymbol{z})\,p_\theta(\boldsymbol{x}|\boldsymbol{z})$，对于长度为$T$的序

列$\boldsymbol{x}$,$p_\theta(\boldsymbol{x}|\boldsymbol{z})$为:

$$p_\theta(\boldsymbol{x}|\boldsymbol{z})=\prod_{i=1}^{T} p_\theta(x_i|x_{<i},\boldsymbol{z}) \tag{1}$$

由于隐变量$\boldsymbol{z}$是高维变量,其后验分布$p_\theta(\boldsymbol{z}|\boldsymbol{x})$难以求解,因此使用一个较为简单的分布$q_\theta$来近似$p_\theta(\boldsymbol{z}|\boldsymbol{x})$.为了使得$q_\theta$和$p_\theta$这两个分布尽可能的相似,最小化两个分布之间的KL散度(Kullback-Leibler divergence),其证据下界为:

$$L=E_{q_\varphi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]-D_{\mathrm{KL}}\left[q_\varphi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})\right] \tag{2}$$

其中,右侧第一项表示重建误差,第二项表示$\boldsymbol{z}$的后验分布$q_\varphi(\boldsymbol{z}|\boldsymbol{x})$和先验分布$p(\boldsymbol{z})$的正则项.

而dVAE模型通过编码器学习中间编码d,然后通过最近邻搜索将中间编码映射为词表中k个向量之一,最终通过解码器对中间编码进行重建.由于中间编码为离散变量,因此$q_\varphi(\boldsymbol{d}|\boldsymbol{x})$不可导,所以使用Gumbel-Softmax技巧使得解码器$q_\varphi(\boldsymbol{d}|\boldsymbol{x})$可微[25],最终式(2)转化为:

$$L\geqslant E_{q_\varphi(\boldsymbol{d}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{d})\right]-\beta\cdot D_{\mathrm{KL}}\left[q_\varphi(\boldsymbol{d}|\boldsymbol{x})\|p(\boldsymbol{d})\right] \tag{3}$$

其中,$q_\varphi(\boldsymbol{d}|\boldsymbol{x})$表示dVAE的编码器生成的中间编码$d$的分布,$p_\theta(\boldsymbol{x}|\boldsymbol{d})$表示其解码器生成的序列的分布.$\beta$表示正则项的权重,仅当$\beta=1$时,式(3)中的等号成立.由此,可以使用dVAE模型将连续的梅尔谱转换为离散的编码序列.dVAE模型示意图见图2.

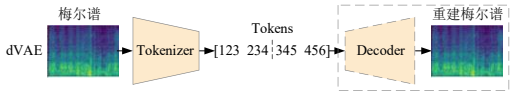


图2　dVAE模型示意图

2.2　BERT模型

BERT模型主要由多层双向Transformer组件(Transformer Encoder)组成,能够生成双向语言表征,在许多下游任务中表现出了出色的性能[26],其系统结构如图3所示.在训练BERT时使用了两个训练任务:屏蔽语言模型(Masked Language Model,MLM)和下句预测(Next sentence prediction,NSP).在MLM任务中,随机遮挡一个训练序列中15%的数据用于预测,假如是第i个标记被选中,则有以下三种处理方式:一是有80%的概率将其替换为[MASK],二是有10%的概率将其随机替换为其他标记,三是有10%的概率保持不变.而后再用上下文的标记预测出被掩蔽的标记.采用MLM对双向的Transformers进行训练,可以生成深层的双向语言表征,学习标记的前后依赖关系.为了直接获取整个序列层次的表征,使用NSP任务进行训练.在数据集中挑选出序列A和序列B,拼接后生成新的序列,其中50%的长序列中,序列B是序列A的下半部分,剩下50%的长序列中,序列B是数据集中的随机的序列,且与序列A不连续.在两个选定的序列之间插入了一个[SEP]标记,而后进行二元分类,判断第一个序列和第二个序列是否连续.

