量子卷积神经网络在医疗保健应用语音任务中的鲁棒性量化

Ha Tran[1], *IEEE*学生会员，Bipasha Kashyap[1], *IEEE*会员，Pubudu N. Pathirana[1], *IEEE*高级会员

arXiv:2601.02432v1 [cs.SD] 5 Jan 2026

摘要——**基于语音的机器学习系统对噪声敏感，这使其难以可靠地应用于情感识别和嗓音病理检测。我们在干净训练/受扰测试的设置下，评估混合量子机器学习模型——量子卷积神经网络（QNN）相对于经典卷积神经网络（CNN）在四种声学扰动（高斯噪声、音高偏移、时间偏移和速度变化）下的鲁棒性。我们使用 AVFAD（嗓音病理）和 TESS（语音情感）数据集，通过准确率和扰动指标（CE、mCE、RCE、RmCE），将三种量子神经网络模型（随机、基础、强）与简单的 CNN 基线模型（CNN-Base）、ResNet-18 和 VGG-16 进行比较，并分析架构因素（电路复杂度或深度、收敛速度）以及各情感类别的鲁棒性。量子神经网络在音高偏移、时间偏移和速度变化下总体优于 CNN-Base（严重时间偏移下 CE/RCE 最多降低 22%），而 CNN-Base 对高斯噪声的抵抗力更强。在量子电路中，QNN-基础在 AVFAD 上的整体鲁棒性最佳，QNN-随机在 TESS 上表现最强。按情感类别来看，恐惧 最具鲁棒性（在严重扰动下准确率为 ≈ 80–90%），中性在强高斯噪声下可能完全失效（准确率为 ≈5.5%），而快乐 最容易受到音高、时间和速度失真的影响。量子神经网络的收敛速度也比 CNN-Base 快至六倍。据我们所知，这是首项针对语音在常见非对抗性声学扰动下的 QNN 鲁棒性开展的系统研究，表明浅层纠缠量子前端能够提升抗噪能力，但对加性噪声的敏感性仍是一项挑战。**

索引词—**量子机器学习、语音情感识别、语音病理检测、鲁棒性**

一、引言

机器学习（ML）方法的强大能力推动了许多任务取得重大进展，例如自动语音识别 [1], [2],说话人识别 [3], 语音情感识别 [4], 和嗓音病理检测[5], [6]。然而，与人类不同，这类机器学习方法对噪声数据可能非常敏感。在现实世界的部署中，这种敏感性尤为突出，因为声学信号会受到加性、频谱、时间和信道扰动的影响。在日常语音通信中，声学信号会受到各种损坏的影响，包括背景噪声、混响、信道失真，以及说话人引起的变化，例如音高偏移、速度变化和韵律波动。这些损坏是开发适用于语音任务的有效机器学习框架时面临的主要挑战之一。这一点尤其适用于旨在提取健康相关信息并辅助诊断的医疗应用，例如语音情感识别和嗓音病理检测。语音情感识别旨在识别说话人的情绪状态，而嗓音病理检测则侧重于利用声学线索检测嗓音异常，并采用信号处理和机器学习方法。这两种应用常常需要应对噪声数据，且在模型不足时性能可能下降。因此，这些系统需要的不仅是准确的机器学习框架，还要具备抗噪能力。

许多研究分析了机器学习在含噪信号上的鲁棒性，并提出了减轻噪声影响的技术。许多技术旨在设计能够抵抗噪声的特征，例如信号级降噪 [7]–[9]。本研究不着重于设计鲁棒特征，而是通过设计合适的模型架构来提升鲁棒性。具体而言，我们不使用带有鲁棒特征的噪声数据训练机器学习模型，而是使用干净数据训练模型，并评估其对未见过的受损数据的鲁棒性。这种方法有助于提升模型在真实世界数据上的韧性和一致性；这些数据会受到各种类型噪声的干扰。这对于处理多样化噪声数据（例如医疗数据）且要求高度可靠性的应用很有益处。研究人员指出，深度学习模型的鲁棒性源于其学习不变表示的能力。例如，卷积神经网络和循环神经网络在处理噪声环境中的识别任务时，已在噪声数据上展现出良好表现 [10], [11]。残差网络（ResNet）采用残差连接，使模型能够聚焦于相关的时频模式、缓解梯度消失问题并抑制无关噪声 [12], [13]。VGG 类卷积神经网络等架构则进一步利用分层特征提取来增强这种鲁棒性 [14]。

近年来，量子机器学习（QML）作为一种颇具前景的技术应运而生，它结合了机器学习的灵活性与学习能力、量子算法的鲁棒性以及量子硬件的高效性。QML旨在提升机器学习方法在多种语音应用中的安全性、计算效率和准确率。研究表明，QML方法在语音识别 [15], [16],和语音情感识别 [17]–[19]方面的表现优于经典机器学习方法。研究人员还证明，尽管目前仅在有限的语音数据集上进行验证，QML仍能取得更优的性能；而这本身就是许多应用（如医疗保健）面临的一项重大挑

[1]所有作者均来自迪肯大学工程学院网络感知与生物医学工程（NSBE）研究实验室，澳大利亚维多利亚州沃恩庞兹（电子邮箱：thi.n.tran@deakin.edu.au、b.kashyap@deakin.edu.au、pubudu@deakin.edu.au

This 本文是目前正在 IEEE 音频、语音与语言处理汇刊接受评审的预印本

战 [16]。然而，与经典机器学习相比，QML在语音领域的应用仍十分有限。此外，在语音等需要具备良好抗噪能力的应用中，QML的鲁棒性尚未得到探索。我们开展这项研究，正是因为缺乏对QML在语音领域鲁棒性的分析。

在本研究中，我们针对语音领域的量子机器学习模型提出了一个鲁棒性评估框架。我们的主要贡献如下：

- 两个语音任务上的鲁棒性评估：我们在干净训练/损坏测试设置下，系统评估了量子神经网络在语音病理检测（AV-FAD）和语音情感识别（TESS）任务上的表现。对于情感识别，我们进一步分析了类别级鲁棒性，以确定在每种扰动下哪些情绪保持稳定，以及哪些情绪的性能下降最为明显。
- 四种扰动类型下的量子—经典比较：我们使用准确率和扰动指标，在四种非对抗性声学扰动（高斯噪声、音高偏移、时间偏移和速度变化）的六个严重程度下，对量子神经网络和卷积神经网络进行了比较。
- 量子电路设计对鲁棒性的影响：我们量化了量子电路及电路深度对鲁棒性的影响，报告了架构特定的权衡，并确定了在每个数据集上实现最佳鲁棒性表现的配置。

本研究的其余部分组织如下。第二节介绍量子机器学习算法及其鲁棒性的相关研究。接下来的第三节阐释量子计算的基本概念和三种量子电路。第四节介绍语音信号处理和模型架构。第五节描述数据集、四种破坏类型、训练和测试流程。第六节展示结果，并对分类准确率和鲁棒性进行讨论。最后，第七节总结主要发现。

II. 相关工作

A. 量子机器学习的效率

许多研究探讨了广受关注的量子机器学习算法，并证明它们在语音处理方面表现出良好的效果，例如量子支持向量机 [16]，量子神经网络 [20]，量子卷积神经网络 [17]。为结合卷积神经网络的表达能力和鲁棒性，以及量子机器学习的效率，研究人员探索了混合量子机器学习方法。这类混合方法利用经典模型和量子机器学习模型各自的优势来提升性能。一个值得关注的例子是混合量子卷积神经网络（QNN）模型 [21]。通过在卷积变换中利用量子力学的特性，QNN 旨在进一步提升机器学习应用中的特征提取能力和整体信息处理水平 [15], [22], [23]。因此，有必要研究 QNN 在各种机器学习应用中的鲁棒性。

B. 量子机器学习的鲁棒性

含噪中等规模量子设备（NISQ） [24]目前仍受到设备产生的噪声影响，例如退相干、门错误以及环境干扰。许多研究利用量子框架 [25], [26]，致力于减轻含噪中等规模量子设备中的这类噪声。一项研究 [27] 考察了 QNN 对图像数据中量子噪声的鲁棒性，结果表明，与其他量子机器学习模型相比，QNN 具有更高的鲁棒性。这展现了 QNN 抵御噪声的潜力。然而，本研究关注的是干扰输入信号的噪声，而非量子设备产生的噪声。

我们首先考察对抗攻击，这类攻击会故意扰动输入数据。对抗攻击是指向输入信号添加微小且有意的扰动。这类噪声通常难以被人类察觉，但可能导致机器学习模型做出错误预测。已有研究探讨了量子模型抵御对抗扰动的鲁棒性，不过大多数研究都聚焦于图像数据。例如， [28]，指出，量子机器学习与经典机器学习类似，也容易受到特定类型的对抗攻击影响。在 [29], 中，作者发现，基于量子机器学习构建的变分分类器能够学习经典神经网络无法检测到的特征。这些额外特征能够提升量子分类器的鲁棒性，减轻对抗攻击的影响。然而，他们的方法计算成本极高，需要 $10^3 - 10^4$ 个量子门，超出了当前量子硬件的能力范围。在 [30], 中，研究人员评估了量子神经网络抵御三种对抗攻击的鲁棒性。研究表明，与经典卷积神经网络相比，增强量子神经网络的量子纠缠特性可以提升其鲁棒性。尽管成像领域已有许多关于对抗攻击的研究，但针对量子机器学习在语音处理应用中所受影响的研究仍然有限。例如， [31]研究了量子神经网络在白盒攻击和黑盒攻击下保护视听语音隐私的能力，但只分析了一种量子电路配置。语音应用中其他量子电路架构的鲁棒性仍未得到充分探索。

除了对抗扰动之外，另一种会影响输入信号的噪声是由现实世界中的变化产生的随机噪声。这可能包括背景噪声、说话人之间的差异以及处理过程中的变化。 [32] 提出了受量子性质启发、用于声学场景分类的 Transformer 模型。在广泛的声学环境中，通过在受到白高斯噪声扰动的语音数据上进行评估，结果表明，与经典模型相比，所提出的 Transformer 模型具有更强的鲁棒性和更高的准确率。与声学场景分类不同，我们的研究关注影响人类语音的自然语音损坏。据我们所知，这是首项研究语音分类任务中量子神经网络抵御自然声学扰动能力的研究。本研究旨在建立实验基线，以便在现实的输入退化条件下比较量子与经典机器学习方法的鲁棒性。

