arXiv:2603.02877v1 [eess.AS] 3 Mar 2026

DBMIF：一种用于气导与骨传导语音增强的深度平衡多模态迭代融合框架

YileiWu[1, 2†], Changyan Zheng[1, 2, 3†], Xingyu Zhang[1,2], Yakun Zhang[1,2], ChengshiZheng[4], Shuang Yang[5], Ye Yan[1, 2*], Erwei Yin[1,2]1军事科学院（AMS）国防创新研究院，北京，100071，中国。2军事科学院（AMS）智能博弈与决策实验室，北京，100071，中国。3高科技研究院，潍坊市范公亭南街12号，261000，中国。4中国科学院声学研究所，北京，100190，中国。5中国科学院计算技术研究所，北京，100190，中国。

*通讯作者。电子邮箱：yy taiic@163.com；贡献作者：wyl21w@126.com；echoaimaomao@163.com；zhangxingyu1994@126.com；ykzhang1222@126.com；cszheng@mail.ioa.ac.cn；shuang.yang@ict.ac.cn；yinerwei1985@gmail.com；

†这些作者对本研究的贡献相同。

摘要

在极低信噪比（SNR）环境中，空气传导（AC）麦克风会受到环境噪声的严重干扰，传统语音增强系统的性能因此急剧下降。尽管骨传导（BC）传感器能够提供互补且耐噪的信息，但现有融合方法难以在广泛的信噪比条件下保持稳定的性能。为解决这一局限，我们提出了深度平衡

多模态迭代融合框架（DBMIF），这是一种三分支架构，旨在通过严格的跨模态交互重建高保真语音。具体而言，该框架以多尺度交互式编码器–解码器骨干网络为基础，采用迭代注意力模块和跨分支门控模块，实现自适应加权与双向交换。为补充这种动态交互，框架还集成了平衡交互瓶颈，以学习紧凑、稳定的融合表示。大量实验表明，在多种噪声类型下，无论是语音质量还是可懂度，DBMIF 的表现均优于近期的单模态和多模态基线。在下游自动语音识别任务中，与竞争方法相比，所提方法至少将字符错误率降低了 2.5%。这些结果证实，DBMIF 能够有效利用骨传导语音的鲁棒性，同时保留空气传导语音的自然度，从而确保其在真实场景中的可靠性。源代码已公开发布于 https://github.com/wyl516w/dbmif。

关键词：语音增强、多模态融合、骨传导、跨模态交互、迭代融合

1 引言

语音增强旨在从噪声录音中恢复清晰自然的语音，是远程通信、可穿戴助听和人机交互的关键前端技术 [1–3]。借助深度学习，单通道系统在处理非平稳噪声以及适应不同说话人方面取得了显著进展，近期基准测试的迭代（如深度噪声抑制（DNS）挑战赛 [4]）也体现了这一点。然而，真实声学场景往往具有极低的信噪比条件和高度非平稳的噪声。在这些条件下，仅依赖空气传导麦克风的方法往往仍会残留噪声并引入语音失真，因为传感器直接暴露于环境干扰之中 [5]。

骨传导（BC）传感器通过颅骨振动采集语音，并能自然抑制空气传播噪声，因此在恶劣环境下具有很强的鲁棒性 [6]。然而，与空气传导语音相比，骨传导语音通常在约 $\sim$1 kHz 以上呈现明显的高频滚降，从而限制了其自然度和可懂度 [7, 8]。除了带宽有限之外，骨传导语音还会受到传播路径和换能器与皮肤接触的影响，导致频率相关的频谱着色和时间错位，而这些现象无法通过简单的增益调整来解释。这些特性使空气传导与骨传导语音融合并非易事：直接混合两种模态可能导致特征校准不当或失衡，尤其会影响对辅音清晰度至关重要的高频线索。以往研究尝试补偿缺失频带和传播路径引起的失真，方法包括频谱均衡 [9] 、带宽扩展，以及近期基于生成对抗网络（GAN）的 [10, 11] 恢复技术，该技术能够恢复更丰富的高频结构。尽管取得了这些进展，带宽损失与模态特定失真的综合影响仍然