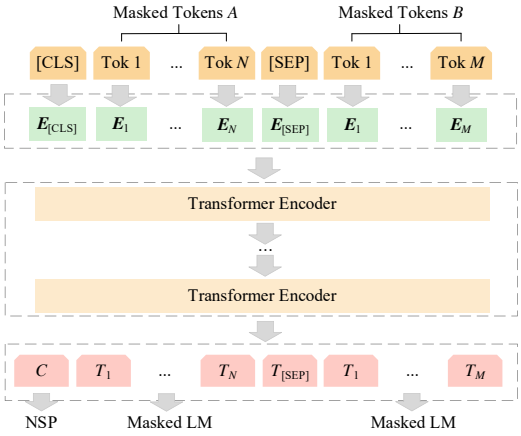


图3　BERT结构示意图

2.3　dVAE-BERT模型

dVAE-BERT模型的训练分为三个阶段.第一阶段,训练dVAE模型来压缩梅尔谱,将其转换为一维的伪音素序列,序列中的元素为离散变分自编码器词表的索引.dVAE模型的编码器为多层一维跨步卷积层,并且在每一个跨步卷积层后加入一个残差块,从而增强模型的学习能力.通过编码器的转换,实现梅尔谱图到dVAE模型词表的映射,并且生成的序列为独热编码的形式.第二阶段,使用dVAE模型的编码器生成的伪音素序列训练BERT模型.将dVAE模型的词表作为主体,并加入[PAD]、[CLS]、[SEP]、[MASK]4个标志,作为BERT模型的词典,并使用该词典生成词嵌入向量.对于MLM任务,按照与原BERT模型相同的方式进行掩蔽.对于NSP任务,将输入BERT模型的一个批次的序列,按照序列长度的一半进行分割,然后从中随机挑选序列进行拼接,按照1:1的比例生成前后连续、前后不连续的序列作为训练数据,并在每个序列前都添加[CLS]标志,在前后句之间加入[SEP]标志,而后进行下句预测.第三阶段,在下游任务中,冻结dVAE模型和BERT模型的参数,使用dVAE-BERT模型提取语音的高维表征,而后进行根据时间维度进行平均池化(Average-pooling),使用平均池化后的表征训练由两个全连接层(FC layer)组成的二元分类网络,进行阿尔茨海默症检测.

2.4　自监督语音预训练模型

一些研究者也尝试使用自监督语音预训练模型进行阿尔茨海默症检测[30].目前语音自监督预训练模型快速发展,通过使用多达上万小时的语音数据进行训练,使得模型在语音识别、情感计算等下游任务上取得较好的性能.为提高检测效果,我们尝试将dVAE-BERT模型提取的语义维度的表征和自监督语音预训练模型提取的表征

相融合. Wav2vec2.0模型[16]是自监督语音表示学习(Self-supervised Speech representation Learning, SSL)的代表. 其主要由特征编码器、上下文编码器和量化模块三部分组成. 在向前传递中,将原始语音波形输入特征编码器,得到潜在表征序列,然后利用量化模块,将潜在表征离散为有限的量化表征集,然后将它们送到上下文编码器中,以生成最终的上下文表示,从而捕获整个序列上的长距离依赖关系. HuBERT模型[28]使用了k-means方法对梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient, MFCC)特征进行聚类,将特征转换为离散化的目标序列,而后使用屏蔽语言模型的自监督预训练方法预测掩码位置的目标,从而提取语音的高维表征.

2.5 早融合、晚融合

目前已经有许多研究者使用融合的方法进行阿尔茨海默症检测,通常情况下,可以通过融合的方法将不同模型、不同模态的互补信息结合起来,使得系统有更好的表现,融合后的效果要好于单个模型的结果[31~33],而常用的融合方法主要有早融合、晚融合等[34]. 早融合,主要是对不同模态,或者不同模型提取的特征执行拼接、逐元素求和等操作,从而对不同的特征进行灵活的组合. 晚融合,主要采用平均、加权投票或多数投票的方法,将来自单模态,或者是单模型的分类预测结果集成在一起,以做出最终的预测.