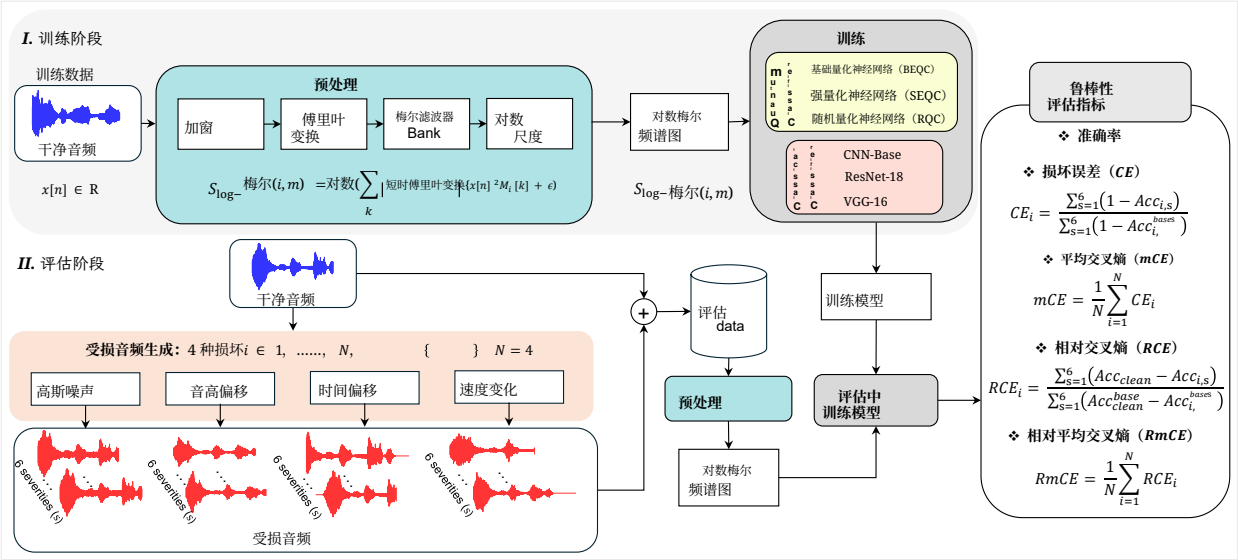


图 1：所提出的用于语音分类的量子-经典鲁棒性评估框架。干净音频被转换为对数梅尔频谱图，并用于训练量子神经网络模型（QNN-Basic、QNN-Strongly、QNN-Random）和经典卷积神经网络（CNN-Base、ResNet-18、VGG-16）分类器。使用交叉熵、mCE、RCE 和 RmCE 等指标，在受损音频（高斯噪声、音高偏移、时间偏移、速度变化）上评估模型鲁棒性。

C. 研究空白与动机

现有研究尚缺乏：（1）对用于语音处理的量子电路架构进行系统比较；（2）对量子模型在自然声学扰动下的鲁棒性进行全面分析。这促使我们评估电路设计如何影响现实语音条件下的 QNN 鲁棒性。量子神经网络通过空间局部量子变换和适用于含噪中等规模量子设备的浅层电路，为处理高维语音数据带来优势。其混合架构能够与经典深度学习无缝集成，同时实现量子增强特征提取。受 [33]，（其对图像模型的鲁棒性进行了基准评测）的启发，本研究考察 QNN 在面对四种常见声学扰动时的鲁棒性：高斯噪声、音高偏移、时间偏移和速度变化。

III. 背景

量子机器学习（QML）是量子计算与机器学习（ML）的结合。因此，量子机器学习利用量子特性（如叠加和纠缠）来增强机器学习。本节将介绍量子计算的基础知识以及三种量子电路类型。

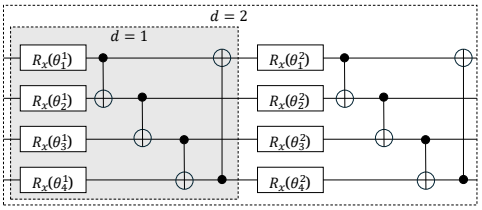
A. 量子计算基础

量子计算利用量子力学原理执行计算 [34]。与经典计算中的比特类似，量子比特（qubit）是量子计算的基本概念。不过，它可以处于 0 和 1 的线性组合状态。这种叠加构成了 Grover 搜索算法等算法实现量子加速的基础 [35]。在量子机器学习中，叠加使经典数据能够嵌入高维希尔伯特空间，从而更轻松地进行模式分类，如支持向量机[36]。量子门是作用于量子比特的基本操作，也是量子电路的基本构建模块，类似于经典计算中的逻辑门。量子门分为两大类：单量子比特门和多量子比特门。单量子比特门作用于单个量子比特，例如哈达玛门（H）和旋转门（R_x（θ）、R_y（θ）、R_z（θ））。相比之下，多量子比特门可同时作用于两个或更多量子比特，包括受控非门（CNOT）、受控 Z 门（CZ）、交换门（SWAP）和 Toffoli（T）门。某些多量子比特门，例如 CNOT 和 CZ，可在量子比特之间建立关联，从而产生纠缠特性。在量子机器学习中，纠缠通过在特征之间建立关联来增强量子电路的表达能力 [30]。量子电路是一种可编程的量子门序列，作用于量子比特，在希尔伯特空间中处理数据，并利用叠加和纠缠特性。

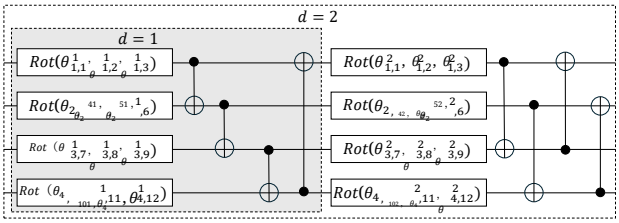
B. 量子电路架构

本节考察 PennyLane [37] 中常用于量子神经网络研究[15], [27], [30]的三种代表性量子电路模板：基本纠缠（BEQC）、强纠缠（SEQC）和随机量子电路（RQC）。每种架构都具有独特的结构和连接模式，影响量子信息在各层中的变换方式。本文并不提出新的架构，而是着重分析它们在不同电路深度（d）下的行为和鲁棒性，其中电路深度定义为重复电路层的数量。

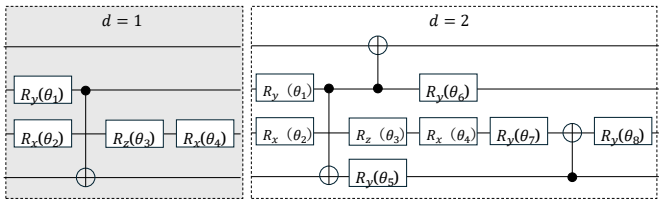
1) 基本纠缠量子电路（*BEQC*）：BEQC采用结构化设计，将参数化单量子比特旋转（R_x）与针对 n个量子比特的CNOT门闭环配置相结合，如图2a所示。



(a) BEQC。随着深度增加，重复使用相同的电路层。



(b) SEQC。旋转块（Rot），包括由R_z, R_y和 R_z构成的序列，会重复使用；而 CNOT 连接则随电路层变化。



（c）RQC。每增加一层，都会随机采样，因此在更高深度下会产生不同的电路结构。

图 2：用于 n =4 个量子比特的三种量子电路架构示例：(a) BEQC、(b) SEQC 和 (c) RQC，分别展示了电路深度为 $d = 1$ 和 $d = 2$ 的情况。

纠缠模块由酉算符定义

$$E^{(i)}_{\text{basic}} \;=\; \text{CNOT}_{n\to 1}\left(\prod_{q=1}^{n-1}\text{CNOT}_{q\to q+1}\right), \qquad (1)$$

形成环形拓扑，确保循环式最近邻纠缠。通过堆叠相同的层 d 次，可获得深度为 d 的 BEQC:

$$U^{(d)}_{\text{basic}}(\boldsymbol{\theta}) \;=\; \prod_{i=1}^{d}\left[E^{(i)}_{\text{basic}}\prod_{q=1}^{n}R_x^{(q)}\big(\theta_q^{(i)}\big)\right], \qquad (2)$$

其中 $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{(1)},\ldots,\boldsymbol{\theta}^{(d)}\}$ 表示所有层中的完整旋转参数集。

2）强纠缠量子电路（*SEQC*）：SEQC 架构通过采用密集纠缠，并在各层之间交替改变门连接方式，最大化量子比特间关联。每一层首先对 n 量子比特依次施加三个参数化单量子比特旋转（R_z、R_y、R_z），统称为 Rot:

$$U^{(i)}_{\text{Rot}}(\boldsymbol{\theta}^{(i)}) \;=\; \prod_{q=1}^{n}R_z^{(q)}(\theta^{(i)}_{q,1})\,R_y^{(q)}(\theta^{(i)}_{q,2})\,R_z^{(q)}(\theta^{(i)}_{q,3}), \qquad (3)$$

表 I：比较三种量子电路（BEQC、SEQC、RQC）的旋转门、旋转门数量（$N_{Rotation}$）和 CNOT 门数量（N_{CNOT}），其中 d 为量子电路的深度。

量子电路	旋转门	$N_{Rotation}$	N_{CNOT}
BEQC	R_x	$4d$	$4d$
SEQC	R_z, R_y	$12d$	$4d$
RQC	$R_x,\ R_y,\ R_z$	$4d$	$\sim 1.71d$

随后施加纠缠模块，在不同层之间交替改变 CNOT 的方向:

$$E^{(i)}_{\text{strongly}} \;=\; \prod_{q=1}^{n}\text{CNOT}_{\,q\to\left((q+i-1)\bmod n\right)+1}. \qquad (4)$$

由此得到的深度为d的电路表示为

$$U^{(d)}_{\text{strongly}}(\boldsymbol{\theta}) \;=\; \prod_{i=1}^{d}\Big[E^{(i)}_{\text{strongly}}\,U^{(i)}_{\text{Rot}}(\boldsymbol{\theta}^{(i)})\Big]. \qquad (5)$$

交替拓扑使所有量子比特能够跨层相互作用，增强了电路捕捉复杂多量子比特关系的能力。

3）随机量子电路（*RQC*）：RQC通过采样幺正操作并将其排列成随机参数化的层来构造。每层通常将单量子比特旋转门与随机选取的多量子比特CNOT门组合起来。在本研究中，固定单一RQC配置并使用10个随机种子，以隔离由数据损坏引起的输入随机性的影响，而非电路拓扑本身的随机性。与结构化的BEQC不同，随着深度增加，RQC不会重复使用相同的配置。例如，在 $d = 1$时，电路对应于图2c中的布局；而在 $d = 2$时，则会附加一个额外随机采样的层，而不是重复先前的结构。为确保公平比较，RQC中CNOT门与旋转门的比例固定为0.3/0.7，因此约有 $1.71d$ 个CNOT门，因为RQC中的旋转门数量为 $4d$。与使用更多双量子比特连接、因而实现更强量子比特耦合的BEQC和SEQC相比，这种配置的纠缠密度较低。由于RQC本质上具有随机性，因此不明确表示其解析酉形式；相反，比较结果以门组成的形式报告（见表I）。