在要求高感知质量时限制着骨传导语音，因此有必要采用学习式自适应融合机制，以校准跨模态表征。

为同时利用骨导语音的抗噪鲁棒性和空气传导语音的高频细节，多模态融合日益受到关注 [12]。研究人员已开发出将加速度计与骨传导麦克风集成的硬件系统，以实现鲁棒的语音活动检测，并为增强提供辅助语音参考信号 [13, 14]。在架构设计上，在时域全卷积网络中，先进行独立建模再融合的效果优于早期拼接 [15]。在复数谱域中，结合半监督训练的注意力机制能够提升低信噪比性能 [16]。DenGCAN 等轻量化设计将注意力机制嵌入跳跃连接，在效果与计算成本之间取得平衡 [17]然而，现有多模态方法在实际部署中常见的超低信噪比条件下仍然脆弱，难以在多种噪声类型下保持适应性。例如，在低信噪比条件下训练时，空气传导语音质量较差，而骨导通道的抗噪性相对较强，因此骨导语音分支往往更容易更新。结果，能立即带来最大损失下降的 BC 模态会主导梯度并抑制其他模态，导致表征坍缩为单一模态，最终降低整体性能。

这些问题要求构建一个能够维持模态间持续、平衡交换的框架。跨模态交互使各模态能够反复相互优化，从而保留并发挥其互补性。相比之下，单阶段融合通常只在一个集成节点处合并各模态，随后仅处理融合后的流，因此无法持续进行跨模态交换。与这种单阶段策略相比，迭代交互能够更有效地抑制任何单一模态占据主导，并促进信息丰富、抗噪的特征学习。

因此，为解决上述问题，我们提出深度平衡多模态迭代融合框架（DBMIF）。与传统的静态融合方法不同，DBMIF 建立了一种持续的跨模态交互机制，将多尺度编码与均衡驱动的再平衡相结合。这种设计使系统能够根据模态的即时可靠性动态调整，确保充分利用空调语音和骨传导语音的互补优势，同时不引入冗余参数。

本文的主要贡献总结如下：

1. 提出了一种用于稳健空气传导与骨传导语音融合的三分支多尺度交互式编码器–解码器架构。通过该框架，跨尺度特征一致性得以保持，从而有助于在真实噪声条件下有效利用两种模态之间的互补性。2. 开发了一种早期深度迭代注意力融合机制，以适应瞬时模态可靠性。具体而言，在恶劣噪声下对 BC 线索赋予更高注意力，而随着条件改善，注意力逐步转向空气传导高频线索。3. 引入了参数共享的平衡交互瓶颈，以递归地重新校准融合表征。因此，稳定性在较宽的

信噪比范围内得到提升，并在不增加参数数量的情况下缓解了模态主导。4. 在公开数据集上进行了广泛实验。所提出的方法被证明在语音质量和可懂度方面相较强基线取得一致增益，同时显著降低了下游自动语音识别（ASR）的字符错误率（CER）。

2 相关工作

2.1 气传导语音增强

基于深度学习的语音增强通常分为时频（T-F）域方法和时域方法 [1]。时频域方法通常利用短时傅里叶变换来估计频谱掩码，或直接回归干净频谱。代表性架构包括 CRN [18] 及其复数值变体 DCCRN [19], ，后者对幅度和相位进行联合建模，以缓解相位失配伪影。为进一步捕捉长程时间依赖关系，GaGNet 等先进模型采用注意力机制，将全局上下文建模与局部频谱细化相结合 [20]。此外，SEGAN 等时域模型 [21] 和 Demucs [22] 直接处理原始波形。虽然这有助于隐式地进行相位重建，但通常会带来显著的计算成本。近年来，人们还引入了基于自监督学习（SSL）的方法，如 HuBERT，通过在预训练特征空间中对表征进行正则化来提升泛化能力 [23]。尽管取得了这些进展，但在信噪比极低的条件下，仅依赖 AC 信号的单模态增强仍存在固有局限，此时目标语音会被非平稳噪声严重掩蔽。

