基于非侵入式 PESQ-DNN 的编码语音质量测量

Ziyi Xu、Ziyue Zhao 和 Tim Fingscheidt，*IEEE* 高级会员

arXiv:2304.09226v1 [eess.AS] 18 Apr 2023

摘要—**AMR-WB 或 EVS 等宽带编解码器广泛应用于（移动）语音通信。编码语音质量通常通过绝对类别评分（ACR）听音测试进行主观评估。然而，ACR 测试不适用于语音通信网络的在线监测。语音质量感知评估（PESQ）是广泛使用的客观指标之一，可用于预测 ACR 测试结果。然而，PESQ 算法需要原始参考信号，而该信号在网络监测中通常无法获取，因此限制了其适用性。NISQA是一种新型的非侵入式基于神经网络的语音质量度量方法，主要针对超宽带语音信号。不过，在本研究中，我们旨在通过非侵入式PESQ-DNN模型预测广为人知的 PESQ 指标。我们通过预测在不同比特率下运行的 AMR-WB 或 EVS 编解码器所生成的宽带编码语音的 PESQ 分数，展示该模型在噪声、串联和易出错传输条件下的潜力。我们通过测量平均绝对误差（MAE）和线性相关系数（LCC），将所提出的方法与用于 PESQ 预测的先进网络拓扑结构QualityNet、WaveNet和DNSMOS进行比较。所提出的PESQ-DNN在无帧丢失条件下取得了最佳总体 MAE 和 LCC，分别为 0.11 和 0.92；纳入帧丢失后仍然表现最佳。请注意，我们的模型也可用于非侵入式预测 POLQA 或其他（侵入式）指标。文章获接收后，代码将在 GitHub 上提供。**

索引词—**客观语音质量度量、感知语音质量评估、POLQA、语音编解码器、语音通信。**

I. 引言

经语音编码器、传输系统和语音解码器处理的语音信号称为编码语音。语音编码器和解码器共同构成语音编解码器，其目标是以尽可能低的比特率将语音信号表示为数字形式。然而，编码语音的质量通常会受到远端加性背景噪声、量化噪声和传输错误的影响。现代传输系统，例如视频会议系统、数字蜂窝通信网络和互联网协议语音（VoIP），支持以 16 kHz 或更高采样率采样的宽带（WB）语音。

国际电信联盟电信标准化部门 G.722 [1] 是首个标准化的宽带语音编解码器，工作速率为 64 千比特每秒（或更低），广泛应用于视频会议系统和无绳电话。自适应多速率宽带（AMR-WB）编解码器 [2] 是应用最广泛的移动电话语音编解码器之一，提供九种工作模式，比特率范围为 6.6 至 23.85 千比特每秒。

Ziyi Xu 和 Tim Fingscheidt 任职于德国布伦瑞克工业大学通信技术研究所，邮编 38106（电子邮件：{ziyi.xu,t.fingscheidt}@tu-bs.de）。Ziyue Zhao 任职于荷兰埃因霍温 GN Hearing，邮编 5612AB（电子邮件：zzhao@gnhearing.com）。

自 AMR-WB 引入第三代（3G）通信系统及后续数字蜂窝服务以来，卓越的语音质量和语音自然度将人们对语音通信系统的期望提升到了更高的水平 [3]。与仅专注于语音信号的 G.722 和 AMR-WB 相比，增强语音服务（EVS）[4]是一种现代音频编解码器，可为语音和音乐信号提供高清（HD）音质。EVS 支持从窄带（NB）到全频带（FB）的音频信号，并应用于第四代（4G）通信系统及后续系统的 LTE 语音（VoLTE）服务。EVS 为宽带语音信号提供 11 种工作模式，比特率范围为 5.9 至 96 千比特每秒。

在通信系统的设计阶段，编码语音质量通常通过主观绝对类别评分（ACR）听音测试进行评估。在此测试中，未经训练的听音者（受试者）对每段语音片段的感知质量进行评分，分值范围为 1（差）至 5（极好）。所有受试者对某段语音片段或测试条件给出的 ACR 评分平均值称为平均意见分（MOS）。这类 ACR 测试通常在实验室环境中进行，以便按照国际电信联盟 T. P.800 [5] 标准控制影响受试者判断的外部因素。然而，准备和实施这类受控 ACR 测试耗时，而且招募未经训练的受试者也需要较高成本。

对于语音通信网络的在线监测，例如在运行期间根据编码语音质量动态调整编解码器的比特率模式，主观 ACR 测试并不实用。语音质量感知评估（PESQ） [6]，、感知客观听音质量评估（POLQA） [7], 和虚拟语音质量客观听音器（ViSQOL） [8] 都是用于评估（编码）语音质量的知名仪器指标。PESQ 旨在预测 ACR 听音测试的结果，因此得到了广泛应用。然而，PESQ、POLQA 和 ViSQOL 算法都需要原始参考信号（侵入式方法），通过测量经感知加权的距离来估计编码语音的感知质量。网络监测中无法获取此类参考信号，因此 PESQ、POLQA 和 ViSQOL 的适用性受到限制。

近年来，数据驱动方法在逼近高度非线性函数方面备受关注，甚至被用于估计人类感知。Abel等人训练了一种基于支持向量机的 MOS 预测器，用于侵入式预测窄带到宽带人工语音带宽扩展的主观 MOS 分数 [9]。深

度神经网络（DNN）进一步提升了性能，并使语音质量深度神经网络能够通过端到端训练来预测仪器指标或主观听音测试结果，甚至可以采用非侵入式 方式，无需参考信号 [10]–[23]。Soni等人 [10] 使用简单的自编码器从语音信号中提取特征，再利用这些特征训练单层神经网络，以预测固定长度窄带（NB）语音信号的主观质量分数。Fu等人 [11]提出了一种基于双向长短期记忆网络（BLSTM）结构的 QualityNet，用于根据宽带语音信号的幅度谱图预测感知语音质量评估（PESQ）分数。得益于 BLSTM 的循环结构，QualityNet 可以预测不同长度语音信号的 PESQ 分数。为预测传输后超宽带（SWB）语音信号的主观质量分数，Mittag 和 M¨oller 提出了一种基于 BLSTM 的深度神经网络，并将其命名为 NISQA [13], [18]。文献 [18] 提出了一种先进的 NISQA 版本，可预测全带（FB）语音信号的三种正交质量属性（噪声性、染色和不连续性）。大多数非侵入式语音质量预测深度神经网络都将语音信号的时频（TF）域表示作为输入。然而，Catellier等人 [16]提出使用名为 WaweNet 的卷积神经网络（CNN），直接根据宽带语音信号的波形预测 PESQ 分数。由于缺少循环结构，WaweNet 只能处理固定长度为三秒的语音信号。较长的语音信号会先被切分为若干个 3 秒片段，然后对所有片段估计出的 PESQ 分数取平均，作为该长语音信号的最终估计结果。

在微软深度噪声抑制（DNS）挑战赛中[24], [25], , Reddy 等人提出了一种基于深度神经网络的语音质量测量方法，称为 DNSMOS [19]。DNSMOS 深度神经网络模型专门经过训练，可依据国际电信联盟电信标准化部门 P.808 [26] 标准，为 DNS 任务中的增强语音信号预测主观评分。为更全面地呈现增强语音信号的感知质量，作者提出了 DNSMOS 的改进版本 [22], , 该版本分别评估语音成分、背景噪声以及整体增强语音的质量，并遵循国际电信联盟电信标准化部门 P.835 [27] 标准。两个版本的 DNSMOS 均要求输入语音信号的时长固定为九秒。在近期工作中 [20], [21], [23],, 我们提出了面向 DNS 应用的端到端 PESQNet，该网络改编自基于双向长短期记忆网络的语音情感识别深度神经网络，用于预测增强语音信号的 PESQ 分数。在这些工作中，训练好的 PESQNet 被用作中介器，在语音增强深度神经网络的训练过程中提供可微PESQ损失，以期最大化增强语音信号的 PESQ 分数。然而，该方法仅在 DNS 深度神经网络与 PESQNet 语音质量预测器遵循交替学习协议时才有效，而这在我们当前的工作中难以实现：我们需要一个已训练好的深度神经网络，用于对经由各种不同语音编解码器及模式转码的语音信号进行非侵入式质量预测。

大多数语音质量预测深度神经网络都侧重于估计因加性噪声而劣化的语音信号，或通过不同 DNS 方法增强的信号的感知质量——法。只有少数研究（例如 WaweNet [16] 和 NISQA [13], [18],的两个版本）估计编码语音信号的感知质量；其中后者明确聚焦于编码超宽带（SWB）语音信号。

在本研究中，我们提出了一种端到端非侵入式深度神经网络模型，并将其称为 PESQ-DNN，用于显式估计由各种宽带编解码器处理的编码语音信号的 PESQ 分数。我们还考虑了不同传输条件的影响，包括易出错传输中的丢包、远端加性噪声以及串联传输（即两个宽带编解码器串联）。据我们所知，此前尚未提出专门针对编码语音、并考虑广泛实际条件而开发的非侵入式 PESQ 估计深度神经网络。此外，与 WaweNet 或 DNSMOS 网络不同，我们提出的 PESQ-DNN 能够估计不同长度编码语音信号的 PESQ 分数。与大多数其他科学研究一样，我们没有遵循国际电信联盟电信标准化部门 P.565 [29] 和 P.565.1 [30] 中针对语音质量估计机器学习模型所标准化的评估流程。原因之一是，这些国际电信联盟电信标准化部门建议未考虑残余背景噪声条件，而我们认为这一条件非常重要。此外，建议中定义的模型要求将辅助信息作为语音质量预测模型的输入，例如丢包标志；然而，这与我们的目标相矛盾，因为我们希望仅 依据测量得到的语音信号来预测语音质量。我们指出，尤其是在易出错传输条件下，我们显然追求的是一个更具挑战性的目标。