3 实验设置与结果分析

3.1 数据集

我们使用INTERSPEECH2021 ADReSSo数据集[27]对模型进行测试. ADReSSo数据集主要为健康的受试者和诊断为阿尔茨海默症的患者描述图片的语音录音,共由237个音频文件组成,训练集中的166个语音片段包括AD患者的87个语音片段和来自健康对照组的79个语音片段,测试集中的71个片段中,AD患者和健康人的语音片段数量分别为35、36个. 数据集中每条音频时长约为22~268 s,总时长约为5.05 h,见表1. 采样率为44.1 kHz,并使用去除固定噪声、音频音量归一化等方式降低录音条件的产生的影响. 所有参与者的母语都是英语,年龄在53岁到84岁之间,语音内容为描述盗窃饼干的画面.

表1 INTERSPEECH 2021 ADReSSo数据集统计信息

类别		音频数量	时长/s	总时长/h
训练集	AD	87	35~268	2.12
	健康人	79	22~168	1.50
测试集		71	22~150	1.43

3.2 实验设置

我们将数据集音频的采样率降为16 kHz,而后使用长度为10 s、窗移为6 s的时间窗对训练集和测试集进行分割. 对于训练集中分割后片段数量较少的音频文件,采用加噪的方式进行数据增强,用以得到均衡的训练数据. 而后按照9∶1的比例将训练数据分割为训练集、验证集. 在提取梅尔谱时,使用帧长25 ms、帧移10 ms的汉宁窗对音频进行分帧,再用torchlibrosa工具提取128维的梅尔谱. 而后将提取的梅尔谱送入dVAE模型进行训练. dVAE模型的编码器为2层一维跨步卷积层,卷积核大小为3、步长为2,且在每个一维跨步卷积层后面接一个残差块. 训练时使用Adam优化器更新模型的权重,学习率设置为0.000 1,批大小为16,共迭代20轮次. 经过编码器中的多层卷积,将输入的梅尔谱转换为长度为250的伪音素序列. BERT模型主要由6层Transformer Encoder组成,每层有12个注意力机制,特征维度设置为768,也使用Adam优化器进行训练,学习率为0.000 03,迭代30轮次. 在使用dVAE模型训练BERT模型时,冻结dVAE模型的参数.

在进行阿尔茨海默症检测时,使用dVAE-BERT模型提取各片段的高维表征,而后将表征按照时间维度进行平均池化,得到768维的向量,而后使用该向量训练由两个具有ReLU激活层的全连接层组成的二元分类网络,全连接层的神经元个数分别设置为256、2. 训练分类网络时,采用使用动量的随机梯度下降方法,学习率设置为0.000 1,动量为0.9,迭代20轮次. 在测试时,对分割后的测试集片段分别进行测试,将模型输出的所有片段的预测结果进行归一化,而后将一条音频文件各片段的结果对应相加,作为该条音频的得分. 基线设置为比赛提供的基线.

为进一步提升测试结果,我们使用dVAE-BERT模型和Wav2vec2.0(https://huggingface.co/facebook/wav2vec2-base)、HuBERT (https://huggingface.co/facebook/hubert-base-ls960)模型进行了晚融合、早融合实验. Wav2vec2.0、HuBERT模型均使用960 h的Librispeech数据集进行训练. 晚融合时,使用dVAE-BERT模型和自监督预训练模型测试每条音频,然后将两个模型的预测结果相加. 早融合时,首先使用两个模型分别提取表征,并进行平均池化,两个表征拼接后,再输入两个全连接层组成的分类网络进行分类. 同时,作为对照实验,将Wav2vec2.0、HuBERT模型分别提取表征,并使用同样的二元分类网络进行分类.