2.2 骨传导语音增强

骨传导语音具有显著的高频衰减和非线性失真，这是声音经颅骨和软组织传播所致，因此其恢复颇具挑战。为减轻这些退化，早期研究探索了均衡 [9]和线性相位重建 [24] 等线性方法。然而，这些方法难以对复杂的频谱映射进行建模。因此，研究者引入了基于高斯混合模型的统计映射方法，以提高可懂度，但这类方法往往会产生过度平滑的频谱 [25, 26]。随着深度学习的兴起，研究者利用去噪自编码器的强大建模能力，同时实现噪声抑制和带宽扩展 [27, 28]。在此基础上，研究者开发了融入注意力机制的基于BLSTM的模型，以更好地利用时间依赖关系，从而更精细地恢复频谱细节 [29, 30]。然而，基于回归的深度模型往往会生成平均频谱。为克服这一局限，研究者探索了基于GAN的方案，以重建逼真的高频结构 [31, 32]。近期，为降低端到端映射的优化难度，研究者提出了顺序式两阶段流程，将去噪和带宽扩展的目标解耦 [33]。尽管取得了这些进展，BC 信号固有的带宽限制以及

传输引起的失真仍会限制恢复质量，尤其是在要求高感知保真度和自然度时。

2.3 多模态特征融合

空气传导语音包含丰富的高频细节，但容易受到噪声影响；骨传导语音则具有较强的抗环境干扰能力，但带宽有限。因此，多模态融合旨在整合两种模态的优势。早期研究采用输入级融合，将空气传导和骨传导特征拼接起来，并报告称在强噪声下性能有所提升 [34]。后续方法采用后期融合架构，先通过独立分支处理各个模态，再将其合并，使可穿戴系统能够同时进行骨传导带宽补偿和去噪 [6]。Yu 等人 [15] 指出，在融合之前对两种信号进行独立建模，在时域全卷积网络中的表现显著优于简单的输入拼接。在复数谱域中，Wang 等人 [16] 提出了一种基于注意力的融合模块，可根据噪声水平自适应地调整空气传导和骨传导信号的权重，从而在低信噪比条件下显著提高可懂度。近年来，DenG- CAN 等轻量化设计将注意力机制融入跳跃连接，在保持有竞争力的感知质量的同时，支持实时移动端部署 [17]。与此同时，利用自监督学习（SSL）的多模态增强方法日益受到关注。例如，视听语音增强利用自监督预训练编码器，降低了对大型配对数据集的依赖 [35]。类似地，近期的空气传导-骨传导语音融合研究强调将模态专属增强与噪声自适应融合相结合，这与可靠性感知集成的发展趋势一致 [36]。

尽管取得了这些进展，但在现实应用中常见的极低信噪比条件下，性能往往会下降。具体而言，现有方法通常缺乏细粒度的可靠性感知融合和稳定的跨模态交互。此外，平衡表征学习往往被忽视，导致一种模态在优化过程中占据主导地位。

3 方法

3.1 问题表述

令$\mathbf{x}_a \in \mathbb{R}^N$ 和$\mathbf{x}_b \in \mathbb{R}^N$ 分别表示含噪气导语音和对应的骨传导语音信号，其中 N 表示信号长度。令$\mathbf{y} \in \mathbb{R}^N$ 表示干净目标语音波形。直接进行时域处理通常会对全带信号一视同仁，往往忽略语音和噪声的频率相关特性。为此，我们采用伪正交镜像滤波器组（PQMF） [37] 将输入信号分解为子带。

给定离散时间输入模态 $c \in \{a, b\}$，第 m 个子带信号 $\mathbf{s}_{c,m}$ 通过信号 $\mathbf{x}_c$ 与分析滤波器 h_m 进行卷积运算得到：

$$\mathbf{s}_{c,m}[n] = (\mathbf{x}_c * h_m)[n], \quad m \in \{0, \dots, M-1\}, \tag{1}$$

