病变语音识别：基于传统声学参数和发声参数的联合深度学习

一、 引言

语音参数作为以非侵入型方式获取的参数，在与发声相关的病证诊断和病症分类任务中发挥了重要的作用。在近年的多项研究中，均体现出语音参数对于病变判断的准确性和稳健性。人类的发声过程涉及到许多器官的配合，比如肺部、咽部、喉部、呼吸道、鼻腔、口腔等，许多病症的表现都涉及到了发声器官。病变状态下的发声器官所产生的语音，同正常的连续语流相比，会有明显的语音差异。这些由于病理所带来的语音变化，可以被多种声学参数量化为客观的语音指标，成为病症诊断的重要依据。语音参数的适用范围广泛，在多类发声相关病症的诊断中均得到了显著的成果，目前得到较多应用的病症主要包括：哮喘病、新冠肺炎、帕金森综合征、重度抑郁、流行性感冒、阿尔茨海默病等多类涉及发声改变的病症。

传统声学参数主要包括梅尔倒谱参数(MFCC)、梅尔频谱(Mel-Spectrogram)、声谱图(Spectrogram)和FBank等。部分学者从语音中提取梅尔倒谱参数(MFCC)、过零率(zero crossing rate)、均方能量(root-mean-square energy)等参数，采取深度神经网络和卷积神经网络等方法进行哮喘患者的识别，获得了较高的准确率。在新冠肺炎方面，有学者从新冠肺炎患者的语音中提取出梅尔倒谱参数和能量相关参数等，采用支持向量机和深度神经网络模型，获得了90%以上的识别准确率。短期傅里叶变换后的频谱特征也被纳入帕金森综合征的识别模型中，借助卷积神经网络的多层次卷积层和池化层，帕金森综合征患者的识别准确率达到89%，但少有研究纳入发声相关参数。

发声参数主要包括频谱斜度参数和非周期性参数，它可以衡量声门作为发声源的状态。根据欧洲喉科学会官方指南，发声参数确实可以稳定反映发声状态。频谱斜度参数可以体现声门的发声状态、开合状态、闭合程度等信息。非周期性参数反映的是声门不稳定性以及嗓音产出中的噪声成分。目前已有研究证明，病变语音的基频扰动和振幅扰动显著高于正常组别，噪声能量相比于常态发音人显著升高。病变患者的第一第二谐波差值、第一谐波与第一共振峰近邻谐波差值、第一谐波与第二共振峰近邻谐波差值等参数同正常发音人相比同样具有显著差异。

遗憾的是，发声参数并没有在病变语音识别的应用研究中得到广泛应用。本研究希望在病变语音识别使用的常见声学参数的基础上，引入语言学方向常用的发声参数

作为补充指标，以提升神经网络模型对于病变语音的识别准确率，将语言学研究和应用研究相结合，为深度学习背景下的病变语音研究提供新的研究思路和指标借鉴。

二、 实验设计

2.1 数据准备

本研究采用自建语音数据库的形式，选取了1260名语音病变患者和正常发音人的元音样本，作为实验材料。其中，正常发音人共计687名，病变发音人共计575名。男性发音人共计481名，女性发音人共计783名。录音信号为专业录音设备录制，信号的采样率和采样深度分别为50kHz和16bit固定值。录音过程有专业录音人员的陪同指导和即时反馈，每个录音的稳态发声元音在录音结束后被手工截取，单个音频样本的发声时长约为1.5秒左右，数据库总大小约为490MB。

2.2 参数提取

发声参数可以反映声门功能和声门状态，作为非侵入性的诊断方式，发声参数可以灵敏捕获到嗓音产出过程中的湍流，以及嗓音障碍的严重程度。发声参数跟人耳的听感效果密切相关，可以为发声相关病症提供重要参考。语音相关病症会极大程度地影响发声参数的数值表现，因此发声参数可以很好地衡量人体的病变状态和病变程度。本次研究中纳入的发声参数主要包括频谱斜度参数和非周期性参数两部分。

我们使用UCLA语音实验室提供的Voice Sauce软件以及美国MathWorks公司出品的MATLAB 2021软件来进行频谱斜度参数的提取，窗口长度设置为25毫秒，帧移为10毫秒。基频测算使用Straight算法，基频区间为40到500赫兹。共振峰和带宽使用Snack算法。音频预加重系数为0.96，线性预测编码系数为12阶。非周期性参数使用Python工具包OpenSMILE来进行提取，我们采用的特征集合为eGeMAPSv02，特征类别为Low Level Descriptors，帧移保持为10毫秒。