IV. 方法

A. 语音处理

在音频处理中，可根据模型架构和应用领域采用多种数据表示方式。音频既可以原始波形（1D 时域信号）的形式输入网络，也可以采用二维时频表示，例如梅尔频谱图、对数梅尔频谱图或梅尔频率倒谱系数（MFCC）。在近期的大多数研究中，量子神经网络和卷积神经网络都倾向于使用二维声学特征，因为这类特征更能有效地捕捉时序和频谱依赖关系。在这些表示方式中，对数梅尔频谱图在语音识别和语音病理分类等任务中始终表现出色

[4], [15], [38], [39]。因此，本研究采用对数梅尔频谱图作为所有模型的输入表示。

对数梅尔频谱图将音频波形$x[n] \in \mathbb{R}$ 转换为二维时频表示，捕获感知上相关的频谱信息。如图 1 所示，首先使用窗函数将波形划分为相互重叠的帧，以保留局部连续性。随后，通过短时傅里叶变换（STFT）将每帧转换至频域，生成随时间变化的频谱。频谱能量经过梅尔滤波器组处理，并按照对数尺度转换为

$$S_{\text{log-Mel}}(i,m) = \log\left(\sum_k |\text{STFT}\{x[n]\}|^2 M_i[k] + \epsilon\right),$$

其中， $M_i[k]$ 表示 i^{th} 梅尔滤波器， ϵ是用于避免数值不稳定的小常数。本研究使用 Librosa 库计算对数梅尔频谱图，参数包括 512 个采样点的快速傅里叶变换大小、128 个采样点的跳跃长度、25 毫秒的窗长、40 个梅尔滤波器，以及 $\epsilon = 10^{-10}$。生成的频谱图经过归一化，并调整为固定的 40×128 （高度×宽度），作为后续章节所述量子神经网络架构的标准化二维输入。

B. 量子卷积神经网络

量子神经网络是量子-经典混合神经网络，它们使用量子变换层作为卷积层 [21]。在经典卷积神经网络中，卷积层将滤波器应用于输入数据以提取特征。类似地，量子卷积（量子卷积层）用量子电路取代这些滤波器。因此，在量子神经网络中，第一个卷积层被替换为量子卷积层，其余架构仍为经典架构。在本研究中，我们评估了三种量子神经网络配置：QNN-Basic、QNN-Strongly 和 QNN-Random。这些配置在量子电路变换和深度方面有所不同，用于评估电路复杂度对模型鲁棒性的影响。

1) 量子卷积层： 量子卷积层利用量子电路变换对数梅尔频谱图的局部块。频谱图被划分为较小的局部区域，并分别由量子卷积层进行处理。在本研究中，选择输入的一个 2×2 块，将其映射到四个量子比特。先前的研究表明，与 1×1或 3×3 [15]、 [22]等备选卷积核相比，2×2卷积核在表示能力和计算可行性之间实现了更有利的权衡。此外，该配置与当前含噪中等规模量子设备的能力相匹配，因此是适用于实际实现的选择。量子卷积层由三个主要部分组成：编码（S），即将经典数据嵌入量子态；量子电路变换（U），即应用幺正操作以捕获高维特征关系；以及解码（$\mathcal{M}$），即执行测量以获得经典特征图。经过量子神经网络（QNN）的量子卷积层和卷积神经网络（CNN）基线模型的卷积层处理后得到的对数梅尔特征表示，如补充材料中的图 9 所示。

a) 编码： 给定的经典输入块由 $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\intercal \in \mathbb{R}^n$ 表示，并使用编码函数$S : \mathbb{R}^n \to \mathcal{H}$编码到 n量子比特希尔伯特空间 $\mathcal{H}$ 中的量子态。每个分量 x_i 是块 $\mathbf{x}$ 的第 i^{th} 个元素，并被分配给一个单独的量子比特；因此，对数梅尔频谱图中的一个 2×2区域包含四个值，并使用四个量子比特（$n = 4$）表示，这些量子比特被初始化为基态 $|0\rangle$。采用角度编码时，每个归一化像素$x_i \in [0,1]$ 决定一个单量子比特门R_y的旋转角度，并通过 π 进行缩放，以确保参数范围在物理上有效。该编码方案保留了输入块的局部空间结构，并与当前含噪中等规模量子设备兼容。通过 π 进行缩放可确保旋转覆盖布洛赫半球 $[0,\pi]$，从而在量子特征空间中产生不冗余的量子比特状态。完整的编码定义如下

$$|\phi(\mathbf{x})\rangle = S(\mathbf{x}) = \bigotimes_{i=1}^{n} R_y(\pi x_i)\,|0\rangle,$$

其中， $\bigotimes$ 表示 n个单量子比特态的张量积。

b) 量子电路变换：随后，编码后的量子态由量子电路进行变换，该电路由酉算符 U 表示，并由单量子比特门和多量子比特门构成。在本研究中，我们通过评估三种量子架构——BEQC、SEQC 和 RQC（如第 III-B 节所述，分别对应 QNN-Basic、QNN-Strongly 和 QNN-Random 配置）——分析电路复杂度对用于音频分类的量子模型鲁棒性的影响。此外，我们将电路深度（d）从 1 层调整至 50 层，以研究量子变换深度增加所产生的影响。通过叠加和纠缠，该过程将编码后的数据块转换为更丰富的量子表示（$|\psi(\boldsymbol{\theta},\ \mathbf{x})\rangle$），其表达式为

$$|\psi(\boldsymbol{\theta},\mathbf{x})\rangle = U^{(d)}(\boldsymbol{\theta})|\phi(\mathbf{x})\rangle, \tag{6}$$

其中 $\boldsymbol{\theta}$ 表示一组固定的量子参数。先前研究表明，不可训练的量子卷积层能够取得与可训练层相当的性能 [21], [22], [30]。基于这些研究，我们在量子卷积层中保持量子参数固定，以尽可能降低计算成本，同时维持其有效性。未来工作可以探索可训练的变体。

c) 解码：解码（或测量， $\mathcal{M}$）从电路变换 U后的量子态中提取经典信息。在本研究中，我们对每个量子比特应用 Pauli-Z 可观测量，生成期望值，以表示多次投影测量的平均结果。我们在 PennyLane 后端中使用 shots=None 配置，该配置可提供解析期望值，而不会受到有限采样带来的统计噪声影响 [22]。四个期望值，记作 $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$，分别对应四个量子比特，其取值范围均在 $[-1,1]$内，如公式 7 所定义。这些值被分配到图3中的四个输出通道（ch.1–4），其中每个 2×2输入块都转换为多通道特征表示，作为量子卷积层的输出。

$$\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}) = \langle\mathcal{M}\rangle_{\mathbf{x},\boldsymbol{\theta}} = \langle\psi(\mathbf{x},\boldsymbol{\theta})|\sigma_z|\psi(\mathbf{x},\boldsymbol{\theta})\rangle. \tag{7}$$

2) 量子卷积网络：与经典卷积层类似，量子卷积层可以轻松集成到各种神经网络架构中。例如，该层后面可以接一个简单的经典全连接层 [22], [30], ，也可以接由卷积层、池化层和全连接层组成的串行层 [21]。本研究旨在发挥量子卷积层和经典卷积层的互补优势。因此，我们的量子神经网络架构由一个量子卷积层、一个经典卷积层、一个池化层、一个展平层和两个全连接层（FC）组成，如图3所示。量子卷积层包含一个卷积核尺寸为 2×2的滤波器。后续卷积层采用 ReLU激活函数、卷积核尺寸为 3×3，并使用 32 个滤波器。池化层使用 3×3 卷积核，将空间维度缩小至原来的三分之一。展平后，特征图被传入全连接（FC）模块，其中第一个全连接层包含 64 个采用双曲正切激活的隐藏单元，第二个全连接层作为输出层，其单元数量等于类别（标签）数量。

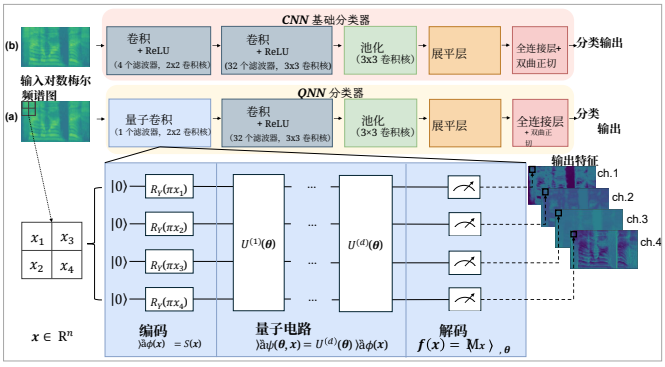


图3： (a) 量子神经网络和 (b) CNN-Base 分类器的架构。在 (a) 中，量子卷积层包含编码、量子电路和解码阶段，其中输入由旋转门（R_y）编码，经处理后解码为四个经典通道（ch.1–ch.4）。随后，这些通道依次经过卷积、池化、展平和全连接层。主要区别在于第一层：量子神经网络采用量子卷积层，而 CNN-Base 采用经典卷积层。

C. 基准*CNN*模型

本研究的主要目标是评估量子神经网络相较于经典网络的鲁棒性。主要基准模型是一个简单的CNN基线模型（CNN-Base），其架构与量子神经网络相同，但将第一个量子卷积层替换为传统卷积层。如图3所示，该卷积层采用ReLU激活函数、一个 2×2 卷积核和四个滤波器。此设计确保了公平的一对一比较，从而分离出量子特征提取对模型鲁棒性的贡献。

为进一步将结果置于背景中进行分析，我们还纳入了两种最先进的经典架构 ResNet-18 [40] 和 VGG-16 [41],，作为额外基线，用以代表语音处理中广泛使用的先进卷积神经网络模型。两种模型均采用以 ImageNet 权重初始化的预训练网络，并针对语音分类进行微调。我们修改了它们的第一个卷积层，使其能够接收单通道对数梅尔频谱图，并将最后的分类层替换为输出相应类别的新线性层。此外，VGG-16 还采用自适应平均池化层，在分类器前统一特征图的尺寸。这些调整使两种预训练架构都能处理二维声学表征，从而能够比较量子神经网络、CNN-Base 和更深层预训练模型在各种声学损坏下的效率与鲁棒性。