其中，$*$ 表示卷积运算。分析滤波器 h_m 和相应的合成滤波器 g_m 由长度为 L 的原型低通滤波器 $p[n]$ 推导得到，该滤波器采用凯撒窗加权 sinc 函数设计：

$$\begin{aligned} h_m[n] &= 2p[L-1-n]\cos\left[\frac{(2m+1)\pi}{2M}\left(n-\frac{L-1}{2}\right)+(-1)^m\frac{\pi}{4}\right], \\ g_m[n] &= 2Mp[n]\cos\left[\frac{(2m+1)\pi}{2M}\left(n-\frac{L-1}{2}\right)-(-1)^m\frac{\pi}{4}\right], \end{aligned} \tag{2}$$

其中，$0 \le n \le L-1$ 是抽头索引，M 是子带数量。

基于分解后的子带特征，我们提出了一种深度神经网络，将其表示为参数化映射函数 $f_{\boldsymbol{\theta}}$，用于估计干净子带分量。这里，$\boldsymbol{\theta}$ 表示可学习参数的集合。该网络以两种模态的拼接子带作为输入，以利用全频带上下文信息。令$\mathbf{S}_c = [\mathbf{s}_{c,\ 0},\ \ldots,\ \mathbf{s}_{c,\ M-1}]$ 表示模态 $\hat{}c$的子带集合。估计得到的干净子带集合$\mathbf{S}$ 如下：

$$\hat{\mathbf{S}} = f_{\boldsymbol{\theta}}(\mathbf{S}_a, \mathbf{S}_b). \tag{3}$$

随后通过 PQMF 合成重建增强后的时域波形 $\hat{}\mathbf{y}$：

$$\hat{\mathbf{y}}[n] = \sum_{m=0}^{M-1} (\hat{\mathbf{s}}_m * g_m)[n], \tag{4}$$

$\hat{}$其中，$\hat{}\mathbf{s}_m$ 是$\mathbf{S}$中的第 m 个估计子带。最后，通过最小化估计语音与真实值之间的损失函数 $\mathcal{L}$来训练网络：

$$\min_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}). \tag{5}$$

在实验中，我们使用 PQMF 滤波器组将输入信号分解为 $M = 4$ 个子带，原型滤波器的长度为 $L = 64$。

3.2 模型概述

如图1所示，DBMIF 由两个主要组件构成：生成器和多尺度判别器。

生成器负责执行核心增强任务。首先，它使用 PQMF 分析滤波器组将输入的时域语音分解为四个子带，配置遵循 [11] 中的建议。随后对这些子带信号进行处理，生成增强后的估计信号，再通过 PQMF 合成将其重建为全频带波形。