在拓扑结构方面，我们以 [21], 为基础，但必须进行多项修改，使深度神经网络能够满足本研究所针对的语音通信监测需求：(1) 与 PESQNet 相比，新颖的 PESQ-DNN 采用复数频谱图作为输入，以明确考虑相位影响对感知语音质量的作用。除少数研究（例如 WaweNet）外，大多数语音质量预测深度神经网络都采用幅度或功率谱图作为输入，这导致无法测量由相位失真引起的语音质量下降。(2) 受 [18], 启发，我们引入了自注意力池化层，以替代 PESQNet 中使用的统计池化（平均值、标准差、最小值和最大值）。注意力层置于 BLSTM 层之后，使 PESQ-DNN 能够动态调整对从不同时间点获取的特征的关注重点，这有助于稳定训练过程并提升推理性能。(3) 不同于 PESQNet，我们研究了借鉴 [11], 的不同训练损失函数，为 PESQ-DNN 内部的中间嵌入赋予物理意义。随后，在所研究的损失函数中对中间嵌入的数值进行控制，并认为这能够稳定训练进程。除引入这些修改外，一项全面的消融研究还阐明了所引入修改的影响和价值，因此也构成了本研究的另一项核心贡献。

本文其余部分安排如下：第二节介绍信号模型和本文使用的数学符号。

第三节介绍我们提出的非侵入式 PESQ-DNN。随后，第四节说明实验设置，包括数据库、训练、验证、测试条件、基线方法和性能指标。第五节对结果进行全面的消融研究和讨论，第六节总结本文工作。

II. 信号模型与符号说明

我们假设发送端语音编码器的输入信号$y(n)$由原始语音信号 $s(n)$和可能存在的残余[1] 噪声 $d(n)$组成:

$$y(n) = s(n) + d(n), \tag{1}$$

其中， n 为离散时间采样索引。随后，信号 $y(n)$由某个语音编码器进行处理以获得比特流，该比特流随后经过传输系统，并由相应的语音解码器进行处理。最后，在接收端获得的编码语音记为 $\hat{y}(n)$。

所提出的非侵入式 PESQ-DNN 基于编码语音信号 $\hat{y}(n)$ 的离散傅里叶变换（DFT）频谱图 $\hat{Y}_\ell(k)$来估计其 PESQ 分数，其中帧索引为ℓ，频率箱索引为 $k \in \mathrm{K} = \{0, 1, \ldots, K-1\}$，且 K为 DFT大小。该变换始终通过快速傅里叶变换（FFT）执行，并且连续的 FFT 帧在时间上存在重叠。

III. 提出的非侵入式感知语音质量评估-DNN

在这项工作中，我们提出一种端到端的非侵入式PESQ-DNN，该模型以国际电信联盟电信标准化部门 P.862.2 感知语音质量评估为基础，但不使用参考信号输入，用于估计（无线）语音通信系统中编码宽带语音话语的 PESQ 分数。

A. PESQ-DNN拓扑结构

我们提出的端到端非侵入式 PESQ-DNN 如图1所示。每一层输入和输出特征图的维度表示为特征数× 时间帧数 × 特征图数量（如适用）。首先，我们研究了两种 PESQ-DNN，分别采用幅度谱图或复数频谱图作为输入。对于第一种 PESQ-DNN，网络的输入是编码语音的幅度谱图 $|S_\ell(k)|$，其中$\ell \in \mathrm{L}_u = \{1, 2, \ldots\ldots, L_u\}$，且 L_u表示索引为 u 的话语中的帧数。相应地，图1中的输入通道数设为 $C = 1$ 。另一方面，采用复值输入的动机很明确：以幅度谱图为输入的算法无法测量相位失真造成的语音质量下降。然而，原始的国际电信联盟电信标准化部门 P.862.2 感知语音质量评估函数在 PESQ 估计中会考虑相位影响，以某种程度.....

范围，例如通过估计退化语音信号样本与其对应参考信号样本之间的时间延迟。复值输入 PESQ-DNN 使用编码语音频谱图的实部和

[1]由于 $y(n)$ 并不表示麦克风信号，而是语音编码器的输入信号，因此 $d(n)$ 对潜在噪声抑制算法产生的残余噪声以及轻微的语音失真进行建模。

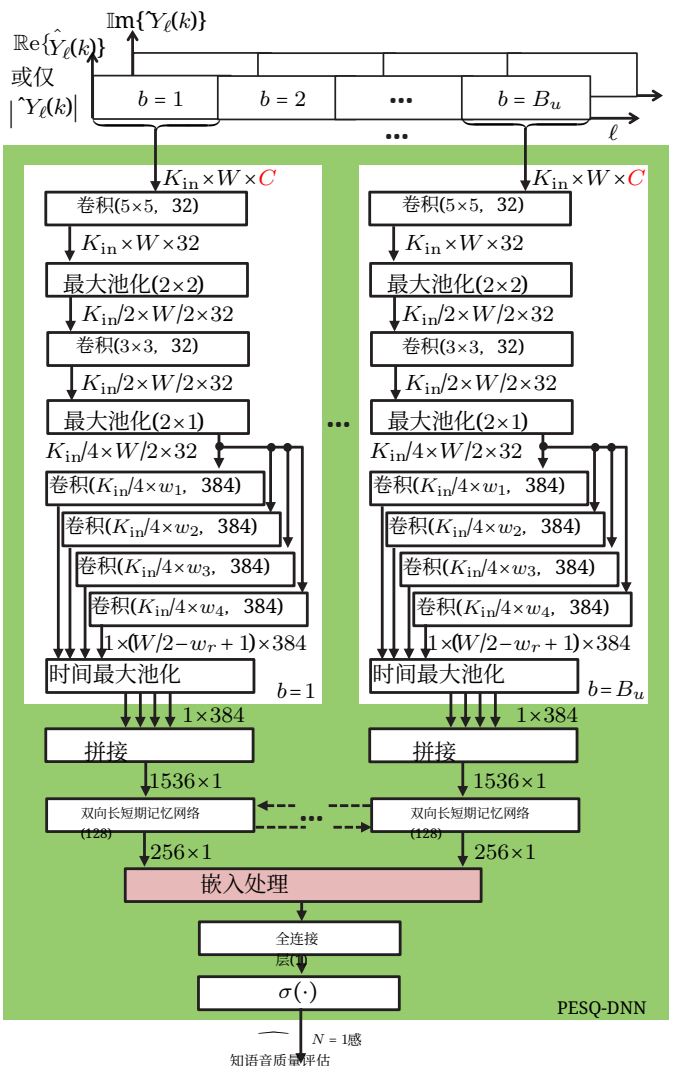

图1：所提出的 PESQ-DNN 的拓扑结构。对于分别以幅度或复数频谱图作为输入的 PESQ-DNN，输入通道数分别设为 $C = 1$ 或 $C = 2$ 。基线 PESQNet [21] 中使用的“嵌入处理”模型在图2中重述，而所提出的 PESQ-DNN 中使用的包含帧级或块级嵌入的模型如图3所示。

虚部，分别表示为 $\mathbb{R}\mathrm{e}\{\hat{Y}_\ell(k)\}$ 和 $\mathbb{I}\mathrm{m}\{\hat{Y}_\ell(k)\}$，作为两个独立的输入通道，从而得到图1中的 $C = 2$ 。对于这两种 PESQ-DNN，输入被划分为若干个特征块（特征矩阵），其索引为 $b \in \mathrm{B}_u = \{1, 2 \ldots, B_u\}$，其中 B_u 为话语 u 的块总数。每个特征块具有相同的维度 $K_{\text{in}}\times W\times C$，其中 K_{in} 和 W 分别表示每个块中的输入频率 bin 数和时间帧数。

随后，相同的子网络并行处理这些特征块：采用基于 CNN 的编码器，从输入特征块中提取质量相关特征。二维卷积层表示为 Conv($h \times w$， f)，其中 f 和 ($h \times w \times$ 输入特征图数量) 表示——

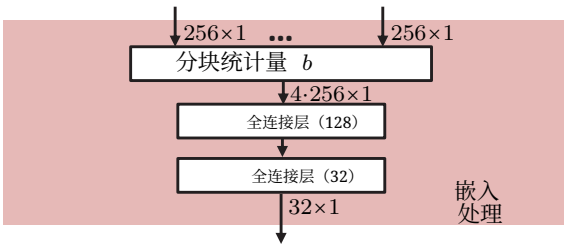

图2：图1中所示的、用于 [21]所提出的PESQNet的“嵌入处理”模型。对块 b上的双向长短期记忆网络输出应用四种统计量（平均值、标准差、最小值和最大值），以生成固定长度的特征向量。

分别表示滤波器核的数量及相应的核尺寸。在此 CNN编码器中，我们使用两种不同核尺寸的最大池化层，分别表示为（2×1 ×输入特征图的数量）和（2 ×2 ×输入特征图的数量）。提取的特征由多宽度卷积结构进行处理，该结构采用宽度为 w_1、 w_2、 w_3 和 w_4 的卷积核，以提取具有不同时间分辨率的特征。宽滤波器能够捕获长期信息，而窄滤波器则关注短期信息。时间最大池化层沿时间轴选取每个特征图中响应最高的值。随后进行拼接，得到一个维度固定的特征向量并输入 BLSTM 层。该层包含 128 个神经元，用于建模时间依赖关系。之后，我们研究多种“嵌入处理”模型（图1），将 BLSTM 层中每个话语数量不定的输出嵌入转换为定长表示。

图2展示了原始PESQNet中首次研究的“嵌入处理”模型，其中，对块 b 上的双向长短期记忆网络输出应用四种统计量（平均值、标准差、最小值和最大值）。随后，得到的特征向量具有固定维度 $4\cdot256\times1$，并由两个全连接层（FC）处理，这两个层记为FC（N），其中 N 为输出节点的数量。然而，由于存在严重的训练不稳定性问题，基线PESQNet在编码语音话语的PESQ估计方面表现不佳。可以想象，例如，丢包会零星地发生在某些输入块中，这可能会显著改变这些块所对应的双向长短期记忆网络输出的取值范围。因此，即使在同一话语中，双向长短期记忆网络输出也具有很大的动态范围。相应地，根据这些双向长短期记忆网络输出计算出的四种统计量，其取值范围也高度不稳定，最终导致训练稳定性问题。