常用于病变语音识别的传统声学参数，主要包括梅尔倒谱参数(MFCC)、梅尔频谱(Mel-Spectrogram)、FBank等。我们使用Python工具包librosa来进行声学参数的提取，我们所采用的特征提取参数包括采样率22050赫兹，帧长1024个采样点，帧移256个采样点，傅里叶变换的维度为1024，频谱的维度为40维，预加重系数为0.97。此外，我们将训练集和测试集的比例划分为90%和10%。对于长度不足128帧的音频进行右侧补零，对于超过128帧的音频进行随机截断，以保证矩阵运算的输入时间维度同步。数据

处理阶段，我们租用了云服务器平台16核心32线程Intel Xeon Gold 6130，参数提取总耗时30小时。模型训练阶段租用11GB显存GeForce RTX 2080Ti，训练总耗时2小时。

2.3 发声参数详解

频谱斜度参数主要包括第一第二谐波差值(H1-H2)、第二第四谐波差值(H2-H4)、第一谐波与第一共振峰近邻谐波差值(H1-A1)、第一谐波与第二共振峰近邻谐波差值(H1-A2)，以及第一谐波与第三共振峰近邻谐波差值(H1-A3)。频谱斜度参数可以侧面体现出声门的发声状态，反映声门开合状态、闭合完全程度、声门闭合速度等多项信息。第一第二谐波差值(H1-H2)主要跟声门开放程度有关，第二第四谐波差值(H2-H4)体现了声带僵硬程度，第一谐波与第一共振峰近邻谐波差值(H1-A1)跟声门闭合的完全程度和声门后部开放程度有关，第一谐波与第二共振峰近邻谐波差值(H1-A2)以及第一谐波与第三共振峰近邻谐波差值(H1-A3)则是跟声门的闭合速度有关。

非周期性参数包括基频扰动(jitter)、振幅扰动(shimmer)、谐波噪声比(harmonics-to-noise ratio)、倒谱峰值显度(cepstral peak prominence)、声门激励强度(strength of excitation)。基频扰动(jitter)指的是相邻区段周期时长之差的绝对值，除以平均后的周期时长，它反映出的是基频在连续周期内的稳定性；振幅扰动(shimmer)指的是相邻区段振幅之差的绝对值，除以平均后的振幅值，它反映出的是振幅在连续周期内的稳定性。谐波噪声比(harmonics-to-noise ratio)主要指经过对数化计算后的声音内周期性成分能量与非周期性成分能量的比值。谐波噪声比越高，声音周期性成分能量越多。倒谱峰值显度(cepstral peak prominence)则是通过将声学频谱通过逆傅里叶变换的方式转换为倒谱域，然后计算基频所对应谐波的最大峰值减去拟合的倒频谱曲线，即可得到倒谱峰值显度，它可以反映出高频噪声成分的大小。声门激励强度(strength of excitation)主要反映的是声门激励源的能量大小。

表1 发声参数汇总

Phonation Parameters			
基频扰动	Jitter	第一第二谐波差值	H1-H2
振幅扰动	Shimmer	第二第四谐波差值	H2-H4
谐波噪声比	HNR	第一谐波与第一共振峰近邻谐波差值	H1-A1
倒谱峰值显度	CPP	第一谐波与第二共振峰近邻谐波差值	H1-A2
声门激励强度	SOE	第一谐波与第三共振峰近邻谐波差值	H1-A3

2.4 传统声学参数详解

常用于病变语音识别的传统声学参数，主要包括梅尔倒谱参数(MFCC)、梅尔频谱(Mel-Spec)、Fbank。我们首先要对语音信号进行预加重，高频信号在传递过程中衰减快，但是高频部分又蕴含很多对语音识别有利的特征，预加重可以提高高频部分的能量。考虑到语音信号的长时非平稳性和短时平稳性，因此需要分帧来使得语音信号表现出明显的稳定性和规律性。在傅里叶变换之前，我们需要对信号进行加窗，将窗函数与信号在时域相乘以抑制旁瓣的频谱泄露。

实际的语音信号非常复杂，我们需要以适当的频率间隔，将语音信号分解为一组组的基础信号，然后计算出每组信号的幅度和相位。傅里叶变换得到每个频带上信号的能量后，我们仍需要进行继续的处理过程。考虑到人耳对于频率的感知是非线性的，傅里叶变换后的频谱需要与梅尔滤波器组相乘，这样得到的频谱数据更接近人耳的听觉感知。经过梅尔滤波器组之后，我们所得到的就是梅尔频谱，如下图所示。如果继续对梅尔频谱取对数，那么就可以得到Fbank特征。