V. 实验

A. 数据集

在实验中，我们使用了两个音频数据集：一个用于语音病理评估，另一个用于语音情感识别。

- 高级语音功能评估数据集（AVFAD）[42]包含709名参与者的录音（346名语音障碍患者和363名健康对照组成员）。数据包括持续元音（/a/、/e/、/u/）、六个句子、一段语音平衡文本和自发语音。本研究仅使用持续元音/a/，并将每次重复视为独立样本，共获得1,498个音频信号。

- 多伦多情感语音集（TESS） [43] 包含两位女性说话人（年龄分别为 26 岁和 64 岁）录制的 2,800 段话语。每位说话人都针对七种情绪录制了 200 个目标词：愤怒、厌恶、恐惧、快乐、愉快惊讶、悲伤和中性。录音采样率为 24 kHz，广泛用于语音情感识别研究。

两个数据集中的音频均在背景噪声极少的受控环境中录制，因此被视为干净音频；本文将这些未受损的录音称为“干净音频”。随后，我们引入可控的损坏，以模拟特定噪声条件，从而在定义明确的损坏条件下系统地研究 QNN 鲁棒性，而非使用现有的含噪数据集。未来的工作将探索使用含噪数据集。接下来，所有音频信号均按照第 IV-A 节所述转换为对数梅尔频谱图，并作为分类器模型的输入。

B. 四种破坏类型

我们评估模型在四种常见声学扰动下的鲁棒性：高斯噪声、音高偏移、时间偏移和速度变化。每种扰动均设有六个严重程度级别。对于高斯噪声、音高偏移和时间偏移，$\sigma = 0$ 表示无损坏；对于速度变化，$\sigma = 1$ 则表示未修改的信号（见表 II）。为实现公平且可量化的鲁棒性评估，我们为每种扰动类型定义了受控的损坏严重程度级别，详情见表 II。每个严重程度参数的选取旨在

表 II：各扰动类型的缩放因子

扰动类型	干净值	严重程度等级
高斯噪声	$\sigma_n = 0$	σ_n = [0.01、0.05、0.1、0.15、0.2、0.25]
音高偏移	$\sigma_p = 0$	σ_p = [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]
时间偏移	$\sigma_t = 0$	σ_t = [0.025、0.05、0.075、0.1、0.125、0.15]
速度变化	$\sigma_s = 1$	σ_s = [1.05、1.1、1.15、1.2、1.25、1.3]

使输入波形逐渐产生更强但符合物理实际的退化，所选范围涵盖了自然语音中常见的变化。

高斯噪声：为模拟背景干扰，如算法1所述，将标准差为 σ_n 的零均值高斯噪声直接施加到波形上。标准差 σ_n 相对于信号幅度进行归一化，范围为 0.01 至 0.25，从而模拟约 40 dB 至 12 dB 的信噪比，这些数值代表轻度至中度环境噪声。这种损坏会独立扰动每个采样值，从而降低信号质量，但不改变其全局结构。

算法1 高斯噪声

输入：音频信号 $\mathcal{A} = \{a^{(i)}\}_{i=1}^{N}$，其中 $a^{(i)} \in \mathbb{R}^{L_i}$，以及噪声因子 σ_n
输出：含噪信号 $\tilde{\mathcal{A}} = \{\tilde{a}^{(i)}\}_{i=1}^{N}$，其中 $|\tilde{a}^{(i)}| = L_i$

对于 每个 $a^{(i)} \in \mathcal{A}$ **执行**
  计算标准差 $\hat{\sigma}_i \leftarrow \text{std}(a^{(i)})$
  采样高斯噪声：$z^{(i)} = [z_1^{(i)}, \ldots, z_{L_i}^{(i)}]$ 其中 $z_k^{(i)} \sim \mathcal{N}(0,\ \sigma_n)$ 独立地
  添加缩放噪声：$\tilde{a}^{(i)} \leftarrow a^{(i)} + \hat{\sigma}_i \cdot z^{(i)}$
  限制振幅：$\tilde{a}^{(i)} \leftarrow \text{clip}(\tilde{a}^{(i)}, -1, 1)$
结束循环
返回 $\tilde{\mathcal{A}}$

算法 2 音高偏移

输入：音频信号 $\mathcal{A} = \{a^{(i)}\}_{i=1}^{N}$ 使用 $a^{(i)} \in \mathbb{R}^{L_i}$；采样采样率 sr；半音因子 σ_p **输出：**移调后的信号 $\tilde{\mathcal{A}} = \{\tilde{a}^{(i)}\}_{i=1}^{N}$

对于每个 $a^{(i)} \in \mathcal{A}$ **执行**
  采样半音移位 $\Delta_i \sim \mathcal{N}(0,\ \sigma_p)$
  $\tilde{a}^{(i)} \leftarrow$音高移位$(a^{(i)},\ \Delta_i,$ 采样率 sr) ▷ 使用 librosa.effects.pitch_shift)
结束循环
返回 $\tilde{\mathcal{A}}$

算法 3 时间偏移

输入：音频信号 $\mathcal{A} = \{a^{(i)}\}_{i=1}^{N}$ with $a^{(i)} \in \mathbb{R}^{L_i}$；最大移位比例 σ_t
输出：移位后的信号 $\tilde{\mathcal{A}} = \{\tilde{a}^{(i)}\}_{i=1}^{N}$ 使用 $|\tilde{a}^{(i)}| = L_i$

对每个 $a^{(i)} \in \mathcal{A}$ **执行**
  采样整数偏移 $s_i \sim \mathcal{N}(0,\ \sigma_t)$
  如果 $s_i > 0$ **则** ▷ 右移：前导零，丢弃末尾
    $\tilde{a}^{(i)} \leftarrow [\underbrace{0, \ldots, 0}_{s_i}, a^{(i)}_{1:L_i - s_i}]$
  否则如果 $s_i < 0$ **则** ▷左移：尾随零，丢弃 head
    $\tilde{a}^{(i)} \leftarrow [a^{(i)}_{1-s_i:L_i}\ \underbrace{0,\ \ldots\ldots,\ 0}_{-s_i}]$
  else
    $\tilde{a}^{(i)} \leftarrow a^{(i)}$
  结束如果
结束循环
返回 $\tilde{\mathcal{A}}$

音高偏移。为了模拟说话人音高或音调变化的差异，在保持时间对齐的同时，通过σ_p 音高因子修改音频的基频（算法 2）。在频谱图上，这表现为谐波能量在各频率 bin 上的垂直平移。音高因子 $\sigma_p \in [0.05, 0.3]$ 生成的频率缩放因子范围在 ±3% 以内（0.5 个半音）。

时间偏移：为了模拟时间错位，沿时间轴将波形随机移动，其位移比例为总长度的 σ_t（算法 3）。正向 σ_t 会使信号延迟（右移），而负向 σ_t则会使信号提前（左移）。应用零填充以保持信号原始长度。在频谱图上，这对应于连续时间帧的水平平移。严重程度$\sigma_t \in [0.025, 0.15]$ 定义了最大位移，确保波形的移动幅度不超过其持续时间的 15%。

速度变化：为模拟语速变化，我们将信号按因子 $r = e^u$ 进行时间拉伸，其中 $u \sim \mathcal{N}(0,\ \log(\sigma_s))$（算法 4）。这种对数正态分布确保围绕原始播放速率对加速和减速进行对称处理。对于 $\sigma_s \in [1.05, 1.3]$，对数标准差范围约为 0.049 至 0.262，产生的速度调整通常在 ±5–30% 之间。

算法 4 速度变化

输入： 音频信号 $\mathcal{A} = \{a^{(i)}\}_{i=}$ $_1$ 其中 $a^{(i)} \in \mathbb{R}^{L_i}$，速度因子 σ_s

输出： 变速信号 $\tilde{\mathcal{A}} = \{\tilde{a}^{(i)}\}_{i=}$ $_1$ 其中 $|\tilde{a}^{(i)}| = L_i$

对每个 $a^{(i)} \in \mathcal{A}$ **执行**
  采样对数速度：$u \sim \mathcal{N}(0,\ \log(\sigma_s))$
  转换为速度比：$r \leftarrow e^u$
  应用时间拉伸：$\hat{a}^{(i)} \leftarrow$ 时间拉伸$(a^{(i)},\ r)$
  如果 $|\hat{a}^{(i)}| > L_i$ **则**
    截断：$\hat{a}^{(i)} \leftarrow \hat{a}^{(i)}[1{:}L_i]$
  否则如果 $|\hat{a}^{(i)}| < L_i$ **则**
    补零：$\hat{a}^{(i)} \leftarrow \mathrm{Pad}(\hat{a}^{(i)}, L_i)$
  结束如果
  App结束：$\tilde{a}^{(i)} \leftarrow \hat{a}^{(i)}$
结束循环
返回 $\tilde{\mathcal{A}}$

C. 训练与测试流程

每个数据集均按 65%/15%/20% 划分为训练集/验证集/测试集。量子神经网络和卷积神经网络模型均在相同的数据集上训练。为评估干净训练/损坏测试场景下的鲁棒性，评估时仅对留出的测试集施加各项损坏。我们还通过 10 次独立运行（不同随机种子）进一步验证模型性能，每次运行采用不同的训练-测试划分和随机参数初始化，例如损坏随机性和权重初始化，以支持泛化并确保评估可靠。该方案（图 1）全面评估了模型在不同损坏类型和严重程度下的性能。每个模型最多训练 10,000 训练轮次，批次大小为 20，使用 Adam（学习率为 10^{-5}，权重衰减为 10^{-2}）。如果验证损失连续 30 个训练轮次未改善，则应用早停，并使用验证集表现最佳的检查点进行测试。所有量子层均通过 PennyLane [37] 在经典设备上模拟。量子卷积实现使用 JAX 进行即时编译和向量化，以缩短编译时间 [44]。模型使用支持 CUDA 12.6 的 PyTorch 2.7.0 和 PennyLane 0.38.0 实现。训练在一台配备 AMD Ryzen 5 9600X 6 核中央处理器和 NVIDIA RTX 4060 Ti 图形处理器（8 GB GDDR6显存）的计算机上进行。

D. 评估指标

为了评估模型性能，我们首先计算模型在干净测试集上的分类准确率，以及在四种损坏类型的各个严重程度等级下的分类准确率。这样，我们便能观察干净条件与受损条件下准确率的变化，如第VI-A节所示。

