基于深度学习的病理语音二分类研究

余颖轩
西京学院，中国 ·陕西 西安 710123

摘要：嗓音疾病的临床诊断长期依赖于喉镜观察与医师的主观诊断，传统喉镜检查虽应用广泛，但存在侵入性强、基层医疗采纳度低等局限。本文提出了一套基于深度学习方法的二分类框架，旨在通过常规语音信号对健康个体与声带病变患者进行区分。该框架首先提取梅尔频率倒谱系数、基频扰动率以及谐波噪声比等声学参数，随后采用卷积神经网络与长短时记忆网络相融合的混合结构进行分类建模。在公开的病理语音数据库上进行评估后，该方法实现了 92.7% 的分类准确率。

关键词：病理语音识别；二分类问题；深度学习；声学参数；卷积神经网络

Research on Pathological Speech Binary Classification Based on Deep Learning

Yu Yingxuan
Xijing College, China Shaanxi Xi'an 710123

Abstract: The clinical diagnosis of vocal disorders has long relied on laryngoscopic observation and physicians' subjective assessment. Although traditional laryngoscopy is widely used, it has limitations such as high invasiveness and low adoption in primary healthcare. This study proposes a binary classification framework based on deep learning methods, aiming to distinguish between healthy individuals and patients with vocal cord lesions through conventional speech signals. The framework first extracts acoustic parameters including Mel-frequency cepstral coefficients, fundamental frequency perturbation rate, and harmonic-to-noise ratio, then employs a hybrid structure combining convolutional neural networks and long short-term memory networks for classification modeling. Evaluated on a publicly available pathological speech database, this method achieved a classification accuracy of 92.7%.

Keywords: Pathological speech recognition; Binary classification problem; Deep learning; Acoustic parameters; Convolutional neural network

1 引言

1.1 研究背景及意义

声带相关疾患在耳鼻喉科诊疗中十分常见，并且其发病率近年来呈现持续上升的态势。目前，该领域的诊断金标准仍为喉镜检查，但是此项检查具有侵入性，往往给患者带来一定程度的痛苦与不适，同时检查费用较高，难以适用于大规模群体的初期筛查。病理语音通常表现出气息声、粗糙感或嘶哑度等听觉特征，而这些特征可以使用基频抖动幅度、振幅扰动程度以及谐波噪声比例等客观指标来予以量化。因此，基于语音信号的无创分类方案具备双重优势：一方面能够消除患者的身体痛苦与检查风险，另一方面借助智能手机等常见终端即可实现低成本的早期筛查。

1.2 相关研究回顾

Godino-Llorente 及其团队在提取 MFCC 特征后引入高斯混合模型进行分类，在 MEEI 数据集上获得了约 85% 的准确率。 Uloza 等人则采用多参数回归模型开展分析，研究结果表明基频扰动和谐噪比能够有效区分声带息肉患者的语音与正常语音。

Fang 等人通过一维卷积神经网络直接对原始语音波形进行处理，从而规避了人工特征设计过程中可能出现的主观偏差。 VásquezCorrea 等人则将 CNN 与 RNN 结合起来加以利用，借助时序依赖关系提升了模型的整体稳健性。尽管如此，现有工作中仍有两个亟待解决的问题： 一是特征融合的策略较为简单，二是面对小样本数据集时容易出现过拟合现象。本文在上述研究的基础上，设计了一种轻量化的混合神经网络架构，同时引入了多级特征融合机制，以提升病理语音二分类的性能。

2 病理语音特性分析

2.1 发声生理机制与病变影响

在健康发声状态下，声带会在气流的驱动下呈现周期性的开闭运动，从而产生近似周期性的声门脉冲信号。声

带的任何病变都会破坏这种周期结构：比如，声带小结或息肉会导致一侧声带出现异常的质量负荷，进而引发振幅调制现象。上述异常在时域波形上体现为周期性的破坏，在频域谱图上则表现为谐波结构模糊以及噪声本底抬升。

2.2 核心声学参数

（1）梅尔频率倒谱系数：能够模拟人耳听觉系统在不同频率上的非线性分辨特征，对声道滤波效应具有良好的紧凑表达能力。当前最常用的短期谱特征维数为 12 维或 13 维。

（2）基频相关指标：包括基频的整体均值、抖动（反映相邻周期之间的微小波动）以及闪烁（反映振幅扰动变化）。

（3）谐波噪声比：指语音信号中谐波成分的能量与宽带噪声能量的比率。如果 HNR 较低，往往提示声门闭合不充分或存在明显的湍流噪声。

（4）梅尔频谱图：将 MFCC 信息以二维图像形式呈现出来，能够保留语音信号在时间维度上的动态演化过程，非常适合输入到卷积网络中进行处理。

2.3 特征融合方式

单纯使用帧级别的特征难以完整捕捉病理模式所对应的全局结构。为了解决这一问题，本文采用了一种分层融合策略：首先在每一帧上提取 MFCC、jitter、shimmer 和 HNR，构成底层特征向量；与此同时，对整段语音计算这些特征统计量的全局平均值和方差，形成高层统计特征；最后，将上述两类特征在通道维度上进行拼接，再送入分类器。

3 研究方法

3.1 系统总体框架

（1）数据预处理：包括噪声抑制、端点检测以及分帧加窗操作。

（2）特征提取：主要包括 MFCC、jitter、shimmer、HNR 以及梅尔频谱图。

（3）混合神经网络：其中 CNN 分支负责提取局部谱模式，LSTM 分支用于捕获长程时序依赖。

（4）决策融合与分类：将双分支输出的特征向量进行拼接，再经全连接层输出健康 / 病理两类概率。

3.2 数据预处理流程

将原始语音信号重新采样至 16 kHz。采用双门限端点检测方法定位出有效的语音片段，去除静音段和呼吸声所造成的干扰。对信号进行分帧时选择 25 ms 的汉明窗，帧移设为 10 ms，从而在相邻帧之间保留一定的连续性。