在基于传统声学特征的模型中,我们提取语音的对数梅尔谱特征后,输入MobileNetV2[35]、ResNet[36]模型,使用学习率为0.000 1的Adam优化器进行训练,分别训练40轮次.

3.3 结果分析

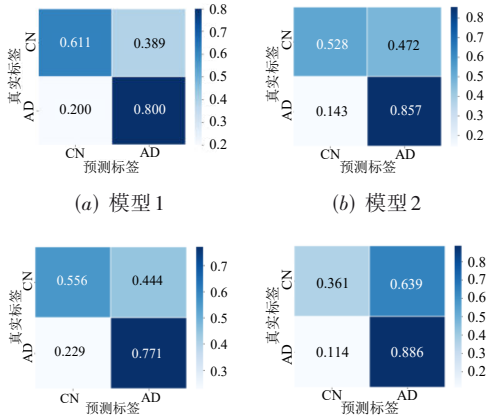
为较为全面地评价模型的分类效果,我们以准确率作为评价指标,并将采用Micro规则计算得到的精度、召回率、$F1$值作为补充指标.

表2中模型1的准确率比基线高5.63%,说明dVAE-BERT模型可以从语音中提取语言维度的表征,分辨出阿尔茨海默症患者和健康人.而从表2中模型1和模型2的结果,以及图4(*a*)和图4(*b*)的混淆矩阵可以看出,当dVAE模型的编码器层数增加时,结果准确率稍有下降.因此,我们对dVAE模型生成序列的能力进行分析.由图5可以看出,对于输入的同一段梅尔谱,当dVAE编码器为2层一维跨步卷积层时,生成的序列更加密集,从而BERT模型能够学习到长度较长的伪音素序列的高维表征.

表2　dVAE-BERT模型实验结果

编号	实验类型	模型	准确率	精度	召回率	*F*1值
模型1	基线实验	dVAE-2,BERT-6	0.704 2	0.666 7	0.800 0	0.727 3
模型2	参数对照实验	dVAE-3,BERT-6	0.690 1	0.638 3	0.857 1	0.731 7
模型3		dVAE-2,BERT-4	0.662 0	0.627 9	0.771 4	0.692 3
模型4		dVAE-2,BERT-8	0.619 7	0.574 1	0.885 7	0.696 6
模型5	消融实验	梅尔谱 + k-means + BERT	0.662 0	0.622 2	0.800 0	0.700 0
模型6	方法对照实验	eGeMAPS + SVM[27]	0.647 9	—	—	—
模型7		paralinguistic features + SVM[15]	0.647 9	0.916 7	0.314 3	0.468 1
模型8		MobileNetV2[35]	0.549 3	0.542 9	0.542 9	0.542 9
模型9		ResNet[36]	0.619 7	0.617 6	0.600 0	0.608 7
模型10		Wav2vec + BERT[37]	0.732 4	0.734 0	0.731 8	0.731 6

注:"dVAE-2,BERT-6"中,"2"表示dVAE模型的编码器为2层一维跨步卷积层,且每层一维跨步卷积后接一个残差块,"6"表示BERT模型使用了6层Transformer Encoder