训练稳定性问题的另一个原因可能是，在没有直接控制或指导的情况下，双向长短期记忆网络的输出不具备物理意义，因此会导致较大的动态范围

值范围。为缓解这一问题，我们采用了图 3 所示的改进“嵌入处理”模型，该模型的灵感来自 QualityNet [11]。其思路是为中间嵌入赋予实际含义，例如帧级嵌入（FLE）PESQ 分数或块级嵌入

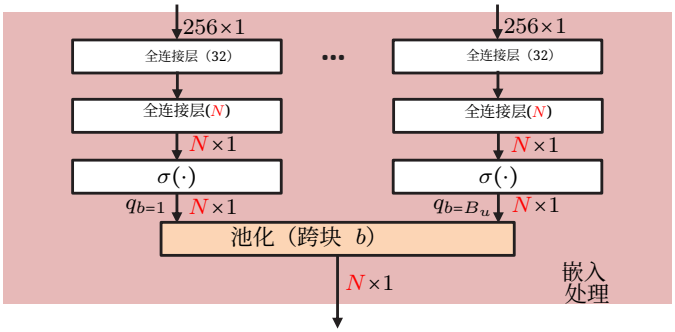

图3：图1中用于我们所提出的 PESQ-DNN 的“嵌入处理”模型。“池化”函数通过平均池化或注意力池化实现，如图4所示。

（BLE）PESQ 分数，尽管训练目标 PESQ 分数仍按语句准备。这个中间 FLE 或 BLE PESQ 分数用 q_b 表示，其中 $b\in B_u$ ={1，2..... B_u}。为估计 FLE PESQ 分数，我们设定$N=W$，表示第二个全连接层中的神经元数量 N 等于每个块中的时间帧数量 W。据此，我们估计对应输入块中每一帧的 PESQ 分数。对于 BLE PESQ 分数，我们设定 $N=1$，使其为每个块输出一个 PESQ 值。随后，全连接层的输出由门函数处理

$$\sigma(x)=3.6\cdot\mathrm{sigmoid}(x)+1.04 \tag{2}$$

以将估计的 FLE 或 BLE PESQ 分数限制在 1.04 到 4.64 之间，这与由国际电信联盟电信标准化部门 P.862.2 [6] 确定的 PESQ 范围一致。因此，如果某个输入块中发生丢包或出现噪声，则其对应的 FLE 或 BLE PESQ 分数将在有限范围内降低。随后，使用池化层将各块的 FLE 或 BLE PESQ 分数映射为单个值，如图3所示。对于池化层，我们研究了平均池化和注意力池化，分别称为“AV”和“AT”，其中所采用的注意力池化如图4所示。

注意力池化的思想受到 [18]，[31],的启发，旨在明确考虑从语音质量较差甚至存在语音停顿的输入块中获得的 FLE 或 BLE PESQ 分数对最终 PESQ 估计的贡献。如图4所示，使用 Softmax 作为输出激活函数的全连接层为每个块估计注意力分数，记为 a_b，其中$a_b\in[0,1]$ 和 $b\in B_u$。随后，将每个块的 FLE 或 BLE PESQ 分数乘以估计的注意力分数，并在各块上求和，如图4所示。

如图1所示，无论是图2中的嵌入处理模型，还是图3中提出的模型，其输出均由包含单个输出节点的全连接输出层进行处理，随后通过门函数 (2) 估计最终的话语级 PESQ 分数。

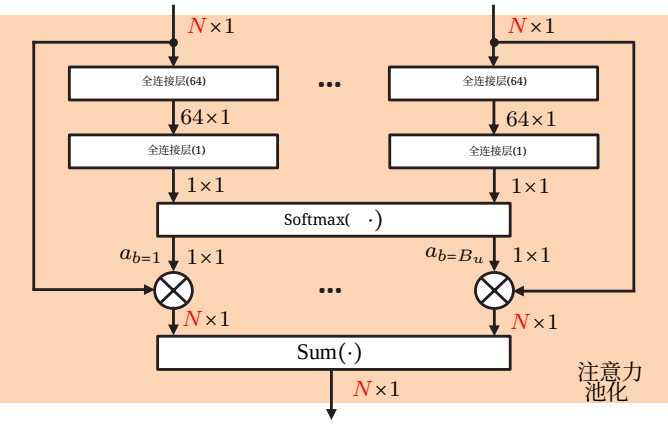

图4：图3所示的池化方式，此处为注意力池化。我们分别将 $N = W$ 和 $N = 1$ 设为帧级嵌入（FLE）和块级嵌入（BLE）。

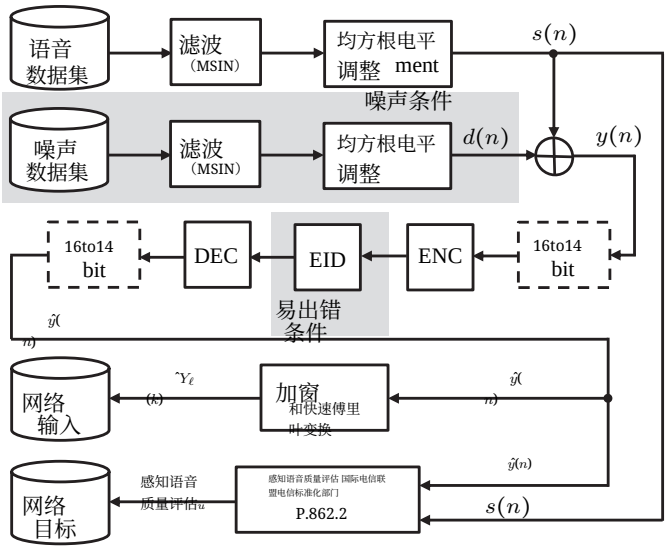

图5：针对所提出的 *PESQ-DNN*，在纯净、有噪声、易出错传输或串联条件下采用各种编解码器时，对训练、验证和测试数据的处理。函数“ENC”和“DEC”分别表示所研究编解码器的编码器和解码器。

B. 训练损失选项

编码语音话语的估计PESQ分数应尽可能接近根据国际电信联盟电信标准化部门 P.862 测得的真实PESQ分数。2 [6] 因此，训练我们提出的 PESQ-DNN 的一种方法是采用所谓的感知语音质量评估损失，其定义如下：

$$J_u^{\text{PESQ}} = \left(\widehat{\text{PESQ}}_u - \text{PESQ}_u\right)^2, \tag{3}$$

$\widehat{\text{PESQ}}_u$，其中 PESQ_u 和 PESQ_u 分别为根据国际电信联盟电信标准化部门 P.862 测得的估计 PESQ 分数和对应的真实 PESQ 分数。2 [6], 分别表示，u表示话语索引。

为训练采用中间 FLE PESQ 分数的 PESQ-DNN，我们在损失函数(3)中引入一个额外约束：

$$J_u^{\text{PESQ}} = \left(\widehat{\text{PESQ}}_u - \text{PESQ}_u\right)^2 + \frac{\alpha_u}{B_u \cdot L_b} \sum_{b \in \mathcal{B}_u} \sum_{\ell \in \mathcal{L}_b} \left(q_b(\ell) - \text{PESQ}_u\right)^2, \tag{4}$$

其中，$q_b(\ell)$ 是针对索引为 b 的块所预测的中间帧级 PESQ 分数，帧索引为 $\ell \in \text{L}_b = \{1, 2, \ldots\ldots, W\}$。参数 $B_u = |\text{B}_u|$ 和 $L_b = |\text{L}_b| = W$ 分别表示索引为 u 的话语所包含的块总数，以及每个块中的帧数。话语级加权因子表示为

$$\alpha_u = 0.9^{|\text{PESQ}_u - \text{PESQ}_{\max}|}, \tag{5}$$

其中，$\text{PESQ}_{\max} = 4.64$ 是国际电信联盟电信标准化部门 P.862.2 [6]中定义的最大感知语音质量评估分数。其思想是，在损失函数（4）中，α_u 所反映的约束对于质量更好的语音话语应当更高：随后，所有中间帧级或块级感知语音质量评估分数都会被促使等于话语级感知语音质量评估分数。可以设想，一段整体感知质量完美的语音话语，应当在各处都标注为同样的高质量，即每个块中都具有相同的高 PESQ 分数。请注意，当 $\alpha_u = 0$时，损失函数（4）将简化为话语级感知语音质量评估损失（3）。

采用中间 BLE 的 PESQ-DNN 得分方面，损失函数（4）可简化为：

$$J_u^{\text{PESQ}} = \left(\widehat{\text{PESQ}}_u - \text{PESQ}_u\right)^2 + \frac{\alpha_u}{B_u} \sum_{b \in \mathcal{B}_u} \left(q_b - \text{PESQ}_u\right)^2, \tag{6}$$

其中 q_b 是中间的块级 PESQ分数，对应索引为 b的块，参见图3。

对于损失函数（4）和（6），话语级加权因子 α_u 由（5）计算得出。

IV. 实验设置与数据库

A. 数据库与预处理

本文用于训练、开发和测试的语音信号采样率为16kHz，取自NTT宽带语音数据库 [32],。该数据库包含21种语言，每种语言由4名女性和4名男性说话人代表。每位说话人提供12段语音片段，每段时长为8秒。美式英语和德语的所有语音片段均用于测试集。我们的训练集由其余19种语言中每种语言前三名女性和男性说话人的所有语音片段构成。同时，开发集使用这19种语言中其余女性和男性说话人（标记为“f4”和“m4”）的所有语音片段。因此，本文进行的测试完全独立于说话人，且在一定程度上独立于语言：英式英语是用于训练和开发的19种语言之一，而测试中使用的是美式英语。

数据预处理流程如图5所示。在本研究中，训练、开发和测试数据根据用于评估编码语音信号的质量评估计划 [33]–[36]进行预处理。图5中的函数取自 ITU-T G.191 软件工具库[37]。

对于我们提出的 PESQ-DNN，在图5所示的“加窗与快速傅里叶变换”函数中，我们采用帧长为 512、重叠率为 50% 的周期性汉恩窗。得到的离散傅里叶变换频谱图 $\hat{Y}_\ell(k)$ 将保存为 PESQ-DNN 的输入。训练所需的相应真实 PESQ目标值取自 ITU-T P.862。2 [6], 如图5所示，输入为编码语音信号及其原始参考语音信号。