为在输入退化情况下进行标准化、全面的鲁棒性评估，我们采用交叉熵（CE）、平均损坏错误率（mCE）、相对损坏错误率（RCE）和相对平均损坏错误率（RmCE）指标，遵循 [33]。这些指标可公平比较量子神经网络与卷积神经网络的鲁棒性，相关结果如第VI-A节所示。CE 指标通过将给定模型（$E_{i,\ s}$）的错误率相对于基线模型（$E^{\text{base}}_{i\ s\ ,}$）的错误率进行归一化，来量化模型在所有损坏类型 i的严重程度级别下的鲁棒性，定义见公式 8：

$$\mathrm{CE}_i = \frac{\sum_{s=1}^{6} \mathrm{E}_{i,s}}{\sum_{s=1}^{6} \mathrm{E}^{\text{base}}_{i,s}} = \frac{\sum_{s=1}^{6} \left(1 - \mathrm{Acc}_{i,s}\right)}{\sum_{s=1}^{6} \left(1 - \mathrm{Acc}^{\text{base}}_{i,s}\right)}, \tag{8}$$

其中，$\mathrm{Acc}_{i,\ s}$ 表示模型 i 在严重程度等级$s \in [1, 6]$下的准确率，而 $\mathrm{Acc}^{\text{base}}_{i\ s\ ,}$ 表示基线模型的准确率。为衡量模型在所有 $N = 4$ 种损坏类型下的鲁棒性，我们计算平均损坏错误率（mCE），如公式9所示：

$$\mathrm{mCE} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{CE}_i. \tag{9}$$

虽然 CE 和 mCE 能反映模型在损坏条件下经基线归一化的性能，但无法体现模型相对于自身干净数据准确率的性能下降程度。为评估这一方面，我们使用相对损坏错误率（RCE），衡量相较于干净基线的鲁棒性下降情况，定义如公式10所示：

$$\mathrm{RCE}_i = \frac{\sum_{s=1}^{6} \left(\mathrm{Acc}_{\text{clean}} - \mathrm{Acc}_{i,s}\right)}{\sum_{s=1}^{6} \left(\mathrm{Acc}^{\text{base}}_{\text{clean}} - \mathrm{Acc}^{\text{base}}_{i,s}\right)}, \tag{10}$$

其中，$\mathrm{Acc}_{\text{clean}}$ 表示模型在干净测试集上的模型准确率。将这一指标扩展至所有损坏类型后，相对平均腐蚀误差（RmCE）便可整体衡量相对鲁棒性，如公式11所示：

$$\mathrm{RmCE} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{RCE}_i. \tag{11}$$

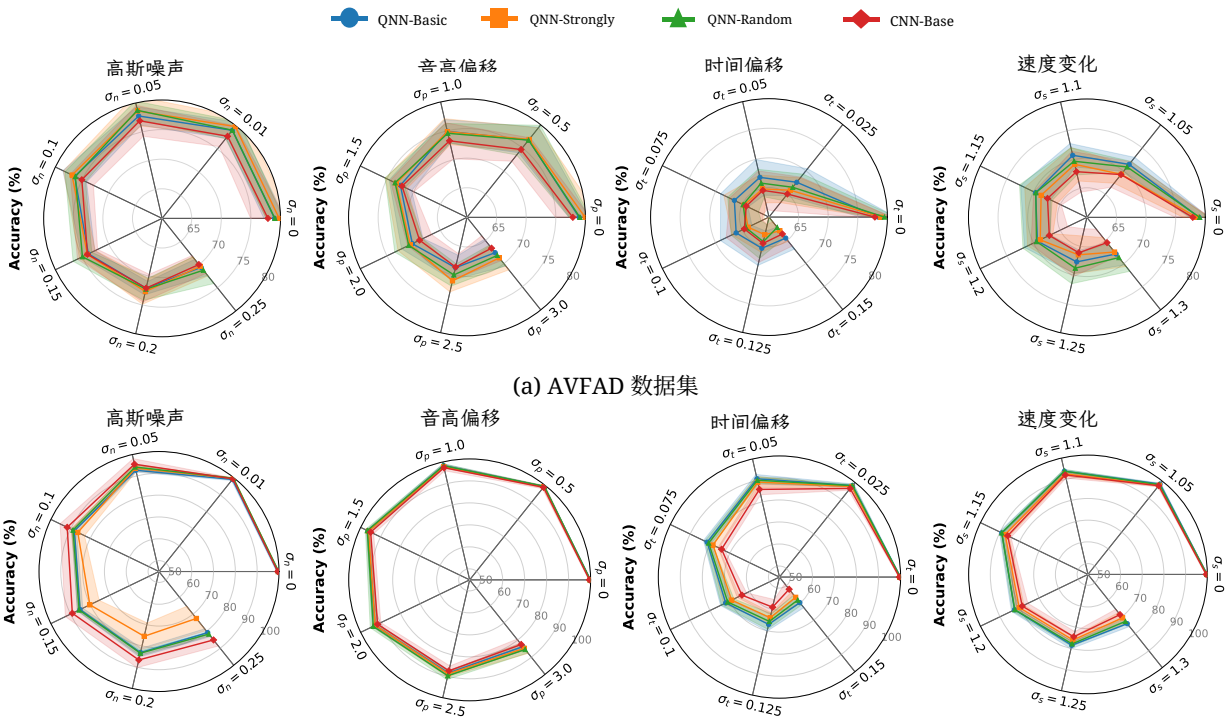
在本研究中，CNN-Base 作为基线模型。因此，其交叉熵、mCE、RCE 和 RmCE 值均归一化为 1。若 QNN 的这些值低于 1，则表示其鲁棒性优于 CNN-Base；若高于 1，则表示其鲁棒性较低。

此外，我们使用混淆矩阵分析不同情绪下的分类结果（第 VI-C 节），以识别哪些情绪在遭受损坏时退化最严重，哪些情绪仍能得到最可靠的分类。此外，我们还考察训练动态，包括每个训练轮次的训练损失和验证准确率，以评估模型收敛速度（第 VI-E 节）。消融研究还考察了量子电路深度对 QNN 性能的影响（第 VI-D 节）。其他章节的结果则展示并讨论了本研究确定的性能最佳配置（见表V）。

VI. 结果与讨论

A. 鲁棒性分析

图4展示了三种量子神经网络和 CNN-Base 在四种损坏类型下的准确率，而表III列出了两个数据集上的量子神经网络和 CNN-Base 的鲁棒性指标，例如 CE、mCE、RCE 和 RmCE。正如预期，随着腐蚀严重程度增加，所有模型的准确率都逐渐下降。



(a) AVFAD 数据集

(b) TESS 数据集

图4：QNN（QNN-Basic、QNN-Strongly 和 QNN-Random）与 CNN-Base 在 AVFAD、TESS 数据集上面对四种损坏类型（高斯噪声、音高偏移、时间偏移和速度变化）时的准确率。QNN-Basic、QNN-Strongly 和 QNN-Random 对应的电路深度（d）汇总于表V。表III：CNN-Base 与量子神经网络模型在 AVFAD 和 TESS 数据集上的鲁棒性指标对比，数值为 10 个随机种子下的均值 ± 标准差。QNN-Basic、QNN-Strongly 和 QNN-Random 对应的电路深度（d）汇总于表V。

AVFAD 数据集

扰动类型	CE				RCE			
	CNN-Base	QNN-Basic	QNN-Random	QNN-Strongly	CNN-Base	QNN-Basic	QNN-Random	QNN-Strongly
高斯噪声	1.00 ± 0.00	0.98 ± 0.08	**0.96 ± 0.07**	0.96 ± 0.10	1.00 ± 0.00	1.69 ± 1.68	1.86 ± 2.64	**1.68 ± 1.44**
音高偏移	1.00 ± 0.00	0.96 ± 0.06	0.95 ± 0.07	**0.94 ± 0.06**	1.00 ± 0.00	1.08 ± 0.46	**0.99 ± 0.27**	1.08 ± 0.30
时间偏移	1.00 ± 0.00	**0.95 ± 0.04**	1.00 ± 0.08	1.01 ± 0.08	1.00 ± 0.00	**0.93 ± 0.16**	1.19 ± 0.41	1.19 ± 0.44
速度变化	1.00 ± 0.00	**0.93 ± 0.04**	0.93 ± 0.08	0.97 ± 0.05	1.00 ± 0.00	0.95 ± 0.39	**0.95 ± 0.30**	1.02 ± 0.36
mCE	1.00 ± 0.00	**0.96 ± 0.03**	0.96 ± 0.04	0.97 ± 0.04	-	-	-	-
RmCE	-	-	-	-	1.00 ± 0.00	**1.16 ± 0.62**	1.25 ± 0.79	1.24 ± 0.53

TESS 数据集

扰动类型	CE				RCE			
	CNN-Base	QNN-Basic	QNN-Random	QNN-Strongly	CNN-Base	QNN-Basic	QNN-Random	QNN-Strongly
高斯噪声	1.00 ± 0.00	1.52 ± 0.55	**1.40 ± 0.42**	1.99 ± 0.93	1.00 ± 0.00	1.59 ± 0.67	**1.51 ± 0.53**	2.16 ± 1.14
音高偏移	1.00 ± 0.00	0.89 ± 0.16	**0.79 ± 0.13**	0.88 ± 0.13	1.00 ± 0.00	0.94 ± 0.14	**0.83 ± 0.13**	0.90 ± 0.13
时间偏移	1.00 ± 0.00	**0.78 ± 0.05**	0.81 ± 0.02	0.86 ± 0.10	1.00 ± 0.00	**0.79 ± 0.05**	0.82 ± 0.02	0.87 ± 0.10
速度变化	1.00 ± 0.00	**0.82 ± 0.07**	0.84 ± 0.07	0.94 ± 0.13	1.00 ± 0.00	**0.84 ± 0.07**	0.86 ± 0.06	0.96 ± 0.12
mCE	1.00 ± 0.00	1.00 ± 0.11	**0.96 ± 0.08**	1.17 ± 0.19	-	-	-	-
RmCE	-	-	-	-	1.00 ± 0.00	1.04 ± 0.14	**1.01 ± 0.11**	1.22 ± 0.24

1）高斯噪声：在 AVFAD 数据上，对于大多数噪声等级，量子神经网络的准确率略高于 CNN-Base，且三种量子神经网络模型的表现相近。它们较低的 CE 值进一步证实，与 CNN-Base 相比，其抗腐蚀鲁棒性有所提升。不过，在最严重的几个等级下，CNN-Base 的准确率与量子神经网络相同。由于 CNN-Base 的干净数据准确率较低，其相对性能下降不那么明显，因此 RCE 值低于量子神经网络。因此，量子神经网络展现出更优的抗损鲁棒性较高，但在 AVFAD 上的相对鲁棒性较差。

相比之下，在 TESS 数据集上，面对高斯噪声时，CNN-Base 的表现明显优于所有量子神经网络。值得注意的是，QNN-Strongly 的准确率从 100% 大幅降至 70%，对应的交叉熵为 = 1.99，RCE 为 = 2.16；而 CNN-Base 在所有噪声水平下的准确率均超过 85%。这些结果表明，在情感语音分类中，量子神经网络对高斯噪声的抵御能力较弱。这种脆弱性可以解释为，噪声可能会放大量子卷积层量子编码中的随机波动