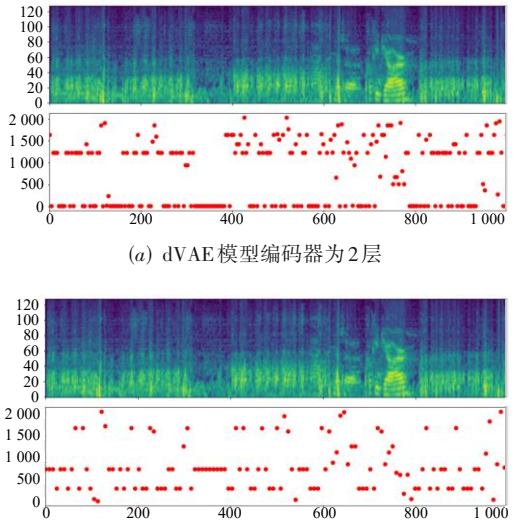


(*a*) 模型1　(*b*) 模型2　(*c*) 模型3　(*d*) 模型4

图4　表1中模型1至模型4的混淆矩阵

为了分析不同深度的BERT模型的学习能力,我们对比了Transformer Encoder层数不同时模型的表现.通过表2中模型1、模型3、模型4结果和图4(*a*)、图4(*c*)、图4(*d*)可以看出,当模型较浅时,准确率有所下降,可能是由于浅层模型学习能力不足,而当模型较深时,准确率下降较多,可能是由于BERT模型复杂度较高,而数据集较小,造成模型难以训练.

为验证dVAE模型生成的伪音素序列是否有效,我们使用模型5进行了对比.模型5是在提取训练集语音的梅尔谱特征后,使用k-means方法对特征序列进行聚类,类别数量设置为256,然后将梅尔谱生成的类别序列按照与dVAE-BERT模型相同的掩蔽方式输入BERT模型进行训练,而后进行阿尔茨海默症检测.从结果可以看出,使用



(*a*) dVAE模型编码器为2层

(*b*) dVAE模型编码器为3层

图5　输入梅尔谱与dVAE模型生成的序列之间的对应关系

dVAE模型生成的伪音素序列具有更好的检测效果.

模型6是INTERSPEECH ADReSSo比赛基线的结果,其主要使用支持向量机对eGeMAPS副语言特征集进行分类.模型7中,提取了梅尔频率倒谱系数、过零率等特征,形成了168维的特征集,而后使用支持向量机分类.这两种方法有一定的检测效果,可能是由于患者和健康人之间的副语言特征、统计学特征存在一定的差异.

表2的第8、9行分别为MobileNetV2、ResNet模型的结果,在提取音频的对数梅尔谱特征后,输入相应的神经网络进行分类,该方法对于阿尔茨海默症检测的

性能有限，可能是由于传统的卷积神经网络难以提取长距离的特征．第10行是文献[37]中的结果，首先使用Wav2vec语音识别模型将语音转录为文本，而后使用BERT模型提取文本的特征进行分类，该方法具有较高的准确率，说明对于阿尔茨海默症检测，从文本中提取语义信息是可行的，但是模型的性能很容易受到自动语音识别模型的影响[23]．本文提出的方法与模型10相比，结果有些差距，可能是由于dVAE-BERT模型分为多个训练阶段，而不同训练阶段的误差会不断累积，对最终结果造成影响，并且由于不同人的语言表达方式不太一致，而实验数据集较小，因此需要更多的训练数据提高模型的检测效果．

表3、表4分别为dVAE-BERT模型和Wav2vec2.0、HuBERT融合的结果．表中第2行、第3行分别表示dVAE-BERT模型与预训练模型融合方式为晚融合、早融合．图6、图7是分别是两种模型融合之后的实验结果对应的混淆矩阵．从中可以看出，dVAE-BERT模型在与Wav2vec2.0、HuBERT模型融合后，实验准确率有一定提升，说明融合后的模型，能够更深入地挖掘语音的潜在特征，且两种模型能够互补，为阿尔茨海默症检测提供更全面的信息．语音自监督预训练模型使用大量的语音数据进行训练，而dVAE-BERT模型仅使用ADReSSo数据集，就得到了与预训练模型相当的准确率，说明该模型能够较好的学习到伪音素序列的上下文信息．