1）训练与开发条件：语音话语首先经过 MSIN高通滤波器[37]处理。随后，按照国际电信联盟电信标准化部门 P.56 [38], 根据均方根（RMS）电平调整滤波后语音话语的活动语音电平。我们为每段语音话语设定三个 RMS 电平 $s(n)$，即 -36、-26 和 -16 dBov。为明确考虑噪声条件下编码语音质量的劣化，编解码器输入信号 $y(n)$ 包含加性噪声 $d(n)$，见式（1）。噪声信号取自 DEMAND [39] 和 QUT [40], ，共包含训练与开发阶段使用的 35 个不同噪声文件。加性噪声信号经过与语音话语相同的 MSIN高通滤波器处理，并根据所需信噪比（SNR）调整噪声信号的 RMS 电平。对于每个活动语音电平，我们分别模拟 10%、10% 和 80% 的语音话语，使其信噪比为 15、20 和 ∞ dB；其中，∞ dB 的信噪比表示不含加性噪声的无噪声条件。这些相对较高的信噪比数值考虑到了实际情况：麦克风路径信号处理会在语音编码前使用噪声抑制器，通常已能将信噪比提高 10...20 dB。

调整语音话段的活动语音电平，并将其与加性噪声混合（如有考虑）后，所得的发送端语音编码器输入信号 $y(n)$ 将进行编码。图5中的“ENC”和“DEC”分别表示所研究编解码器的编码器和解码器，并包含用于时间对齐的时延补偿。本文考虑训练和开发阶段使用不同比特率的 AMR-WB [2] 和 EVS [4] 编解码器获得的编码语音。对于 AMR-WB 编解码器，我们采用其定点实现，在不启用不连续传输（DTX） [41] 的情况下使用七种比特率（模式）：6.6、8.85、14.25、15.85、18.25、19.85 和 23.05 千比特每秒。对于 EVS 编解码器，我们采用其定点实现 [42] ，使用五种比特率：5.9、8.0、9.6、16.4 和 24.4 千比特每秒。请注意，以 5.9 千比特每秒运行的 EVS 需要启用 DTX，而其他模式均在未启用 DTX 的情况下实现。

图5中的错误插入设备（EID）用于将帧丢失插入从“ENC”获得的比特流中，以模拟易出错传输条件。在本研究中，我们在干净语音话语中使用 EVS 实现帧丢失。我们考虑两种类型的帧丢失：随机帧擦除和突发帧擦除。随机帧擦除基于 Gilbert模型 [43], ，而突发帧擦除采用 Bellcore模型 [44]。两种帧擦除均通过帧错误率（FER）进行衡量，该指标反映失真帧数与所有传输帧数之比。在训练和开发阶段，我们采用 3% 和 10% 的 FER，并将其平均分配给两种类型的帧擦除，应用于每种活动语音电平下 10% 的纯净语音话语。请注意，发生帧丢失时，语音解码器（“DEC”）会提供错误隐藏机制，并仍输出一些估计语音 $\hat{y}(n)$。

在串联（TDM）传输条件下，“EID”被替换为“DEC”及其后续的“ENC”，从而形成两个宽带编解码器的串行拼接。请注意，为了使实验总数保持在可控范围内，本研究在 TDM 评估中未考虑加性噪声和易出错传输。在训练和开发过程中，语音话语首先经过 G.722 [1] 处理，然后经过以 9.6 和 24.4 千比特每秒运行的 EVS[4],处理。图5中的“16to14”函数用于将 16 比特转换为 14 比特，以满足 AMR-WB 和 G.722（在 TDM 条件下）的要求。

在本研究中，训练集和开发数据集分别记为 $\mathrm{D}^{\mathrm{train}}$ 和 $\mathrm{D}^{\mathrm{dev}}$。如上所述，开发集包含与训练集相同的条件，但语音话语来自不同的说话人。我们使用开发集 $\mathrm{D}^{\mathrm{dev}}$ 不仅在训练过程中进行学习率调度和早停，还根据表 I 所示的开发集性能，选出最佳的 PESQ-DNN 方案；这一点将在后文讨论。

2) 测试条件：测试所用的语音话语采用与训练相同的方式进行预处理，包括 MSIN滤波和 RMS 电平调整。每个测试语音话语的活动语音电平均归一化至 -36、-26 和 -16 dBov。我们明确评估了所提出的 PESQ-DNN 在无噪声、含噪、易出错以及 TDM 传输条件下获得的编码语音上的性能。

在无噪声条件下，加性噪声的信号路径和图5中的“EID”功能分别被停用和旁路。对于此无噪声测试条件，我们从每种活动语音电平下的测试语音话语中选取 80%。其余 20% 的测试语音话语将用于制备含噪语音话语，以供含噪条件下的测试使用。对于无噪声测试条件，我们考察了由工作于 12.65 和 23.85 千比特每秒的 AMR-WB，以及工作于 7.2、13.2 和 48 千比特每秒的 EVS 编码得到的编码语音话语。请注意，训练和开发期间未使用 AMR-WB 和 EVS 的这些工作模式，因此可以展示所训练 PESQ-DNN 的编解码器模式泛化能力。所提出的 PESQ-DNN 在无噪声条件下的性能见表 II，后文将对此进行讨论。

在噪声 条件下，加性噪声的信号路径处于启用状态，而图5中的“EID”功能则被旁路。我们排除易出错传输条件，专门研究加性噪声对编码语音的 PESQ 估计的影响。请注意，我们报告的是 EVS 编解码器以 13.2 kbps 运行时的噪声测试条件。我们从 ETSI背景噪声数据库 [45], 中选取三种未见过的 噪声类型，这些类型与训练和开发期间所用噪声数据库中的类型不同。所选的未见过噪声类型包括自助餐厅噪声、时速为 100 km/h 的汽车噪声以及道路交通噪声。为准备噪声测试条件下的语音话语，我们取每种活动语音电平 下剩余的 20% 测试语音话语，并将其与未见过类型的噪声信号混合，以模拟 15 dB 的信噪比。

稍后我们将讨论表 III 所示噪声条件下 PESQ-DNN 的性能。

在易出错传输条件下，图5中的“EID”功能处于开启状态，而噪声信号路径处于停用状态，因此我们在纯净条件下开展研究。此外，为使实验总数保持在可控范围内，我们报告了采用 EVS 编解码器、运行速率为 13.2 和 48 千比特每秒时 EID 的影响。为准备受 EID 损伤的编码语音话语，我们在每种活动语音电平 下选取 50% 的测试语音话语，并模拟 3% 和 6% 的 FER，使其在随机帧擦除和突发帧擦除之间均匀分布。请注意，6% 的 FER 在训练和开发期间是未见过的。

在串联（*TDM*）传输 条件下，我们在所有测试语音话语上采用两种 TDM 情况：对于已见过的TDM 情况，两个编解码器按照训练和开发中使用的相同顺序串联，即先由 G.722 处理语音话语，然后由 EVS 以 13.2 千比特每秒处理。对于未见过的 TDM 情况，我们调换两个编解码器的顺序，先使用 EVS 以 13.2 千比特每秒处理，随后使用 G.722。对于这两种 TDM 传输情况，我们排除加性噪声和易出错传输的影响。

在以下讨论中，所采用的测试集记为 $\mathrm{D}^{\mathrm{test}}$。如表 II、III、IV、V 和 VI 所示，我们分别报告 PESQ-DNN 在 $\mathrm{D}^{\mathrm{test}}$ 各条件下的性能。我们将在第五节中对这些表格进行详细分析。

B. 基线方法

在已有研究中，有两种方法 [11], [16] 直接预测 PESQ，另一种方法则可以轻松改造为预测 PESQ[19]。在使用我们的数据集重新训练后，这三种方法将作为本文的基线，从而能够直接比较实际采用的网络拓扑结构。下面将简要介绍这三种方法。

1) QualityNet DNN： Fu 等人 [11] 提出了所谓的 QualityNet，用于预测受加性噪声干扰的语音信号或由特定 DNS 模型获得的增强语音信号的 PESQ 分数。QualityNet 的输入是语音幅度谱图。作为本文的一种基线方法，我们采用 QualityNet 深度神经网络拓扑结构，通过在训练过程中替换输入及相应目标，来预测编码语音的 PESQ 分数。因此，QualityNet 基线深度神经网络的拓扑结构与 [11]中提出的完全相同。

与我们提出的 PESQ-DNN 类似，QualityNet基线深度神经网络采用双向长短期记忆网络层作为循环结构，以处理长度可变的编码输入语音信号。请注意，损失函数 (4) 和 (6) 在 PESQ-DNN 训练期间显式控制中间 FLE 和 BLE PESQ 分数，其灵感来自 QualityNet 训练中使用的损失函数 [11]。因此，通过与该基线进行比较，我们旨在展示所提出的 PESQ-DNN 仅凭网络结构所获得的优势。该基线在结果表中记为“QualityNet 基线深度神经网络 [11]”。

2) WaweNet 深度神经网络：作为另一种基线，我们采用 WaweNet 在 [16] 中提出的拓扑结构，直接根据编码语音信号的波形预测 PESQ 分数。与本研究类似，[16] 中最初提出的 WaweNet 根据宽带编解码器生成的编码语音信号预测 PESQ 分数，但仅考虑加性背景噪声的影响。请注意，将原始波形乘以 -1 以反转语音信号的波形，不会影响感知质量或测得的 PESQ 分数。这种反转不变性可以作为训练目标之一。为此，我们遵循 [16] 的作者，对 WaweNet 使用的所有训练和开发数据执行逆相位增强（IPA）。此外，由于缺少循环层，WaweNet 只能处理长度固定为三秒的语音信号。在本研究中，将较长的语音片段切分为若干个互不重叠的三秒片段，并丢弃短于三秒的余数。该较长语音片段最终估计的 PESQ 分数等于所有片段估计值的平均值。在本研究中，我们将该基线称为“WaweNet 基线深度神经网络[16]”。