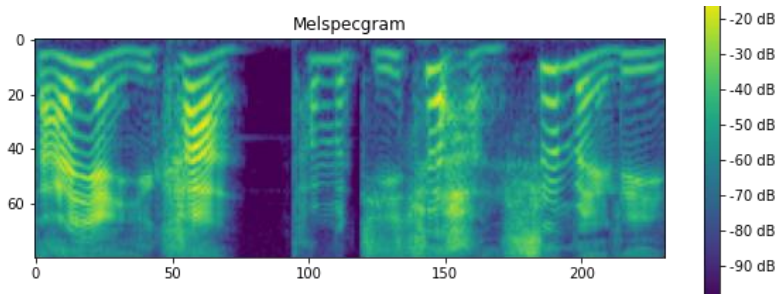

图 1 梅尔频谱图

在早期的语音识别和语音合成任务中，由于受到计算能力的限制，梅尔频谱和Fbank特征的维度较高导致运算耗时过长，因为有学者提出了梅尔倒谱参数(MFCC)来降低特征维度和加快计算速度。梅尔倒谱参数还可以减少特征的内部关联性，后续被用于许多任务中。梅尔倒谱参数的本质是对梅尔频谱和Fbank特征进行 (傅里叶逆变换，原频域上变化慢的共振峰包络会占据新频域上低频部分，原频域上变化快的频谱细节会占据高频部分，这样就可以分离共振峰包络和频谱细节。

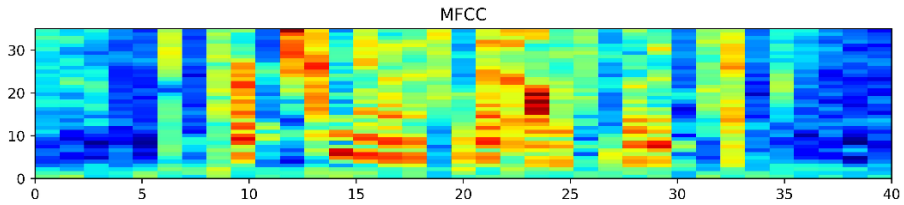

图 2 梅尔倒谱参数图

三、 模型简介

我们的研究中共使用了三个深度学习模型来进行病变语音识别，三个模型分别为卷积神经网络(CNN)、卷积神经网络-长短时记忆(CNN LSTM)、卷积神经网络-双向长短时记忆(CNN Bi-LSTM)。接下来我们对模型的各个组成模块进行解读，首先是模型的输入和输出。我们的模型输入包括两类语音参数，即前文介绍的声学参数和发声参数。为了验证发声参数能否提升病变语音识别的准确率，我们采取对照实验的方式，先加入声学参数(MFCC、Fbank、Melspec)，然后再加入声学参数和发声参数合并而成的联合参数(MFCC + Phonation、Fbank + Phonation、Melspec + Phonation)，比较相同模型下引入发声参数后，模型的识别准确率是否有所提高。考虑到不同类别参数之间的量纲不同，我们在参数合并阶段采取各自归一化的方式，将声学参数和发声参数按照各自的均值的方差进行归一化处理，然后再进行特征合并。

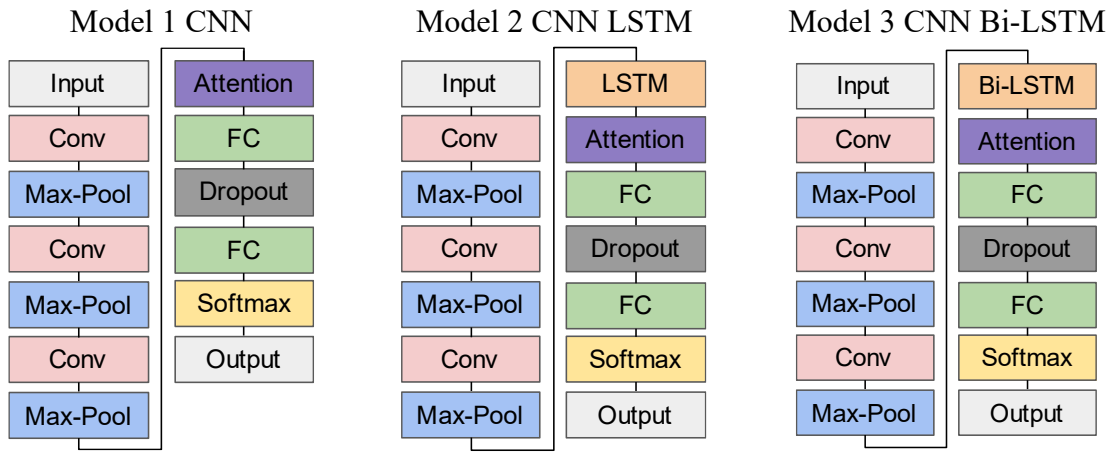

图 3 模型训练流程图

在卷积网络模型(CNN)中，如果我们的输入特征维度为(32, 128, 40)，经过三次卷积层和最大池化层后，特征的维度变为(32, 32, 16, 5)。其中多出来的维度32是由卷积层多个卷积核带来的通道数，输入特征的后两维在经过三次最大池化后，缩减为原来的1/8维度。接下来，我们将特征按照时间序列展开为(32, 16, 160)，采用平均注意力机制确定各时间点所对应特征的注意力，并按照注意力权重进行加权求和。最后，我们采取全连接层和随机失活层逐步缩小神经元节点，以SoftMax层输出二分类的概率(病变或是正常)，随机失活层可以有效抑制训练过程中的过拟合问题。