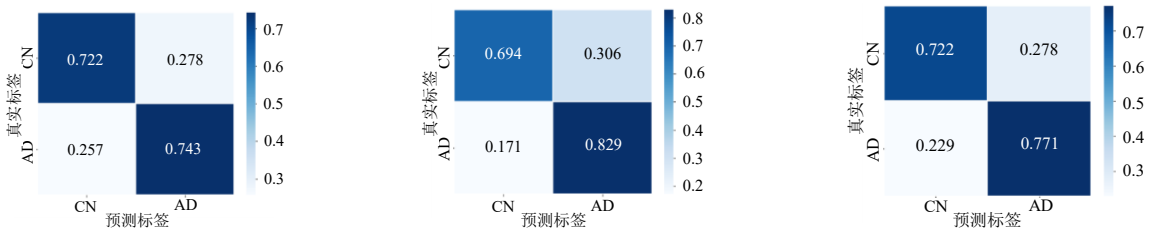
通过表3与表4可以看出，与Wav2vec2.0模型相比，HuBERT模型的准确率有所下降，并且融合后结果准确率也较低．HuBERT模型效果下降原因可能是两个模型的特征提取方式不同，HuBERT模型使用MFCC特征进行聚类，而后输入BERT模型，而Wav2vec2.0模型通过一维卷积方式提取类似Fbank的特征输入Transformer，提取的特征携带更多的上下文信息，对于阿尔茨海默症检测任务，可能使用携带上下文信息的特征更加有效．

表3 Wav2vec2.0模型与dVAE-BERT模型融合的结果

模型		准确率	精度	召回率	*F*1值
1	Wav2vec2.0	0.732 4	0.722 2	0.742 9	0.732 4
2	晚融合	0.760 6	0.725 0	0.828 6	0.773 3
3	早融合	0.746 5	0.729 7	0.771 4	0.750 0

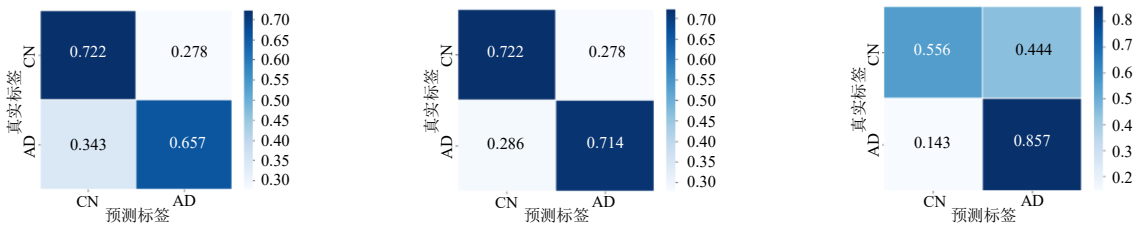
表4 HuBERT模型与dVAE-BERT模型融合的结果

模型		准确率	精度	召回率	*F*1值
1	HuBERT	0.690 1	0.697 0	0.657 1	0.676 5
2	晚融合	0.718 3	0.714 3	0.714 3	0.714 3
3	早融合	0.704 2	0.652 2	0.857 1	0.740 7



(*a*) Wav2vec2.0 (*b*) 晚融合 (*c*) 早融合

图6 Wav2vec2.0模型,及其与dVAE-BERT模型晚融合、早融合之后的混淆矩阵



(*a*) HuBERT (*b*) 晚融合 (*c*) 早融合

图7 HuBERT模型,及其与dVAE-BERT模型晚融合、早融合之后的混淆矩阵

4 结论

本文提出一种基于dVAE-BERT模型的阿尔茨海默症检测方法，使用dVAE模型将梅尔谱转换为伪音素序列后，再使用BERT模型对序列进行建模．该方法在INTERSPEECH2021 ADReSSo数据集上的准确率有明显的提升，表明从伪音素序列中提取的表征可以有效检测阿尔茨海默症，但是由于该方法分为多个训练阶段，因此训练时容易产生累积误差，并且由于不同地区、不同人的语言表达方式存在差异，从而需要更多的训练数据来提高模型的泛化性能．为进一步提高检测

效果,我们将该模型与Wav2vec2.0、HuBERT模型进行融合,结果表明,该方法提取的特征与预训练语音模型提取的特征具有互补性. 未来,我们将探索该模型对于不同语种的阿尔茨海默症检测的效果,实现多语种阿尔茨海默症检测.