3) DNSMOS 深度神经网络： Reddy 等人提出了一种基于深度神经网络的度量方法，称为 DNSMOS [19]，用于根据 ITU-T P.808 [26] 预测由 DNS 模型获得的增强语音信号的人类主观评分。 [22] 中提出的改进版本可按照 ITU-T P.835[27]分别估计语音成分、背景噪声和整体增强语音质量的人类评分。由于本研究不预测 DNSMOS，且噪声抑制算法也不是我们的研究重点，因此我们采用 [19]中提出的 DNSMOS 深度神经网络拓扑结构 作为进一步的基线，但在训练期间替换输入及其对应的目标。请注意，改编后的 DNSMOS 深度神经网络以对数功率谱图作为输入，快速傅里叶变换大小为 $K = 320$。 [19] 中提出的 DNSMOS 深度神经网络采用基于卷积神经网络的结构，用于预测固定长度为九秒的输入语音话语的人类评分。在本研究中，所用语音话语的长度为八秒。因此，在改编后的 DNSMOS 基线的训练、开发和测试中，所有语音话语均进行零填充，使其长度达到九秒。在以下讨论中，此基线称为“DNSMOS 基线深度神经网络 [19]”。

C. 性能指标

在 [9], [11] 和 [19], 之后，通过平均绝对误差（平均绝对误差）测量 PESQ-DNN 和基线模型的性能

$$\mathrm{MAE} = \frac{1}{U_{\mathrm{test}}} \sum_{u \in \mathcal{U}_{\mathrm{test}}} \left| \widehat{\mathrm{PESQ}}_u - \mathrm{PESQ}_u \right| \tag{7}$$

以及线性相关系数（线性相关系数）

$$\mathrm{LCC} = \frac{\sum_{u \in \mathcal{U}_{\mathrm{test}}} \left(\widehat{\mathrm{PESQ}}_u - \hat{\mu} \right) \cdot \left(\mathrm{PESQ}_u - \mu \right)}{\sqrt{\sum_{u \in \mathcal{U}_{\mathrm{test}}} \left(\widehat{\mathrm{PESQ}}_u - \hat{\mu} \right)^2} \cdot \sqrt{\sum_{u \in \mathcal{U}_{\mathrm{test}}} \left(\mathrm{PESQ}_u - \mu \right)^2}} \tag{8}$$

其中，$\widehat{\mathrm{PESQ}}_u$ 和 PESQ_u 分别是语音片段对应的估计 PESQ 分数和真实感知语音质量评估分数

索引为 u，且属于 $u \in \mathrm{U}_{\mathrm{test}}$。在式（7）中，集合 $\mathrm{U}_{\mathrm{test}}$ 中的语音片段总数记为 $U_{\mathrm{test}} = |\mathrm{U}_{\mathrm{test}}|$。集合 $\mathrm{U}_{\mathrm{test}}$ 上估计 PESQ 分数和真实 PESQ 分数的均值分别表示为 $\hat{\mu}$ 和 μ。MAE 和 LCC 均根据 PESQ-DNN 得到的估计 PESQ 分数及其依照 ITU-T P.862.2 PESQ [6]测得的对应真实分数计算。估计的 PESQ 分数越准确，MAE 越低（理想值为 0），LCC 越高（理想值为 1.0）。

D. 训练设置

对于我们提出的 PESQ-DNN 和 QualityNet 基线深度神经网络，网络的输入按照 QualityNet 提出者在 [46]中的建议，使用在训练数据集上收集的统计量进行归一化，使其均值为零、方差为单位值。与此同时，对于 WaweNet 基线深度神经网络和 DNSMOS 基线深度神经网络，相应的波形和对数功率输入不 进行归一化，因此遵循 [16], [19]。

图1中的输入和输出频率箱数量设为 $K_{\mathrm{in}} = 260$。最后 3 个频率箱是冗余的，用于兼容图1所示 PESQ-DNN 中采用的两次最大池化操作。图1所示 PESQ-DNN 中使用的卷积核宽度设为 $w_i = 2^{i-1}, i \in \{1, 2, 3, 4\}$。图1所示每个特征块中的时间帧数量设为 $W = 16$。除 softmax（图4）或门函数（2）（图1和图3）之前采用线性激活函数的层外，PESQ-DNN 的其余层均使用带泄漏修正线性单元（ReLU） [47] 作为激活函数。

在 PESQ-DNN 训练过程中，我们采用损失函数 (3)，也可以分别采用 (4) 和 (6)，以显式控制中间的 FLE 和 BLE PESQ 分数。损失函数 (4) 和 (6) 中的话语级加权因子 α_u 根据 (5) 计算。请注意，(4) 和 (6) 改编自 QualityNet 基线训练所用的损失函数，其中由 α_u 加权的第二个损失项，是对属于当前输入话语的所有帧取平均计算得出的，且不将这些帧分组为块。其他基线模型则采用损失函数 (3) 进行训练，这是其原始研究 [16], [19]中提出的方法。

对于 PESQ-DNN 的训练，我们采用截断时间反向传播（BPTT）训练方案，序列长度（BPTT 的展开深度）等于属于当前输入话语（索引为 u）的特征块总数 B_u。训练 QualityNet 基线时也采用类似的截断 BPTT 方案，但序列长度设为属于当前输入话语的时间帧数。对于本研究中的所有训练，我们均采用 Adam优化器，初始学习率为10^{-4}。当在 $\mathrm{D}^{\mathrm{dev}}$ 上测得的开发集损失连续两个训练轮次没有改善时，学习率乘以 0.6。开发集损失连续六个训练轮次没有改善后，我们停止训练，并保存开发集损失最低的模型。

表 I：对所提出方法进行的消融研究，结果为在开发集 PESQ-DNN 上的表现，使用 **开发集** $\mathcal{D}^{\mathrm{dev}}$。该 PESQ-DNN 变体使用幅度（$C = 1$）或复数谱图（$C = 2$）作为输入。采用块级嵌入（BLE）或帧级嵌入（FLE）的 PESQ-DNN 拓扑结构见图1和图3。“AV”：平均池化；“AT”：注意力池化，见图4。最佳结果以**粗体**显示，次佳结果以下划线标出。最佳和次佳PESQ-DNN 变体分别以 □ 和△标记。

Input	方法	MAE				LCC			
		EVS	AWB	TDM	总计	EVS	AWB	TDM	总计
Amplitude	PESQNet [21], $J(3)$	0.40	0.41	0.21	0.37	0.07	0.10	0.13	0.08
	PESQ-DNN（BLE）， $J(6)$，音视频	0.40	0.39	0.18	0.36	0.32	0.51	0.52	0.44
	感知语音质量评估-DNN (FLE)，$J(4)$, 音视频	0.43	0.41	0.23	0.39	0.37	0.53	0.65	0.48
	PESQ-DNN（BLE）、$J(3)$、音视频	0.41	0.39	0.19	0.37	0.31	0.50	0.53	0.43
	PESQ-DNN（FLE）、$J(3)$、音视频	0.43	0.44	0.10	0.39	0.35	0.49	0.56	0.45
	PESQ-DNN（BLE）， $J(3)$，AT	0.43	0.43	0.23	0.40	0.37	0.53	0.60	0.48
	PESQ-DNN（FLE）， $J(3)$，AT	0.46	0.46	0.25	0.43	0.24	0.44	0.47	0.37
Complex	PESQNet， $J(3)$	0.41	0.41	0.21	0.38	0.08	0.11	0.12	0.08
	PESQ-DNN（BLE）， $J(6)$，音视频	0.20	0.12	0.07	0.14	0.77	**0.94**	**0.94**	**0.88**
	PESQ-DNN（FLE）， J（4），音视频△	**0.19**	0.12	**0.06**	0.14	**0.78**	**0.94**	**0.94**	**0.88**
	PESQ-DNN（BLE）， $J(3)$，音视频	**0.19**	0.12	**0.06**	0.14	**0.78**	0.93	0.93	0.87
	PESQ-DNN（FLE）， $J(3)$，音视频	**0.19**	0.12	**0.06**	0.14	**0.78**	0.93	0.93	0.87
	PESQ-DNN（BLE）， $J(3)$，AT□	**0.19**	**0.11**	**0.06**	**0.13**	**0.78**	**0.94**	0.93	**0.88**
	PESQ-DNN（FLE）， $J(3)$，注意力池化	**0.19**	0.12	**0.06**	0.14	**0.78**	**0.94**	0.93	**0.88**

V. 实验结果与讨论

A. 新 *PESQ-DNN* 的消融研究

本节明确考察我们提出的新型 PESQ-DNN 中哪些改动最有助于稳定训练过程并提升性能。如表I所示，我们对已训练的 PESQ-DNN 变体进行全面的消融研究，分析其在开发集 $\mathrm{D}^{\mathrm{dev}}$ 上的性能。同时，我们将根据这些变体在开发集上的表现，选出两个最佳的 PESQ-DNN 变体作为所提模型，并在后续测试集 $\mathrm{D}^{\mathrm{test}}$ 上与基线方法进行比较。对于每个已训练的 PESQ-DNN 变体，我们报告其在来自 EVS、AMR-WB（AWB）以及串联（TDM）条件（先 G.722，后 EVS）的编码语音上，按相应比特率计算的平均性能。与训练集一样，开发集中的 EVS 编码语音也考虑了噪声和易出错传输条件的影响。PESQNet[21] 和复数输入 PESQNet 使用损失函数（3）。图3所示的采用 FLE 和 BLE 的 PESQ-DNN 变体使用损失函数（3）、（4）或（6）进行训练。标记为“AV”的模型使用平均池化，“AT”表示图4所示的注意力池化。

首先，我们研究不同输入表示形式（幅度谱或复数谱图）对参考PESQNet [21] 和我们提出的PESQ-DNN的影响，另请参见图1中的 $C = 1$ 或 $C = 2$。关于平均绝对误差（7），表I上半部分（幅度谱）中看不出我们新提出的PESQ-DNN变体相较参考PESQNet [21]有明显优势。然而，在LCC（8）方面，新提出的PESQ-DNN有所改进（总体上约为0.40，而参考模型约为0.08）。不过，低于0.50的LCC结果仍不足以构成令人信服的解决方案。表I下半部分显示，