此外，两个数据集上的性能差异源于 AVFAD 语音录音比 TESS 的更加不规则且噪声更多，因此在这种具有挑战性的数据上，量子神经网络相较于 CNN-Base 仍具有一定竞争力。相比之下，TESS 语音更加干净、结构也更规整。因此，叠加的高斯噪声会更严重地扰乱量子卷积层，使 CNN-Base 获得更大的性能优势。

2）音高偏移：随着音高偏移严重程度的增加，在两个数据集上，量子神经网络的准确率和鲁棒性始终优于 CNN-Base，其 CE 值更小。与 AVFAD 数据集相比，这一优势在 TESS 数据集上更加明显。例如，在 TESS 数据集上，QNN- Random 的 CE 为 = 0.79，RCE 为 = 0.83；而在 AVFAD 数据集上，其 CE 为 = 0.95，RCE 为 = 0.99。这些发现表明，量子神经网络能够更有效地捕捉相关性和结构性频率模式，从而增强其抵御音高偏移等频谱域损坏的能力。

3）时间偏移：在四种损坏类型中，时间偏移对模型准确率的干扰最为严重，导致准确率下降 20%–50%。在 AVFAD 数据集上，QNN-Basic 获得了最高准确率，尤其是在中等裁剪级别（$\sigma_t = 0.075$–0.125）下，其 CE（0.95）和 RCE（0.93）值均小于 CNN-Base。相比之下，QNN-Random 和 QNN-Strongly 在 CE 方面与 CNN-Base 表现相近，但相对鲁棒性更差，RCE 为 = 1.19。

在 TESS 数据集上，随着裁剪严重程度增加，量子神经网络始终优于 CNN-Base，其更小的交叉熵值（0.78-0.86）和 RCE 值（0.79-0.87）证明了这一点。在最严重的级别下，量子神经网络的准确率比 CNN-Base 高约 8%。QNN-Basic 再次取得了最强的性能，其交叉熵值（0.78）和 RCE 值（0.79）最低，相比 CNN-Base 实现了 22% 的鲁棒性提升。这些发现表明，量子神经网络在保留全局时间特征方面更为有效，而当时间连续性受到破坏时，CNN-Base 的性能会急剧下降。

4) 速度变化：与 CNN-Base 相比，速度变化在两个数据集上都进一步凸显了量子神经网络的鲁棒性。在 AVFAD 数据集上，QNN-Basic 和 QNN-Random 的准确率最高，其交叉熵值和 RCE 值均小于 CNN-Base，表明它们具有更强的鲁棒性。在 TESS 数据集上，QNN-Basic 表现最为突出，在所有模型中始终取得最低的交叉熵值和 RCE 值。这些结果表明，量子映射对时间特征的重新缩放不太敏感。

总体而言，在音高偏移、时间偏移和速度变化下，量子神经网络的鲁棒性始终优于 CNN-Base，而 CNN-Base 在高斯噪声下则更具韧性。鲁棒性指标还表明，不同量子模型对不同损坏类型的响应各不相同。使用 mCE 和 RmCE 对四种损坏进行评估后发现，QNN-Basic 在 AVFAD 数据集上总体鲁棒性最佳，而 QNN-Random 在 TESS 数据集上表现最佳。相比之下，QNN-Strongly 具有更复杂的纠缠结构，其韧性有所下降

表 IV：经典架构（CNN-Base、ResNet-18、VGG-16）与量子神经网络在 AVFAD 和 TESS 数据集上的模型鲁棒性（mCE 和 RmCE）比较。QNN-Basic、QNN-Strongly 和 QNN-Random 对应的电路深度（d）汇总于表 V。

模型名称	模型大小	AVFAD 数据集		TESS 数据集	
		mCE	RmCE	mCE	RmCE
CNN-Base	1M	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
ResNet-18	11M	1.06 ± 0.04	1.82 ± 1.59	1.17 ± 0.25	1.25 ± 0.28
VGG-16	134M	0.94 ± 0.09	1.95 ± 1.29	1.23 ± 0.39	1.28 ± 0.35
QNN-Basic	1M + 4q（冻结）	0.96 ± 0.03	1.16 ± 0.62	1.00 ± 0.11	1.04 ± 0.14
QNN-Random	1M + 4q（冻结）	0.96 ± 0.04	1.25 ± 0.79	0.96 ± 0.08	1.01 ± 0.11
QNN-Strongly	1M + 4q（冻结）	0.97 ± 0.04	1.24 ± 0.53	1.17 ± 0.19	1.22 ± 0.24

与其他量子模型相比。量子神经网络的 RmCE 值高于 CNN-Base（RmCE $>$ 1），这主要是因为它们在高斯噪声下的相对鲁棒性较差，而高斯噪声对平均值的影响占主导。

B. 与大规模经典网络的鲁棒性比较

表 IV 使用 mCE 和 RmCE 指标，对比了量子神经网络与先进经典架构 ResNet-18 和 VGG-16 在 AVFAD 和 TESS 数据集上的鲁棒性。尽管量子神经网络模型的参数量约为 1M（+4 个量子比特），而 ResNet-18 和 VGG-16 的参数量分别为 11M 和 134M，但所有量子神经网络模型在各种输入扰动下都实现了相当或更优的鲁棒性。在 AVFAD 数据集上，量子神经网络模型的 mCE 保持在 = 0.96，与大型经典网络接近或更低，同时避免了 VGG-16 中观察到的较高 RmCE 方差（1.95 ± 1.29）。在 TESS 数据集上，量子编码的优势更加明显：较深的经典网络性能显著下降（ResNet-18：mCE= 1.17 ± 0.25，VGG-16：mCE = 1.23 ± 0.39），而随机量子神经网络实现了最低的 mCE（0.96 ± 0.08）和 RmCE（1.01 ± 0.11）。这些结果表明，量子特征映射能够产生更具不变性且紧凑的表征，使量子神经网络在保持更高参数效率的同时，达到或超过规模大得多的经典架构的鲁棒性。

C. 多类别准确率

图5中的混淆矩阵展示了TESS 数据集上的分类准确率，分别在干净测试数据和每种损坏最严重的级别下进行评估。为清晰起见，我们重点关注CNN-Base和QNN-Random，后者作为量子神经网络模型的代表；QNN-Basic和QNN-Strongly对应的混淆矩阵见补充材料。在七种情绪中，QNN和CNN-Base在干净数据上都实现了近乎完美的准确率，其中QNN表现出略强的泛化能力。然而，在损坏条件下，性能显著下降。在高斯噪声下，中性情绪受到的影响最为严重：CNN-Base的准确率从99.9%降至28.5%，而QNN从99.9%降至5.5%，因为两个模型都经常将中性 误分类为悲伤。这是因为高斯噪声与中性语音平坦的频谱模式高度重叠，使模型难以

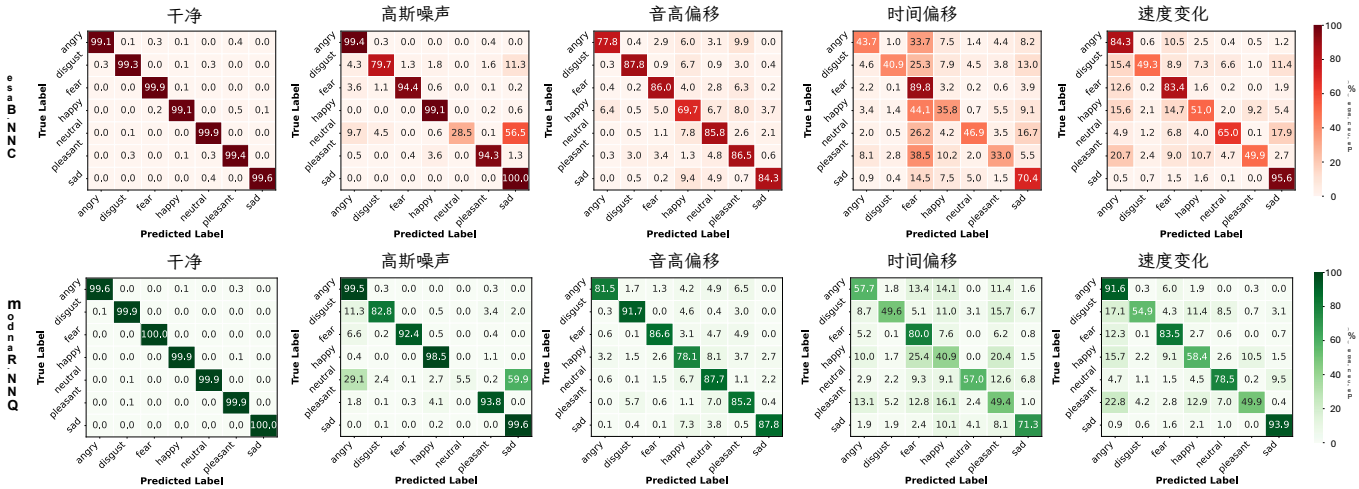


图5：CNN-Base（红色）和 QNN-Random（绿色）在干净条件下以及四种损坏类型最严重等级下的混淆矩阵。QNN-Random 对应的电路深度（d）为 $d = 1$。

区分中性 和悲伤，后者具有相似的低能量频率分布。相比之下，两个模型都能可靠地分类悲伤、快乐和愤怒 。对于音高偏移，快乐 情绪受影响最大，而厌恶仍是最具鲁棒性的类别。这是因为快乐语音高度依赖音高升高这一判别特征，因此对频谱偏移非常敏感 [45]。然而，QNN对快乐 的分类准确率仍然相对较高（78.1%），高于CNN-Base的69.7%。在时间偏移下，大多数情绪都出现了明显的性能下降，只有恐惧例外，其在CNN-Base和QNN中的准确率分别达到89.8%和80%。恐惧 的鲁棒性源于其与持续的高能量片段相关，因此即使只有部分信号，也能保留足够的分类线索。较高的准确率还受到预测偏差的影响：两个模型都倾向于将裁剪后的样本误分类为恐惧，且CNN-Base中的这种偏差强于QNN。对于速度变化，受影响最大的情绪是厌恶、快乐和愉悦，而愤怒、恐惧和悲伤 仍能以相对较高的准确率完成分类。这是因为速度损坏会不成比例地扭曲依赖时间的线索，例如节奏变化和韵律变化，而这些线索对于快乐 和厌恶 [46]等情绪至关重要。相比之下，愤怒 和恐惧 依赖更多全局性的能量和频谱模式，对时间尺度失真不太敏感。