参考文献

[1] JACK C R Jr, ALBERT M S, KNOPMAN D S, et al. Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease[J]. Alzheimer's & Dementia, 2011, 7(3): 257-262.

[2] VILLEMAGNE V L, BURNHAM S, BOURGEAT P, et al. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: A prospective cohort study[J]. The Lancet. Neurology, 2013, 12(4): 357-367.

[3] WORLD HEALTH ORGANIZATION. Dementia: A public health priority[EB/OL]. (2012-04-11) [2022-04-06]. https://www. who. int/mental_health/publications/dementia_report_2012/en/.

[4] POORE Q E, RAPPORT L J, FUERST D R, et al. Word list generation performance in Alzheimer's disease and vascular dementia[J]. Neuropsychol Dev Cogn B Aging Neuropsychol Cogn, 2006, 13(1): 86-94.

[5] REILLY J, PEELLE J E, ANTONUCCI S M, et al. Anomia as a marker of distinct semantic memory impairments in Alzheimer's disease and semantic dementia[J]. Neuropsychology, 2011, 25(4): 413-426.

[6] HOFFMANN I, NEMETH D, DYE C D, et al. Temporal parameters of spontaneous speech in Alzheimer's disease [J]. International Journal of Speech-Language Pathology, 2010, 12(1): 29-34.

[7] LUZ S, DE LA FUENTE S, ALBERT P. A method for analysis of patient speech in dialogue for dementia detection[EB/OL]. (2018-12-25) [2022-1021]. http://arxiv. org/abs/1811.09919.

[8] 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, 2020, 14(2): 272-281.

[9] LUZ S, HAIDER F, DE LA FUENTE S, et al. Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge[C]//Interspeech 2020. Shanghai: ISCA, 2020: 2172-2176.

[10] 陈旭初, 张卫强, 马勇. 基于原始波形的端到端阿尔茨海默症检测方法[J]. 电子学报, 2023, 51(12): 3582-3590. CHEN X C, ZHANG W Q, MA Y. Raw waveform-based end-to-end Alzheimer's disease detection method[J]. Acta Electronica Sinica, 2023, 51(12): 3582-3590. (in Chinese)

[11] EYBEN F, SCHERER K R, SCHULLER B W, et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing[J]. IEEE Transactions on Affective Computing, 2016, 7(2): 190-202.

[12] WARNITA T, INOUE N, SHINODA K. Detecting Alzheimer's disease using gated convolutional neural network from audio data[C]//Interspeech 2018. Hyderabad: ISCA, 2018: 1706-1710.

[13] CHIEN Y W, HONG S Y, CHEAH W T, et al. An automatic assessment system for Alzheimer's disease based on speech using feature sequence generator and recurrent neural network[J]. Scientific Reports, 2019, 9: 19597.

[14] ROHANIAN M, HOUGH J, PURVER M. Multi-modal fusion with gating using audio, lexical and disfluency features for Alzheimer's dementia recognition from spontaneous speech[C]//Interspeech 2020. Shanghai: ISCA, 2020: 2187-02191.

[15] BALAGOPALAN A, NOVIKOVA J. Comparing acoustic-based approaches for Alzheimer's disease detection [C]//Interspeech 2021. Brno: ISCA, 2021: 3800-3804.

[16] BAEVSKI A, ZHOU H R, MOHAMED A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[EB/OL]. (2020-06-20) [2022-10-21]. http://arxiv.org/abs/2006.11477.

[17] MIRHEIDARI B, BLACKBURN D, WALKER T, et al. Detecting signs of dementia using word vector representations[C]//Interspeech 2018. Hyderabad: ISCA, 2018: 1893-1897.

[18] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EBB/OL]. (2013-01-16) [2022-10-21]. http://arxiv. org/abs/1301.3781.

[19] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2014: 1532-1543.