表 I：**干净条件**下的**测试集**$\mathrm{D}^{\mathrm{test}}$ 上的性能。分别报告 EVS 和 AMR-WB (AWB) 在不同测试比特率下获得的编码语音的性能。最佳结果以**粗体**显示，次佳结果加下划线。

	方法	MAE								LCC							
		AWB			EVS				总计	AWB			EVS				总计
	比特率（千比特每秒）	12.65	23.85	Avg.	7.2	13.2	48	Avg.	-	12.65	23.85	Avg.	7.2	13.2	48	Avg.	-
REF	QualityNet 基线深度神经网络 [11]	0.16	0.09	0.13	0.30	**0.08**	0.29	0.25	0.19	0.79	0.83	0.74	0.52	0.61	**0.67**	0.27	0.43
REF	WaweNet 基线深度神经网络 [16]	0.14	0.15	0.15	0.14	0.16	0.20	0.17	0.16	0.76	0.69	0.77	0.51	0.52	0.27	0.84	0.82
REF	DNSMOS 基线深度神经网络 [19]	0.65	0.75	0.70	0.73	0.73	0.76	0.74	0.71	0.80	0.75	0.77	0.48	0.59	0.21	0.83	0.81
NEW	PESQ-DNN（FLE） $J(4)$，音视频△	0.10	**0.08**	**0.09**	**0.09**	**0.08**	0.10	**0.09**	**0.09**	**0.84**	**0.89**	**0.89**	0.70	**0.63**	0.61	**0.94**	**0.92**
NEW	PESQ-DNN（BLE） $J(3)$，AT□	**0.09**	0.09	**0.09**	0.10	**0.080.09**		**0.09**	**0.09**	**0.84**	0.86	0.86	**0.73**	0.61	0.49	0.93	0.91

将原始PESQNet调整为接收复值输入后，其性能实际上仍与接收幅度输入的参考PESQNet [21] 一样差（总体平均绝对误差 = 0.38，总体LCC = 0.08）。然而，有趣的是，我们新提出的PESQ-DNN从采用复数频谱图输入中获益显著，这一点从明显更低的平均绝对误差结果（≈ 0.14）和更高的*LCC*结果（≈ 0.88）中一目了然。另一方面，在六种PESQ-DNN变体之间，我们看不出显著差异，无论是在BLE或FLE方面，还是在损失函数或池化方法方面。因此，在后续实验中，我们保留由绿色符号标记的两种最佳方法（七 1st 个排名和一个 2^{nd}个排名，以及六 1st 个排名和两个 2nd 个排名），二者均采用复数频谱输入。值得注意的是，我们提出的PESQ-DNN相较参考PESQNet [21] 在性能上的大幅提升，可归因于池化前图3中的中间嵌入 q_b；无论训练时是否对图3中的中间嵌入 q_b 进行显式约束（损失函数（4）、（6）），这一设计似乎都能带来优势（未显式约束时使用损失函数（3））。

B. 新的PESQ-DNN 与基线方法的比较

1) 无噪声条件：首先，我们在 $\mathrm{D}^{\mathrm{test}}$ 上评估所提出的PESQ-DNN 变体和所有基线深度神经网络在无噪声条件 下的表现，使用以不同比特率运行的 EVS 和 AWB 得到的编码语音信号，如表 II 所示。为了更全面地展示结果，我们还分别报告 AWB 和 EVS 在所有测试比特率下的平均值性能，以及

总体性能是对所有测试条件的数据进行平均后得到的（不是对平均绝对误差或线性相关系数取平均！）。请注意，表 II 中未考虑加性噪声、易出错传输和串联（TDM）的影响。

在所有方法中，我们发现，以平均绝对误差为评估指标时，DNSMOS 基线深度神经网络最不适合预测PESQ 分数（总平均绝对误差 = 0.71），而其线性相关系数几乎是基线方法中最好的。另一方面，在所有方法中，QualityNet基线的相关性最差（总线性相关系数 = 0.43），因此也不适合预测PESQ。相比之下，WaveNet 基线深度神经网络表现出相当出色且均衡的性能，总平均绝对误差为 0.16，总线性相关系数为 0.82，在基线方法中明显位居 1st。不过请注意，我们新提出的两种 PESQ-DNN 在平均绝对误差和线性相关系数指标上的总体性能甚至优于基线方法。在除一种情况外的每个单独条件下，它们的性能均与基线方法持平或显著更好：对于 48 千比特每秒的 EVS，我们的方法表现稍逊

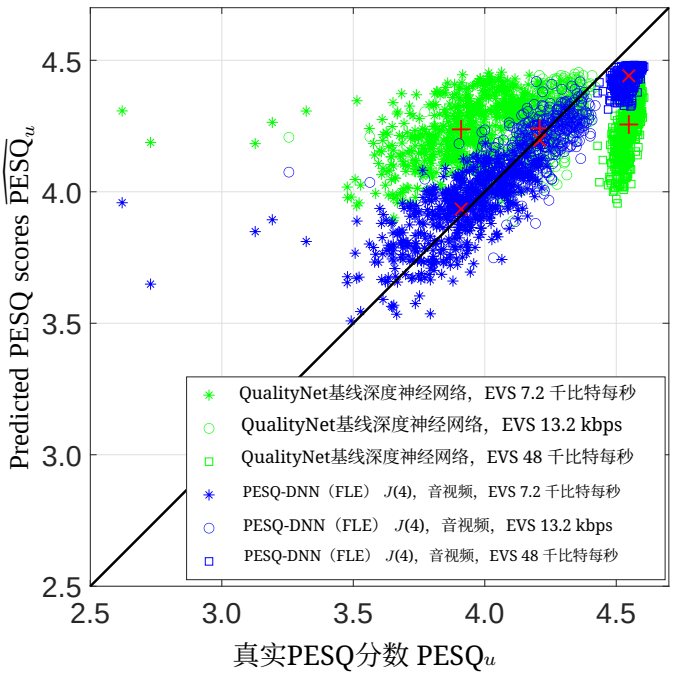

图6：散点图展示了在 $\mathrm{D}^{\mathrm{test}}$ 的**干净条件**下，对**EVS 编码**语音（7.2、13.2 和 48 千比特每秒）进行测量时，由性能最强的 PESQ-DNN 变体和经过训练以预测 PESQ 的 QualityNet 基线深度神经网络得到的预测 PESQ 分数 PESQ_u。标记 + 和 × 分别表示在 EVS 模式为 $j \in \{7.2, 13.2, 48\}$ 千比特每秒时，单条件 LCC_j (8) 对 $\widehat{\mathrm{PESQ}}_u$ 得出的 $\hat{}\mu_j^{\mathrm{QualityNet}}$ 和 $\hat{}\mu_j^{\mathrm{PESQ\text{-}DNN}}$ 的均值（从左到右）。

QualityNet 的 LCC = 0.67性能（尽管存在不平衡）不及我们的方法，但在较低的 EVS 比特率 7.2 千比特每秒时，我们的方法明显更胜一筹（LCC = 0.70/0.73 vs. 0.52）。总而言之，MAE 达到 = 0.09 vs. 0.16，我们的方法几乎将 *MAE* 降至最佳基线方法 *WaweNet* 的一半。在 *LCC* 方面，我们的两种方法均达到 0.90 以上，我们认为与基线方法相比，这是一个令人鼓舞的结果。请注意，由于我们采用了 QualityNet 的 [11] 基于 FLE 的中间嵌入 q_b，因此本实验中出色的相关性性能并非来自这些嵌入，而是源于我们显然更具优势的 PESQ-DNN 网络结构的其余部分。

在表II中，我们观察到 PESQ-DNN 各编解码模式的LCC指标均低于 0.90，但平均值为

表 III：纯净和**未见过的噪声条件**下的性能。性能是在来自**测试集**D^{test}、以 13.2 kbps 编码的 EVS 语音上测得的。最佳结果以**粗体** 显示，次佳结果带下划线。

	方法	纯净		噪声	
		平均绝对误差	线性相关系数	平均绝对误差	线性相关系数
REF	QualityNet 基线深度神经网络 [11]	**0.08**	0.61	1.54	0.23
REF	WaveNet 基线深度神经网络 [16]	0.16	0.52	**0.26**	0.57
REF	DNSMOS 基线深度神经网络 [19]	0.73	0.59	0.28	**0.78**
NEW	PESQ-DNN（FLE），J（4），音视频△	**0.08**	**0.63**	0.31	0.75
NEW	PESQ-DNN（BLE），J(3)，AT□	**0.08**	0.62	0.29	0.72

表 IV：**已见**和**未见串联（TDM）条件**下的性能。在**测试集** D^{test} 上测得性能。最佳结果以**粗体**显示，次佳结果带下划线。

	方法	已见 TDM		未见 TDM	
		平均绝对误差	线性相关系数	平均绝对误差	线性相关系数
REF	QualityNet基线深度神经网络 [11]	0.18	0.48	0.28	0.41
REF	WaweNet 基线深度神经网络 [16]	0.14	0.61	**0.14**	0.66
REF	DNSMOS 基线深度神经网络 [19]	0.72	0.67	0.63	0.76
NEW	PESQ-DNN (FLE) J(4) 音视频△	**0.08**	0.79	0.21	**0.89**
NEW	PESQ-DNN (BLE) J(3)，AT□	**0.08**	**0.80**	0.26	0.85