总体而言，在各种损坏下，恐惧 是分类鲁棒性最高的情绪；而在高斯噪声下，中性 最脆弱，在音高偏移、时间偏移和速度变化下，快乐最脆弱。在各模型中，QNN-Random 在大多数损坏下始终优于 CNN-Base，明显例外是高斯噪声下的中性 类别，此时 CNN-Base 的鲁棒性更高。

D. 消融研究

为评估量子电路深度（d）对量子神经网络鲁棒性的影响，我们通过将 d 从 1 变化到 50 进行了消融研究。表V总结了最优电路

表V：三种量子电路架构的最优电路深度：QNN-Basic 中的 BEQC（d_{BEQC}）、QNN-Strongly 中的 SEQC（d_{SEQC}）以及 QNN-Random 中的 RQC（d_{RQC}）。

AVFAD 数据集			
扰动类型	QNN-Basic d_{BEQC}	QNN-Strongly d_{SEQC}	QNN-Random d_{RQC}
高斯噪声	30	20	25
音高偏移	4	15	15
时间偏移	15	10	1
速度变化	4	1	25
TESS 数据集			
扰动类型	QNN-Basic d_{BEQC}	QNN-Strongly d_{SEQC}	QNN-Random d_{RQC}
高斯噪声	50	10	1
音高偏移	30	30	1
时间偏移	15	50	1
速度变化	15	4	1

三种量子电路架构的深度：QNN-Basic 中的 BEQC（d_{BEQC}）、QNN-Strongly 中的 SEQC（d_{SEQC}）以及 QNN-Random 中的 RQC（d_{RQC}）在 AVFAD 和 TESS 数据集的四种损坏类型下的表现。如图所示，在所有设置下，最佳深度并不存在一致的趋势。无论数据集或损坏类型如何，QNN-Random 在较浅的深度 $d_{\mathrm{RQC}} = 1$ 下始终表现最佳，表明即使电路复杂度最低，其仍具有很强的鲁棒性。相比之下，QNN-Basic 和 QNN-Strongly 的最佳深度会根据损坏类型和数据集而变化。例如，在高斯噪声下，QNN-Basic 更倾向于使用较深的电路（$d_{\mathrm{BEQC}} = 30$–50），而 QNN-Strongly 在中等深度（$d_{\mathrm{SEQC}} = 10$–20）下达到最佳性能。图 6 进一步展示了 QNN-Random 模型在 TESS 数据集四种输入扰动下的准确率，而 QNN-Basic 和 QNN-Strongly 在 TESS、AVFAD 数据集上的相应结果见补充材料。结果表明，电路深度会影响模型鲁棒性，尤其是在高斯噪声下。对于四种损坏类型，QNN-Random 模型在较浅深度下取得了最高的准确率和鲁棒性

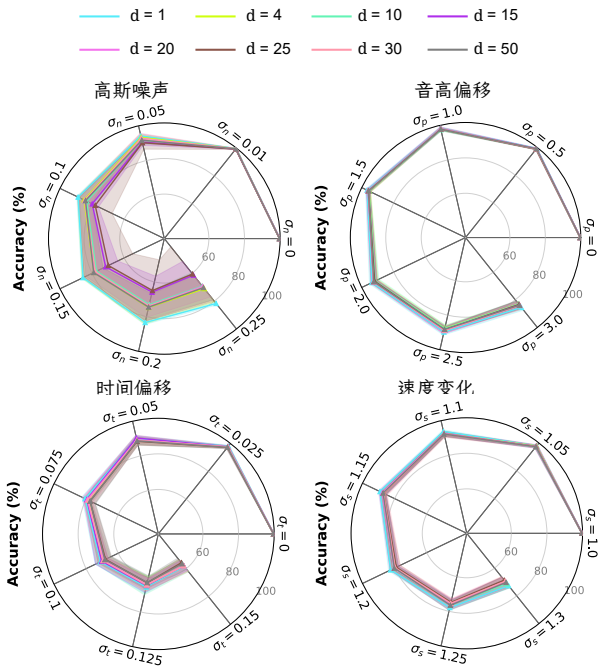


图 6：TESS 数据集上，QNN-Random 模型在不同电路深度（$d \in [1, 4, 10, 15, 20, 25, 30, 50]$）和不同损坏类型下的准确率。模型在较浅的深度 $d = 1$ 时取得最高准确率和鲁棒性，随着深度增加，性能呈非线性下降。

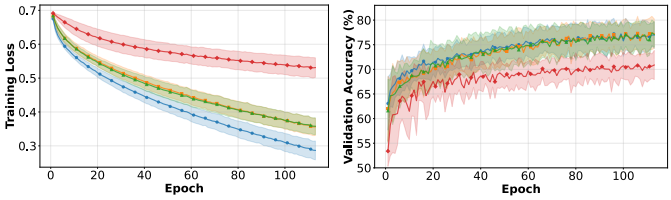
为 $d = 1$。随着深度增加，性能以非线性方式下降，这表明更深的电路并不一定能提升泛化能力。QNN-Basic 和 QNN-Strongly 也表现出类似的非线性下降趋势；然而，它们的最佳深度会因不同的损坏类型而异，具体详见补充材料。

总体而言，改变 QNN-Random 和 QNN-Strongly 的电路深度，对其抵抗音高偏移、时间偏移和速度变化的鲁棒性影响较小，而 QNN-Basic 对深度变化更为敏感。相比之下，三个 QNN 变体在受到高斯噪声影响时均表现出明显的深度依赖性，凸显了电路复杂度对于维持噪声鲁棒性的关键作用。

E. 收敛性分析

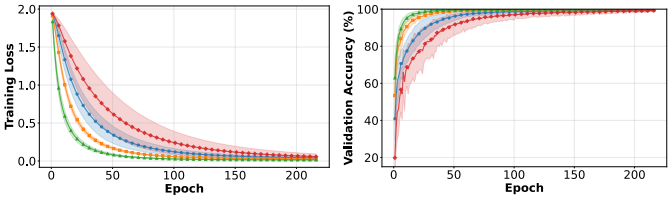
图7和图8展示了 AVFAD 数据集和 TESS 数据集上每个训练轮次的训练损失和验证准确率，所报告的数值为基于 10 个随机种子计算的均值和标准差。训练过程中采用了早停，并展示了最少训练轮次数，以突出收敛行为。正如预期，随着训练轮次增加，训练损失降低，验证准确率提高。在 AVFAD 数据集上，与量子神经网络相比，CNN-Base 始终表现出较低的验证准确率和较高的损失。在两个数据集上，量子神经网络均比 CNN-Base 更快收敛，这种差异在 TESS 数据集上尤为明显。例如，QNN-Random 在约 30 个训练轮次内收敛，而 CNN-Base 则需要近 200 个训练轮次。在 AVFAD 数据集上，QNN-Basic 达到最快收敛，而在 TESS 数据集上，QNN-Random 收敛最快，这与第 VI-A 节报告的鲁棒性结果一致。此外，CNN-Base 在 TESS 数据集上的标准差更大，表明与量子神经网络相比，其不同随机种子之间的变化更大。

总而言之，在两个数据集上，量子神经网络的收敛速度都显著快于 CNN-Base。这一特性在现实场景中具有优势，因为更快的训练能够降低计算成本并提高部署效率。



(a) 训练损失。 (b) 验证准确率。

图 7：AVFAD 数据集上的训练损失和验证准确率。QNN-Basic、QNN-Strongly 和 QNN-Random 的相应电路深度（d）汇总于表V。



(a) 训练损失。 (b) 验证准确率。

图 8：TESS 数据集上的训练损失和验证准确率。QNN-Basic、QNN-Strongly 和 QNN-Random 的相应电路深度（d）汇总于表V。

VII. 结论

本研究使用语音病理（AVFAD）和语音情感（TESS）数据集，系统评估了多个量子神经网络模型在多种声学损坏条件下相较于经典卷积神经网络的鲁棒性。在各种损坏类型下，量子神经网络在音高偏移、时间偏移和速度变化时始终优于 CNN-Base（严重时间偏移下的 CE/RCE 优势最高达 22%）；而 CNN-Base 在高斯噪声下表现更佳，凸显了量子模型与经典模型对噪声的敏感性互为补充。数据集特性也会影响鲁棒性：量子神经网络在 AVFAD 不规则且带噪的语音上保持了有竞争力的表现；而在更干净的 TESS 数据集上，高斯噪声对量子神经网络造成了不成比例的性能下降，这可能放大了量子卷积层中的随机波动。此外，量子神经网络的鲁棒性也达到并超过了两种先进经典架构（ResNet-18 和 VGG-16），同时参数量更少。按情感类别分析发现，恐惧仍是最鲁棒的类别（在严重时间性损坏下准确率约为 80–90 %），而中性 和快乐 分别最易受到高斯噪声和时间偏移的影响。在三种量子电路中，QNN-Basic 的整体鲁棒性最高，尤其是在 AVFAD 上，其 mCE 和 RmCE 值最低。

量子神经网络在 AVFAD 上的整体鲁棒性尤为突出，取得了最低的 mCE 和 RmCE 值。QNN-Random 在 TESS 上展现出更强的泛化能力，在除高斯噪声外的大多数损坏类型下均优于 CNN-Base；而 QNN-Strongly 的抗扰能力较弱，这表明过高的纠缠深度可能会降低鲁棒性。此外，量子电路的深度会影响量子神经网络的表现，尤其是在高斯噪声下。最后，训练动态分析显示，量子神经网络具有显著的收敛优势：约 30 个训练轮次内即可趋于稳定，而 CNN-Base 则需要近 200 个训练轮次。总体而言，这些发现表明，量子神经网络不仅收敛更快，还对结构性和时间性损坏具有内在鲁棒性，进一步凸显了其作为基于语音的量子机器学习中可靠架构的潜力。本研究采用软件模拟的量子层、两个语音数据集和无混响损坏条件；未对室内声学、信道效应、重叠说话人或量子硬件噪声进行建模。

在未来的工作中，我们将评估 QNN 在更真实声学条件下的鲁棒性，例如混响、通道失真和语音重叠，并将量子硬件噪声明确纳入训练和推理。此外，我们计划研究可训练量子卷积层的性能优势，并通过纳入更多数据集开展跨语料库评估，以评估模型在不同语音语料库上的泛化能力。