3.3 具体特征提取过程

MFCC：为每一帧提取 13 维倒谱系数，再结合其一阶差分与二阶差分，构成 39 维的特征矢量。扰动参数：首先通过自相关法估计基频轨迹，随后计算 jitter（反映周期相对扰动）以及 shimmer（反映振幅相对扰动）。HNR：采用互相关法计算谐波成分能量与残差噪声能量的比值。梅尔频谱图：对每一帧的能量在 128 个梅尔频率滤波器组上进行投影，最终得到一幅维度为（时间 × 频率）的二维图像。

3.4 网络结构设计

在 CNN 分支的设计上，三个卷积模块被依次级联起来。每一个模块内部均包含一个卷积层（选用 3×3 的卷积核，激活函数为 ReLU）、 一个批归一化层以及一个最大池化层（池化窗口尺寸为 2×2）。该分支以梅尔频谱图作为输入，经过前述模块处理后输出的特征图会经历一次全局平均池化操作，从而将特征维度压缩至 256 维。

至于 LSTM 分支，其输入形式为由帧级 MFCC 及其动态差分特征共同构成的时序序列。该分支采用了双层双向 LSTM 结构，每一层的隐藏单元数量设定为 128 个，网络最终仅保留最后一个时间步的输出向量。考虑到小样本数据集容易诱发过拟合问题，在特征向量进入全连接层之前，我们设置了 0.5 的 dropout 比率，同时对 LSTM 层施加了循环正则化约束。

在融合阶段， CNN 分支输出的 256 维特征与 LSTM 分支输出的 256 维特征于特征通道上进行拼接，由此得到一个 512 维的联合表征向量。该向量随后依次经过两层全连接网络（维度变化路径为 512 → 128 → 2），并借助 Softmax 函数产生最终的类别概率预测。训练过程中，模型以交叉熵作为损失函数，选用 Adam 优化器进行参数更新，初始学习率设定为 0.001，同时采用余弦退火机制对学习率进行衰减。

4 数据集说明

实验中使用了公开可用的数据库： MEEI 数据库：包含 53 例健康语音（以及 173 例不同类型声带病变的语音样本。为了专注于二分类任务，将所有病理亚型归并为正类，健康样本归为负类。按照 7:2:1 的比例将数据划分为训练集、验证集和测试集，并且确保同一说话人的样本不会出现在不同集合中，以避免数据泄露问题。

5 结果与分析

下表汇总了各模型在 MEEI 测试集上的性能表现。本文所提出的混合模型取得了 92.7% 的准确率以及 0.967 的 AUC 值，明显优于各种传统的基线方法。与单一分支结构相比，混合模型的 F1 分数相比纯 CNN 模型提升了 3.2 个百分点，相比纯 LSTM 模型提升了 4.5 个百分点，这一结果表明局部谱特征与时序动态特征之间具有良好的互补

性。在 Saarbruecken 数据库上，模型的准确率达到 89.3%，略低于 MEEI 数据库上的结果。分析认为，差异可能源于 Saarbruecken 数据库中包含语句级别的语音，而持续元音的发声状态更为稳定，病变信息也更为突出。

模型	准确率(%)	精确率(%)	召回率(%)	F1(%)	AUC
SVM (RBF)	78.4±2.1	80.2	77.5	78.8	0.851
随机森林	76.9±1.8	78.4	75.9	77.1	0.839
单分支 CNN	87.2±1.5	87.9	86.1	87.0	0.928
单分支 LSTM	86.0±1.9	85.4	86.7	86.0	0.915
本文混合模型	92.7±1.2	93.5	92.0	92.7	0.967

6 结语

本文针对声带病变二分类任务，提出了一种融合卷积神经网络与循环神经网络的混合架构。通过在 MEEI 和 Saarbruecken 两个公开数据库上进行系统评估，得出以下几点结论：

（1）将梅尔频谱图与时序 MFCC 特征相结合的双分支结构，与仅使用单一特征相比，能够显著提高分类准确率，最终 F1 分数达到 92.7%，验证了局部谱模式与全局时序动态之间的互补效果。

（2）谐噪比和基频扰动是贡献度最大的两个声学指标，这与病理语音的生理生成机制保持一致，同时也说明模型确实学到了具有临床意义的相关特征。

（3）低信噪比条件下性能出现较为明显的下降，这一现象提示我们：对于实际部署而言，模型鲁棒性仍然是主要的瓶颈，需要引入更先进的噪声抑制技术或领域泛化策略。

综上所述，基于深度学习的病理语音二分类在经过仿真临床条件的测试后，已经展现出了较好的可行性，可以作为喉镜检查之前的一种低门槛预筛工具。下一步工作将聚焦于真实环境下的数据收集以及模型可解释性的提升，从而加速该技术在临床应用中的转化。

参考文献：

[1] 常静雅，张晓俊，顾玲玲等. 小波域能量谱和非线性降维的病理嗓音识别[J]. 计算机工程与应用，2017, 53(2): 166-171.

[2] 陈益，武倩文，姜羽菲等. 基于相关互补非线性特征融合的嗓音疾病分类[J]. 电子设计工程，2024, 32(21): 18-22.

[3] 陈英. 基于语音反演机器学习方法的声道模型研究[D]. 南京邮电大学，2013.

[4] 顾玲玲. 病理喉声源发声系统模型的非线性分析及其优化研究[D]. 苏州大学，2016.

[5] 姜羽菲，石宇，何若男等. 基于卷积神经网络特征提取的病理语音识别[J]. 电子设计工程，2024, 32(20): 26-30.