卷积神经网络-长短时记忆(CNN LSTM)和卷积神经网络-双向长短时记忆(CNN Bi-LSTM)多一层循环神经网络层，卷积和池化后的特征并不直接进入注意力层，而是

按时间序列展开后为(32, 16, 160)后经过长短时记忆变为(32, 16, 256)，其中维度256是我们预先规定的隐含特征维度。在经过长短时记忆层后，我们的特征会进入注意力层，获得各时间点处的向量权重分布并进行加权求和，最后经过全连接层和随机失活层逐步缩小节点，以获得二分类输出概率。

对于模型的损失，我们采取交叉熵损失作为损失函数，依照样本真实标签和SoftMax二分类概率进行模型预测过程的损失计算，并将误差以梯度下降的方式传递给前方神经元节点。模型采用的优化器是Adam优化器，同时还采用了学习率衰减方式来更新优化器的学习率，每隔200步存储一次模型，每隔100步学习率缩减为上阶段学习率的0.95倍，同时每隔100步对于测试集的音频样本进行评估，记录预测成功的概率和模型的损失，保存为日志文件。为了保证模型的可复现性，我们采取随机种子锁定的方式，锁定训练过程中数据集切分、模型参数初始化、批数据读取、随机失活层等具有随机性的过程。

四、 结果和不足

4.1 实验结果

在进行 2000 步训练后(约 20 轮次迭代)，我们将模型结果汇总到下方表格，标红的部分为加入发声参数合并而成的联合参数。从训练结果来看，我们的猜想得到了验证，发声参数的加入可以提高模型的判断准确率。就卷积神经网络的结果来看，相比于单独的声学参数，发声参数的加入使得模型的准确率上升了 1%到 4%左右。在卷积神经网络后接单向长短时记忆的结果来看，发声参数的加入带来的准确率上升不是很明显，大约在 1%到 2%左右。提升效果最明显的模型是卷积神经网络后接双向长短时记忆，发声参数的加入使得模型准确率上升了 6%到 8%左右。实验结果说明发声参数的加入可以提高模型的判断准确率，如果我们未来继续加深网络和精细化调整训练的超参数，我们相信模型会有更好的表现。

表2 模型训练结果

参数类别	CNN	CNN_LSTM	CNN_Bi-LSTM
Fbank	66.32	66.58	63.16
Fbank+Phonation	67.63	67.63	69.21
MelSpec	66.84	66.58	62.11
MelSpec+Phonation	67.37	68.69	70.79
MFCC	65.53	66.32	67.37
MFCC+Phonation	70.79	66.84	73.16

4.2 不足之处

相比于现有病变语音识别模型 80%到 90%准确率，我们的最高准确率为 73.16%，距离我们的理想情况仍然有所差距，我们需要再提升 7%到 10%左右，才能基本与表现较好的现有模型持平。另外，我们还可以在未来采取数据增强的方式，比如说对原始数据进行数据增广，通过变速变调等方式将小体量的数据库进行扩充，以提高模型的识别准确率和泛化能力。除此之外，语言学研究中还有很多可以使用的参数，比如说共振峰、声门开商、速度商等参数。如果加入更多语言学相关的参数，模型的准确率应该可以进一步提升。

五、 结语

本研究在常见声学参数的基础上，引入发声参数作为补充指标，以提升神经网络模型对于病变语音的识别准确率。我们构建了三个神经网络模型，以卷积神经网络(CNN)、卷积神经网络-长短时记忆(CNN LSTM)、卷积神经网络-双向长短时记忆(CNN Bi-LSTM)来验证我们的猜想，通过比较发声参数加入所带来的准确率变动来查看发声参数的贡献度。实验结果证明，发声参数的加入使得模型准确率最高上升了 6%到 8%左右，我们的最高准确率达到了 73.16%。

通过将语言学研究和应用研究相结合，我们的研究为深度学习背景下的病变语音研究，提供了新的研究思路和指标借鉴。通过引入更多语言学特征来辅助深度学习的模型，以特征融合的方式进行模型训练，可以获得更加优良的模型表现。语言学研究中还有更多可以引入的参数，我们在未来的研究中可以继续探索。除此之外，针对病变语音数据的稀缺性和难获得性，我们提出了数据增广的预处理设想，通过变速变调等形式合成近似语音来扩充数据集的不足，同时还能提高模型的泛化能力。