致谢

本研究得到迪肯大学工程学院网络感知与生物医学工程（NSBE）研究实验室的支持。

参考文献

[1] D. Yu 和 L. Deng，《自动语音识别》。施普林格，2016年，第1卷。[2] H. Ahlawat、N. Aggarwal 和 D. Gupta，“自动语音识别：深度学习技术与方法综述”，《国际认知计算工程期刊》，2025年。[3] S. Wang、Y. Yang、Z. Wu、Y. Qian 和 K. Yu，“利用深度生成模型进行基于嵌入的说话人识别的数据增强”，*IEEE/ACM* 音频、语音与语言处理汇刊，第28卷，第2598–2609页，2020年。[4] H. Meng、T. Yan、F. Yuan 和 H. Wei，“利用深度学习网络从三维对数梅尔频谱图进行语音情感识别”，*IEEE Access*，第7卷，第125 868–125 881页，2019年。[5] P. Harar、J. B. Alonso-Hernandezy、J. Mekyska、Z. Galaz、R. Burget 和 Z. Smekal，“利用深度学习进行嗓音病理检测：一项初步研究”，载于*2017*年国际仿生智能会议与研讨会（*IWOBI*）。IEEE，2017年，第1–4页。[6] M. A. Mohammed、K. H. Abdulkareem、S. A. Mostafa、M. Khanapi Abd Ghani、M. S. Maashi、B. Garcia-Zapirain、I. Oleagordia、H. Al-hakami 和 F. T. Al-Dhief，“基于卷积神经网络模型的嗓音病理检测与分类”，《应用科学》，第10卷，第11期，第3723页，2020年。[7] S. Boll，“利用谱减法抑制语音中的声学噪声”，*IEEE*声学、语音与信号处理汇刊，第27卷，第2期，第113–120页，2003年。[8] S. Kamath、P. Loizou 等，“一种增强受有色噪声污染语音的多频带谱减法方法。”载于*ICASSP*，第4卷，2002年，44 164 44 164页。[9] B. Koo、J. D. Gibson 和 S. D. Gray，“用于语音增强与编码的有色噪声滤波”，载于国际声学、语音与信号处理会议。IEEE，1989年，第349–352页。

[10] Y. Qian、M. Bi、T. Tan 和 K. Yu，“用于噪声鲁棒语音识别的超深卷积神经网络”，IEEE/ACM 音频、语音与语言处理汇刊，第 24 卷，第 12 期，第 2263–2276 页，2016年。[11] T. Tan、Y. Qian、H. Hu、Y. Zhou、W. Ding 和 K. Yu，“用于噪声鲁棒语音识别的自适应超深卷积残差网络”，IEEE/ACM 音频、语音与语言处理汇刊，第 26 卷，第 8 期，第 1393–1405 页，2018年。[12] J. Ao、M. S. Yıldırım、R. Tao、M. Ge、S. Wang、Y. Qian 和 H. Li，“Used：通用说话人提取与说话人日志”，IEEE/ACM 音频、语音与语言处理汇刊，2024年。[13] T. Zhou、Y. Zhao 和 J. Wu，“用于说话人验证的 ResNeXt 和 Res2Net 结构”，载于 2021 年 IEEE 口语技术研讨会（SLT）。IEEE，2021年，第 301–307 页。[14] S. Hamsa、I. Shahin、Y. Iraqi、E. Damiani、A. B. Nassif 和 N. Werghi，“利用学习式语音分离和语音 VGG，从情感语音与噪声语音中识别说话人”，《专家系统与应用》，第 224 卷，文章号 119871，2023年。[15] C.-H. H. Yang、J. Qi、S. Y.-C. Chen、P.-Y. Chen、S. M. Siniscalchi、X. Ma 和 C.-H. Lee，“利用量子卷积神经网络实现自动语音识别中的特征提取去中心化”，载于 ICASSP 2021—2021 IEEE 声学、语音与信号处理国际会议（ICASSP）。IEEE，2021年，第 6523–6527 页。[16] S. Sridevi 等，“用于提升语音分类效果的量子核辅助支持向量机分类器”，载于 2023年第 14 届国际计算、通信与网络技术会议（ICCCNT）。IEEE，2023年，第 1–6 页。[17] G. Balachandran、S. Ranjith、G. Jagan 和 T. Chenthil，“采用优化的等变量子卷积神经网络实现先进的语音情感识别，以准确分类情绪状态”，《知识型系统》，第 316 卷，文章号 113414，2025年。[18] R. Soltani、E. Benmohamed 和 H. Ltifi，“用于快速、准确语音情感识别的量子增强型皮层深度回声状态网络”，《量子机器智能》，第 7 卷，第 2 期，文章号 77，2025年。[19] K.-C. Chen、W. Ma 和 X. Xu，“用于语音识别的基于共识的分布式量子核学习”，载于 2025 IEEE 声学、语音与信号处理国际会议研讨会（ICASSPW）。IEEE，2025年，第 1–5 页。[20] M. Esposito、G. Uehara 和 A. Spanias，“用于音频分类及医疗保健应用的量子机器学习”，载于 2022年第 13 届国际信息、智能、系统与应用会议（IISA）。IEEE，2022年，第 1–4 页。[21] M. Henderson、S. Shakya、S. Pradhan 和 T. Cook，“量子卷积神经网络：利用量子电路驱动图像识别”，《量子机器智能》，第 2 卷，第 1 期，文章号 2，2020年。[22] A. Ceschini、A. Carbone、A. Sebastianelli、M. Panella 和 B. Le Saux，“关于混合量子卷积神经网络优化”，《量子机器智能》，第 7 卷，第 1 期，文章号 18，2025年。[23] R. Ahalya、F. M. Almutairi、U. Snekhalatha、V. Dhanraj 和 S. M. Aslam，“Ranet：用于手部热成像中类风湿关节炎自动检测的定制卷积神经网络模型与量子卷积神经网络”，《科学报告》，第 13 卷，第 1 期，文章号 15638，2023年。[24] J. Preskill，“NISQ 时代及之后的量子计算”，《量子》，第 2 卷，文章号 79，2018年。[25] J. Guan、W. Fang 和 M. Ying，“量子分类器的鲁棒性验证”，载于计算机辅助验证国际会议。施普林格，2021年，第 151–174 页。[26] Y. Li、Y. Qu、R.-G. Zhou 和 J. Zhang，“QMLSC：用于情感分类的量子多模态学习模型”，《信息融合》，第 120 卷，文章号 103049，2025年。[27] T. Ahmed、M. Kashif、A. Marchisio 和 M. Shafique，“量子神经网络中的噪声鲁棒性比较分析与评估”，《科学报告》，第 15 卷，第 1 期，文章号 33654，2025年。[28] S. Lu、L.-M. Duan 和 D.-L. Deng，“量子对抗机器学习”，《物理评论研究》，第 2 卷，第 3 期，文章号 033212，2020年。[29] M. T. West、S. M. Erfani、C. Leckie、M. Sevior、L. C. Hollenberg 和 M. Usman，“大规模基准测试对抗鲁棒量子机器学习”，《物理评论研究》，第 5 卷，023186，2023年 2，第 p.。[30] W. El Maouaki、A. Marchisio、T. Said、M. Bennai 和 M. Shafique，“Advqunn：分析量子卷积神经网络对抗鲁棒性的方法”，载于 2024 IEEE 量子软件国际会议（QSW）。IEEE，2024年，第 175–181 页。[31] C. Wang、J. Du、H. Chen、R. Wang、C.-H. H. Yang、J. Zhao、Y. Ren、Q. Li 和 C.-H. Lee，“利用量子计算增强词级音视频语音识别中的隐私保护”，载于 2023年亚洲

太平洋信号与信息处理协会年度峰会与会议（*APSIPA ASC*）。IEEE，2023年，第 635–642 页。[32] M. K. Quan、M. Wijayasundara、S. Setunge 和 P. N. Pathirana，“用于物联网环境中鲁棒声学场景分类的量子增强型 Transformer 模型”，载于 *2025*年国际计算、网络与通信会议（*ICNC*）。IEEE，2025年，第 295–299 页。[33] D. Hendrycks 和 T. Dietterich，“对神经网络应对常见损坏与扰动的鲁棒性进行基准测试”，*arXiv* 预印本 *arXiv:1903.12261*，2019年。[34] M. A. Nielsen 和 I. L. Chuang，量子计算与量子信息。剑桥大学出版社，2010年。[35] L. K. Grover，“用于数据库搜索的快速量子力学算法”，载于第二十八届 *ACM* 计算理论年度研讨会论文集，1996年，第 212–219 页。[36] M. Schuld 和 N. Killoran，“特征希尔伯特空间中的量子机器学习”，《物理评论快报》，第 122 卷，第 4 期，文章号 040504，2019年。[37] V. Bergholm、J. Izaac、M. Schuld、C. Gogolin、S. Ahmed、V. Ajith、M. S. Alam、G. Alonso-Linaje、B. AkashNarayanan、A. Asadi 等，“PennyLane：混合量子-经典计算的自动微分”，*arXiv* 预印本 *arXiv:1811.04968*，2018年。[38] A. Meghanani、C. S. Anoop 和 A. Ramakrishnan，“探索对数梅尔频谱图和梅尔频率倒谱系数特征在自发语音阿尔茨海默病性痴呆识别中的应用”，载于*2021* 年 *IEEE* 口语技术研讨会（*SLT*）。IEEE，2021年，第 670–677 页。[39] K. Bhangale 和 K. Mohanaprasad，“利用梅尔频率对数频谱图和深度卷积神经网络进行语音情感识别”，载于国际未来通信与网络技术会议。施普林格，2020年，第 241–250 页。[40] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于*IEEE* 计算机视觉与模式识别会议论文集，2016年，第 770–778 页。[41] K. Simonyan 和 A. Zisserman，“用于大规模图像识别的超深卷积网络”，*arXiv* 预印本 *arXiv:1409.1556*，2014年。[42] L. M. Jesus、I. Belo、J. Machado 和 A. Hall，“高级语音功能评估数据库（AVFAD）：为嗓音临床医师和语音研究提供的工具”，载于《言语语言病理学进展》。IntechOpen，2017年。[43] M. K. Pichora-Fuller 和 K. Dupuis，“多伦多情感语音集（TESS）”，*Scholars Portal Dataverse*，第 1 卷，文章号 2020，2020年。[44] J. Bradbury、R. Frostig、P. Hawkins、M. J. Johnson、C. Leary、D. Maclaurin、G. Necula、A. Paszke、J. VanderPlas、S. Wanderman-Milne 和 Q. Zhang，“JAX：Python+NumPy 程序的可组合变换”，2018年。 [在线]。网址：http://github.com/jax-ml/jax [45] J. Yuan、L. Shen 和 F. Chen，“中文中愤怒、恐惧、喜悦和悲伤的声学实现”，载于*INTERSPEECH*，2002年，第 2025–2028 页。[46] S. Rigoulot、E. Wassiliwizky 和 M. D. Pell，“倒序感受？语音中的时间顺序如何影响声音情感识别的时间进程”，《心理学前沿》，第 4 卷，文章号 367，2013年。