[20] FRASER K C, MELTZER J A, RUDZICZ F. Linguistic features identify Alzheimer's disease in narrative speech

[J]. Journal of Alzheimer's Disease, 2016, 49(2): 407-422.
[21] MEGHANANI A, ANOOP C S, RAMAKRISHNAN A G. Recognition of Alzheimer's dementia from the transcriptions of spontaneous speech using fastText and CNN models[J]. Frontiers in Computer Science, 2021, 3: 624558.
[22] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Bag of tricks for efficient text classification[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Stroudsburg: Association for Computational Linguistics, 2017: 427-431.
[23] CODINA-FILBÀ J, CÁMBARA G, LUQUE J, et al. Influence of ASR and language model on Alzheimer's disease detection[EB/OL]. (2021-09-20)[2023-01-02]. https://arxiv.org/abs/2110.15704.
[24] RAMESH A, PAVLOV M, GOH G, et al. Zero-Shot text-to-image generation[EB/OL]. (2021-02-26)[2022-10-21]. http://arxiv.org/abs/2102.12092.
[25] ROLFE J T. Discrete variational autoencoders[EB/OL]. (2017-04-22)[2022-10-21]. http://arxiv.org/abs/1609.02200.
[26] DEVLIN J, CHANG M W, LEE K T, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL] (2019-05-24)[2022-12-30]. http://arxiv.org/abs/1810.04805.
[27] LUZ S, HAIDER F, DE LA FUENTE S, et al. Detecting cognitive decline using speech only: The ADReSSo challenge[C]//Interspeech 2021. Brno: ISCA, 2021: 3780-3784.
[28] HSU W N, BOLTE B, TSAI Y H H, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 29: 3451-3460.
[29] KINGMA D P, WELLING M. Auto-encoding variational bayes[EB/OL]. (2013-12-20) [2022-10-03]. http://arxiv.org/abs/1312.6114.
[30] GAUDER L, PEPINO L, FERRER L, et al. Alzheimer disease recognition using speech-based embeddings from pre-trained models[C]//Interspeech 2021. Brno: ISCA, 2021: 3795-3799.
[31] TU Y, LIN S, QIAO J, et al. Alzheimer's disease diagnosis via multimodal feature fusion[J]. Computers in Biology and Medicine, 2022, 148: 105901.
[32] CAMPBELL E L, DOCÍO-FERNÁNDEZ L, RABOSO J J, et al. Alzheimer's Dementia detection from audio and text modalities[A/OL]. (2020-08-11)[2023-01-03]. http://arxiv.org/abs/2008.04617.
[33] ILIAS L, ASKOUNIS D, PSARRAS J. A multimodal approach for dementia detection from spontaneous speech with tensor fusion layer[C]//2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). Piscataway: IEEE, 2022: 1-5.
[34] CUI C, YANG H C, WANG Y H, et al. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: A review[EB/OL]. (2022-03-25)[2022-10-21]. http://arxiv.org/abs/2203.15588.
[35] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018:
[36] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.
[37] ZHU Y X, OBYAT A, LIANG X H, et al. WavBERT: exploiting semantic and non-semantic speech using Wav2vec and BERT for dementia detection[C]//Interspeech 2021. Brno: ISCA, 2021: 3790-3794.

作者简介

陈旭初 男,1992年12月出生于河南省驻马店市. 现为清华大学电子工程系在读硕士研究生. 主要研究方向为音频事件检测、情感识别.
E-mail: chen-xc20@mails.tsinghua.edu.cn

蒲 钰 男,2001年6月出生于四川省成都市. 现为清华大学电子工程系在读硕士研究生. 主要研究方向为阿尔兹海默症检测.
E-mail: puy19@mails.tsinghua.edu.cn

张卫强 男,1979年1月出生于河北省雄县. 2002年于中国石油大学应用物理系获学士学位,2005年于北京理工大学电子工程系获硕士学位,2009年于清华大学电子工程系获博士学位,2017年斯坦福大学访问学者,现为清华大学电子工程系副研究员. 主要研究方向为语音与音频信号处理.
E-mail: wqzhang@tsinghua.edu.cn