高于 0.90。与此相反，对于 QualityNet，所有编解码模式的单独 LCC 值都高于相应的平均值。为分析这一现象，我们在图6中给出一幅散点图，展示采用 FLE 的最强 PESQ-DNN 变体（蓝色）和 QualityNet 基线（绿色）对 EVS 编码语音的预测 PESQ 分数 PESQ_u ，以及由国际电信联盟电信标准化部门 P.862.2 [6]测得的真实值 PESQ_u 。根据公式（8），我们看到，如果真实值 PESQ_u减去其平均值 μ 后的变化，能够由预测 PESQ 分数 PESQ_u 减去 $\hat{\mu}$后的相似数值变化反映出来，就会得到较高的LCC分数。因此，这体现为标记沿对角线分布。为更清晰地展示分布情况，我们分别用标记 + 和 ×绘出了各个 EVS 模式 $j \in \{7.2, 13.2, 48\}$ 千比特每秒下 $\hat{\mu}_j^{\text{QualityNet}}$ 和 $\hat{\mu}_j^{\text{PESQ-DNN}}$ 的平均值，如图6所示。对于 PESQ-DNN 和 QualityNet，各 EVS 模式对应标记的分布趋势都与对角线相似，这与表II中所有编解码模式的单独 LCC 值均高于 0.6 一致。同时，各 EVS 模式相应平均值之间的连线可反映整体分布。令人惊讶的是，QualityNet 各编解码模式的单独平均值（+）几乎呈水平分布，而所提出的 PESQ-DNN 变体对应的平均值（×）则几乎沿对角线散布。这解释了为何在 EVS 编码语音上测得的 QualityNet基线平均LCC远低于表II所示的各编解码模式单独 LCC 值。蓝色点显然分布得更靠近对角线，尽管存在一些离群点；绿色点则分布得更分散。这证实，与表II中的 QualityNet基线相比，我们提出的 PESQ-DNN 的平均绝对误差更低（0.09 对 0.25）。

2）噪声条件：我们在测试集 D^{test} 中的 13.2 kbps EVS 编码语音上，评估所有研究方法在未见噪声条件下的性能。我们分别报告在不含和含加性噪声的编码语音上的性能，在表 III 中分别记为纯净条件和噪声条件。由于我们研究的是加性噪声对 PESQ 估计的影响，因此在本实验中排除了 TDM 和易出错传输条件。

在基线深度神经网络中，结果呈现出不同的情况：QualityNet DNN 在无噪声条件下的平均绝对误差和线性相关系数表现最佳，但在噪声条件下，这两项指标的表现最差。另一方面，DNSMOS 深度神经网络表现最佳的是噪声条件下的表现；就平均绝对误差而言，它在噪声条件下（0.28）的表现甚至优于无噪声条件下（0.73）。然而，与所有其他模型相比，它在无噪声条件下的平均绝对误差表现要差得多。相比之下，我们提出的 PESQ-DNN 变体在无噪声条件下表现出色，在噪声条件下的表现也较为均衡，仅略逊于最佳结果。值得注意的是，采用 J(4) 和平均池化(音视频)的 FLE 变体表现最佳。

3) 串联条件：在表 IV 中，我们展示了所有研究模型在*TDM* 条件下的性能，其中两个宽带编解码器进行了串联。我们在此实验中排除了加性噪声和易出错传输的影响。在已见 TDM 条件下，两个串联编解码器的顺序与训练和开发阶段所采用的顺序相同，即先使用速率为 64 kbps 的 G.722，再使用速率为 13.2 kbps 的 EVS。在未见 TDM 条件下，我们调换两个编解码器的顺序，先使用速率为 13.2 kbps 的 EVS，再使用运行于 64 kbps 的 G.722。

在所有条件和指标下，DNSMOS 和 QualityNet基线的基线深度神经网络的表现均不如我们提出的两种 PESQ-DNN 变体。有趣的是，对于两个指标，WaweNet 基线深度神经网络在已见和未见 TDM 条件下都表现出良好的质量；在后者中，它甚至在平均绝对误差方面是最强的方法（0.14）。另一方面，我们提出的 PESQ-DNN 变体在已见 TDM 条件下（两个指标）以及未见 TDM 条件下的线性相关系数方面都明显表现最强。需要注意的是，未见 TDM 条件下的平均绝对误差表现较强（但并非最佳），尤其值得注意的是，我们提出的两种方法再次展现出均衡的性能。

4) 易出错传输条件：表 V 展示了我们对所有研究模型在易出错传输条件下帧丢失方面的性能评估。为更好地考察帧丢失对 PESQ 估计的影响，本实验排除了噪声条件和 TDM 传输条件。我们分别报告 EVS 编码语音在 13.2 和 48 千比特每秒比特率下、帧擦除率为 0%、3% 和 6% 时的性能，如表 V 所示。在所采用的帧擦除率中，6% 的帧擦除率在训练和开发阶段未曾出现，而 0% 的帧擦除率代表无帧错误的传输条件，因此其性能与表 II 所示的纯净条件下的性能相同。此外，每种模拟的帧擦除率都包含数量相同的随机帧擦除和突发帧擦除。为更全面地展示所有研究模型在易出错传输条件下的表现，我们仅针对 3% 和 6% 的帧擦除率，报告所有所用比特率的平均总体性能，如图所示

表 V：不同帧错误率（FER）下的**易出错条件**性能。性能是在测试集 D^{test} 的 13.2 千比特每秒和 48 千比特每秒 EVS 编码语音上测得的。最佳结果以**粗体**显示，次佳结果带下划线。

	方法	EVS @ 13.2 kbps						EVS @ 48 千比特每秒						总计（帧错误率 3%/6%）	
		MAE			LCC			MAE			LCC			MAE	LCC
	帧错误率：	0%	3%	6%	0%	3%	6%	0%	3%	6%	0%	3%	6%	-	-
REF	QualityNet基线深度神经网络 [11]	**0.08**	0.38	0.70	0.61	0.64	0.51	0.29	**0.23**	0.42	**0.67**	0.53	0.60	0.43	0.54
REF	WaweNet 基线深度神经网络 [16]	0.16	**0.27**	0.49	0.52	0.54	0.46	0.20	0.29	0.47	0.27	0.44	0.42	0.38	0.67
REF	DNSMOS 基线深度神经网络 [19]	0.73	0.39	**0.33**	0.59	0.57	0.42	0.76	0.35	**0.24**	0.21	0.56	0.41	**0.33**	0.61
NEW	PESQ-DNN (FLE) *J*(4), 音视频△	**0.08**	0.32	0.54	**0.63**	**0.67**	0.63	0.11	0.32	0.51	0.61	**0.76**	**0.80**	0.45	**0.75**
NEW	PESQ-DNN（BLE） *J*(3)，AT□	**0.08**	0.33	0.56	0.62	0.66	**0.66**	**0.09**	0.33	0.53	0.49	0.68	0.70	0.46	0.74

表 VI：**总体性能** 对所有测试条件（“total”）以及除易出错传输条件之外的所有条件（“totalwithout EID”）取平均。性能在**测试集** D^{test} 上进行测量。最佳结果以**粗体**显示，次佳结果加下划线。

	方法	总计		总计不含 EID	
		平均绝对误差	线性相关系数	平均绝对误差*	线性相关系数*
REF	QualityNet 基线深度神经网络 [11]	0.27	0.67	0.24	0.44
REF	WaweNet 基线深度神经网络 [16]	0.19	0.79	0.16	0.87
REF	DNSMOS 基线深度神经网络 [19]	0.63	0.78	0.67	0.88
NEW	PESQ-DNN (FLE) *J*(4)，音视频△	**0.16**	**0.84**	**0.11**	**0.92**
NEW	PESQ-DNN (BLE) *J*(3)，AT□	**0.16**	0.82	0.12	0.90

见表 V。

结果表明，易出错传输条件对所有研究模型而言都颇具挑战。对大多数模型来说，帧错误情况下的平均绝对误差远高于帧擦除率为 = 0% 时的结果。令人意外的是，DNSMOS 基线深度神经网络在易出错条件下的平均绝对误差低于帧擦除率为 = 0% 时的结果。这与表 III 中的观察相似：DNSMOS 网络在噪声条件下的平均绝对误差优于纯净条件下的结果。由此可见，它仅在恶劣条件下具有拓扑优势。请注意，尽管我们提出的方法在平均绝对误差方面略逊于基线方法，但在相关性（线性相关系数）方面明显领先：总体而言，我们提出的方法在线性相关系数排名中位列第 1st 和第 2nd，同时平均绝对误差表现也相当不错。

5) 总体性能：最后，我们在表 VI 中报告所研究方法的总体测试性能。这些测量结果基于由以不同比特率运行的 EVS 和 AWB 编解码器生成的编码语音信号计算得出，并涵盖表 II、III、IV 和 V 所示的所有测试条件。由于表 V 中所示的大多数研究模型在易出错条件下的性能存在一定局限性，我们还报告了排除易出错条件后的总体性能。

总的来说，综合所有条件下的数据（而非平均绝对误差或线性相关系数），我们提出的两种方法分别位列第 1st 和第 2nd。排除易出错传输条件下的数据后，我们的两种方法同样分别排名第 1st 和第 2nd。在这种情况下，表现最佳的方法“PESQ-DNN (FLE) *J*(4), 音视频△”取得了 0.92 的高相关性，平均绝对误差仅为 0.11，因此成为我们最终提出的基于深度神经网络的非侵入式 PESQ 预测方法。

六、结论

本研究提出了一种端到端非侵入式 PESQ-DNN，用于评估语音通信系统中的编码语音质量。我们通过非侵入式估计由 AMR-WB 或 EVS 编解码器在不同比特率下、于噪声、串联和易出错传输条件中生成的宽带编码语音的语音质量感知评估（PESQ）分数，展示了所提出的 PESQ-DNN 的潜力。我们通过测量平均绝对误差（MAE）和线性相关系数（LCC），将我们的方法与当前先进的 QualityNet、WaweNet 和用于 PESQ 估计的 DNSMOSDNN 网络进行了比较。作为本研究的核心贡献，我们开展了全面的消融研究，阐明所提出的 PESQ-DNN 中各项改进的影响和价值，包括神经网络的输入类型、PESQ-DNN 拓扑结构的变化（涉及不同的嵌入处理模型），以及训练过程中采用的损失函数。在所有测试条件下取平均值后，所提出的 PESQ-DNN 在不考虑帧丢失的情况下，取得了最佳的总体平均绝对误差和线性相关系数，分别为 0.11 和 0.92。即使存在帧丢失，其线性相关系数也优于所有基线方法。值得注意的是，我们的网络同样可用于非侵入式预测 POLQA 或其他侵入式指标。

参考文献

[1] 国际电信联盟，*64 kbit/s* 内的 *7 kHz* 音频编码，国际电信联盟电信标准化部门（ITU-T），2012年9月。[2] 3GPP，语音编解码器语音处理功能；自适应多速率宽带（*AMR-WB*）语音编解码器；转码功能（*3GPP TS 26.190*，第*15*版），3GPP，2018年6月。[3] B. Bessette、R. Salami、R. Lefebvre、M. Jelinek、J. Rotola Pukkila、J. Vainio、H. Mikkola 和 K. J¨arvinen，“自适应多速率宽带语音编解码器（AMR-WB）”，*IEEE T-ASLP*，第10卷第8期，第620–636页，2002年11月。[4] 3GPP，增强语音服务（*EVS*）编解码器；详细算法描述（*3GPP TS 26.445*，第*16*版），3GPP，2019年6月。[5] 国际电信联盟，建议 *P.800*：传输质量主观测定方法，国际电信联盟电信标准化部门（ITU-T），1996年8月。[6] ——，建议 *P.862.2*：勘误*1*，建议书 *P.862*关于宽带电话网络和语音编解码器评估的宽带扩展，国际电信联盟电信标准化部门（ITU-T），2017年10月。[7] ——，建议 *P.863*：感知客观听觉质量预测（*POLQA*），国际电信联盟，电信标准化部门（ITU-T），2018年3月。[8] A. Hines、J. Skoglund、A. C. Kokaram 和 N. Harte，“ViSQOL：客观语音质量模型”，*EURASIP* 音频、语音与音乐处理期刊，第2015卷第1期，第1–18页，2015年5月。

[9] J. Abel、M. Kaniewska、C. Guillaum´e、W. W. Tirry 和 T. Fingscheidt，“人工扩展带宽的语音信号的客观质量度量”，IEEE/ACM 音频、语音与语言处理汇刊，第25卷第2期，第384–396页，2016年2月。[10] M. H. Soni 和 H. A. Patil，“用于非侵入式语音质量评估的新型深度自编码器特征”，载于欧洲信号处理会议论文集，匈牙利布达佩斯，2016年12月，第2315–2319页。[11] S. W. Fu、Y. Tsao、H. T. Hwang 和 H. M. Wang，“Quality-Net：一种基于双向长短期记忆网络的端到端非侵入式语音质量评估模型”，载于 INTERSPEECH 论文集，印度海得拉巴，2018年9月，第1873–1877页。[12] A. H. Andersen、J. M. De Haan、Z. H. Tan 和 J. Jensen，“使用卷积神经网络进行非侵入式语音可懂度预测”，IEEE T-ASLP，第26卷第10期，第1925–1939页，2018年10月。[13] G. Mittag 和 S. M¨oller，“超宽带语音通信网络的非侵入式语音质量评估”，载于国际声学、语音与信号处理会议论文集，英国布莱顿，2019年5月，第7125–7129页。[14] A. R. Avila、H. Gamper、C. K. A. Reddy、R. Cutler、T. Tashev 和 J. Gehrke，“使用神经网络进行非侵入式语音质量评估”，载于国际声学、语音与信号处理会议论文集，英国布莱顿，2019年5月，第631–635页。[15] H. Gamper、C. K. A. Reddy、R. Cutler、I. Tashev 和 J. Gehrke，“使用卷积神经网络进行侵入式和非侵入式感知语音质量评估”，载于语音与音频处理及听觉会议论文集，美国纽约州纽帕尔茨，2019年10月，第85–89页。[16] A. A. Catellier 和 S. D. Voran，“Wawenets：一种基于无参考卷积波形的方法，用于估计窄带和宽带语音质量”，载于国际声学、语音与信号处理会议论文集，西班牙巴塞罗那，2020年5月，第331–335页。[17] X. Dong 和 D. S. Williamson，“用于真实环境中客观语音评估的注意力增强多任务模型”，载于国际声学、语音与信号处理会议论文集，西班牙巴塞罗那，2020年5月，第911–915页。[18] G. Mittag、B. Naderi、A. Chehadi 和 S. M¨oller，“NISQA：利用众包数据集进行多维语音质量预测的深度卷积神经网络自注意力模型”，载于 INTERSPEECH 论文集，捷克共和国布尔诺，2021年8月，第2127–2131页。[19] C. K. A. Reddy、V. Gopal 和 R. Cutler，“DNSMOS：用于评估噪声抑制器的非侵入式感知客观语音质量指标”，载于国际声学、语音与信号处理会议论文集，加拿大安大略省多伦多，2021年6月，第6493–6497页。[20] Z. Xu、M. Strake 和 T. Fingscheidt，“采用非侵入式 PESQNet 监督的深度噪声抑制，使真实训练数据得以使用”，载于 INTERSPEECH 论文集，捷克共和国布尔诺，2021年8月，第2806–2810页。[21] Z. Xu、M. Strake 和 T. Fingscheidt，“通过非侵入式 PESQNet 介导的最大化不可微的PESQ实现深度噪声抑制”，IEEE/ACM 音频、语音与语言处理汇刊，第30卷第4期，第1572–1585页，2022年4月。[22] C. K. A. Reddy、V. Gopal 和 R. Cutler，“DNSMOS P.835：用于评估噪声抑制器的非侵入式感知客观语音质量指标”，载于国际声学、语音与信号处理会议论文集，新加坡，2022年4月，第886–890页。[23] Z. Xu、M. Strake 和 T. Fingscheidt，“PESQNet（损失）是否需要干净参考输入？原始PESQ需要，但 ACR 听音测试不需要”，载于国际语音增强与噪声消除研讨会论文集，德国班贝格，2022年9月，第1–5页。[24] C. K. A. Reddy、H. Dubey、V. Gopal、R. Cutler、S. Braun、H. Gamper、R. Aichner 和 S. Srinivasan，“ICASSP 2021深度噪声抑制挑战赛”，载于国际声学、语音与信号处理会议论文集，加拿大安大略省多伦多，2021年6月，第6623–6627页。[25] C. K. A. Reddy、H. Dubey、K. Koishida、A. Nair、V. Gopal、R. Cutler、S. Braun、H. Gamper、R. Aichner 和 S. Srinivasan，“INTERSPEECH 2021深度噪声抑制挑战赛”，载于 INTERSPEECH 论文集，捷克共和国布尔诺，2021年8月，第2796–2800页。[26] 国际电信联盟，建议书P.808：采用众包方法进行语音质量主观评估，国际电信联盟电信标准化部门（ITU-T），2018年2月。[27] ——，建议书P.835：勘误1，包含噪声抑制算法的语音通信系统评估主观测试方法，国际电信联盟电信标准化部门（ITU-T），2011年1月。[28] P. Meyer、Z. Xu 和 T. Fingscheidt，“改进用于语音情感识别的卷积循环神经网络”，载于口语技术研讨会论文集，中国深圳，2021年1月，第356–372页。[29] 国际电信联盟，建议书P.565：用于创建和性能测试基于机器学习的模型的框架，以评估传输网络对移动分组交换语音语音质量的影响服务，国际电信联盟电信标准化部门（ITU-T），2021年11月。[30] ——，建议 *P.565.1*：用于评估移动分组交换语音业务语音质量的传输网络影响的机器学习模型，国际电信联盟电信标准化部门（ITU-T），2021年11月。[31] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“Attention Is All You Need”，载于*NIPS* 论文集，美国加利福尼亚州长滩，2017年12月，第1–11页。[32] “电话测量多语言语音数据库”，NTT先进技术株式会社（NTT-AT），日本川崎市，1994年。[33] 3GPP，*AMR-WB* 特性描述处理计划草案（*WB-7c*），*1.0*版，3GPP，2001年9月。[34] A. Ramo 和 H. Toukomaa，“不同窄带和宽带语音编解码器语音质量的比较”，载于国际信号处理应用研讨会论文集，澳大利亚新南威尔士州悉尼，2005年8月，第603–606页。[35] 3GPP，*EVS* 常设文件 *EVS-7c*：特性描述阶段处理功能，*V1.0.0*，3GPP，2014年8月。[36] Z. Zhao、H. J. Liu 和 T. Fingscheidt，“用于增强编码语音的卷积神经网络”，*IEEE/ACM* 音频、语音与语言处理汇刊，第27卷，第4期，第663–678页，2019年4月。[37] 国际电信联盟，语音和音频编码标准化软件工具，国际电信联盟电信标准化部门（ITU-T），2010年3月。[38] ——，建议 *P.56*：活动语音电平的客观测量，国际电信联盟电信标准化部门（ITU-T），2011年12月。[39] J. Thiemann、N. Ito 和 E. Vincent，“多样化环境多通道声学噪声数据库：多通道环境噪声录音数据库”，美国声学学会期刊，第133卷，第5期，第3591–3591页，2013年。[40] D. B. David、S. Sridharan、R. J. Vogt 和 M. W. Mason，“用于评估语音活动检测算法的 QUT-NOISE-TIMIT 语料库”，载于*INTERSPEECH* 论文集，日本幕张，2010年9月，第3110–3113页。[41] 3GPP，自适应多速率宽带（*AMR-WB*）语音编解码器的 *ANSI-C* 代码（*3GPP TS 26.173*，第*14*版），3GPP，2017年4月。[42] ——，增强语音服务（*EVS*）编解码器；*ANSI C* 代码（定点）（*3GPP TS 26.442*，第*16*版），3GPP，2021年2月。[43] M. Mushkin 和 I. Bar-David，“Gilbert-Elliott 信道的容量与编码”，*IEEE* 信息论汇刊，第35卷，第6期，第1277–1290页，1989年11月。[44] E. N. Gilbert，“突发噪声信道的容量”，贝尔系统技术期刊，第39卷，第5期，第1253–1265页，1960年9月。[45] “语音处理、传输和质量方面（STQ）；存在背景噪声时的语音质量性能；第1部分：背景噪声模拟技术与背景噪声数据库”，ETSI，法国索菲亚科技园，ETSI EG 202 396-1，1.2.2版，1994年9月。[46] S. W. Fu、C. F. Liao 和 Y. Tsao，“ 利用学习得到的损失函数进行学习：借助 Quality-Net 提升语音增强的语音质量感知评估”，*IEEE* 信号处理快报，第27卷，第11期，第26–30页，2019年11月。[47] A. L. Maas、A. Y. Hannun 和 A. Y. Ng，“整流器非线性提升神经网络声学模型性能”，载于*ICML* 论文集，第30卷，美国加利福尼亚州长滩，2013年6月，第1–6页。