与生成器相辅相成，判别器采用多分辨率架构。它不仅检查重建的全频带波形，以确保整体一致性，还会检查各个子带信号，以实现细粒度的频谱保真度。

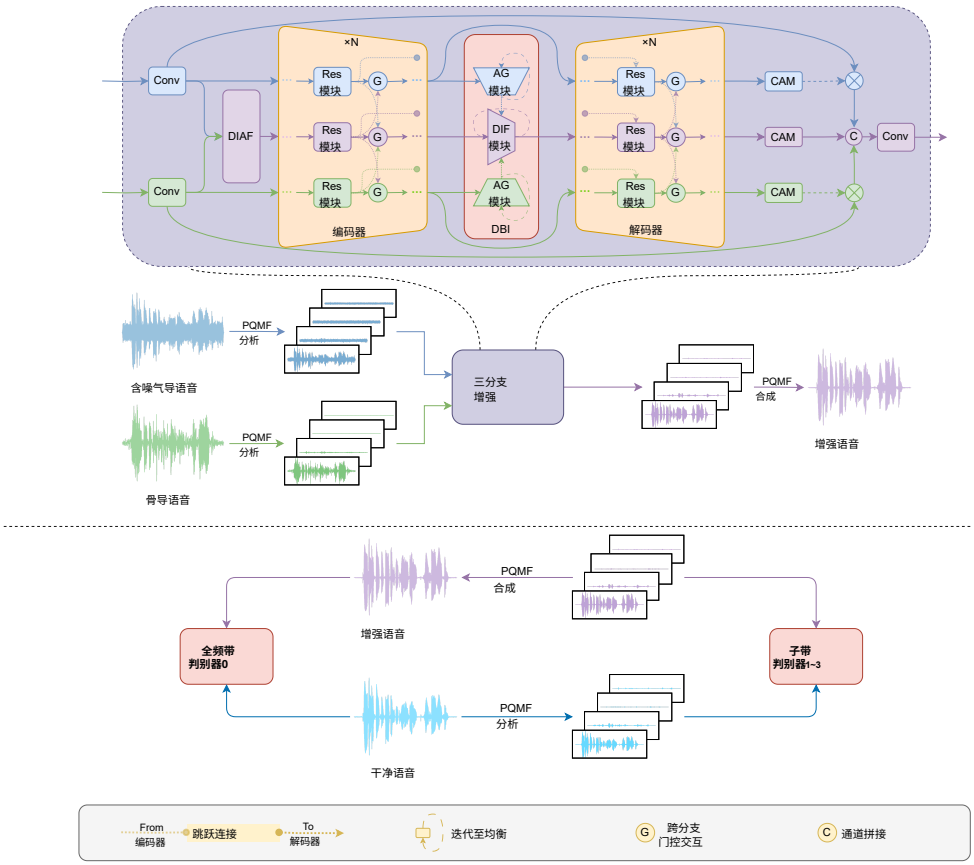


图 1：DBMIF 的整体架构。

以下各节将详细介绍生成器和判别器的具体架构，以及用于优化该框架的损失函数和训练策略。

3.3 生成器

3.3.1 三分支多尺度交互编码器-解码器

我们设计了一种三分支架构，由独立的空调与骨导路径以及中央融合分支组成。该主干在多个层次尺度上运行，以捕获多样的频谱上下文。具体而言，我们在相继各层采用渐进的通道宽度 $\{32,\ 64,\ 128,\ 256\}$ ，确保网络既能捕获细粒度细节，也能捕获高层语义抽象。两个单模态分支用于保留各模态的判别性特征，而融合流促进持续的信息交换，为网络提供均衡且互补的引导。

该架构的核心是跨分支门控交互（CBGI）模块，它实现融合分支与单模态分支之间的双向调制，如图2所示。在第 s层，令$\mathbf{x}_a^{(s)}$，$\mathbf{x}_b^{(s)}$，$\mathbf{x}_f^{(s)}$ 分别表示空调、骨导和融合分支的特征表征。CBGI机制分两个不同阶段运行：

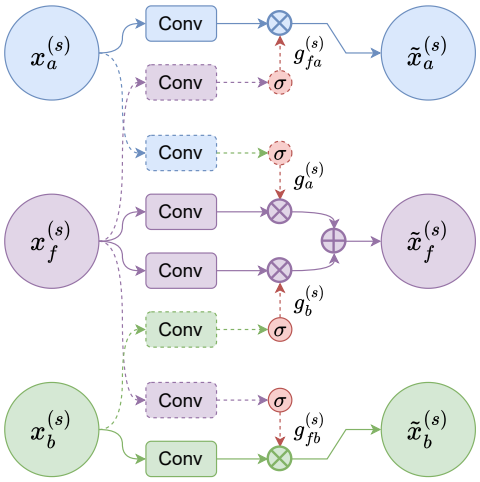


图 2：CBGI模块示意图。实线箭头表示特征流，虚线箭头表示生成的门控信号，用于调制相应的流。

1. 融合到单模态调制。首先，融合分支生成门控信号，以重新校准单模态特征。这使得来自融合流的全局上下文能够抑制单模态分支中的无关成分：

$$\begin{aligned} g_{fa}^{(s)} &= \sigma\big(\mathcal{F}_{\text{conv}}(\mathbf{x}_f^{(s)})\big), \qquad \tilde{\mathbf{x}}_a^{(s)} = g_{fa}^{(s)} \odot \mathcal{F}_{\text{conv}}(\mathbf{x}_a^{(s)}), \\ g_{fb}^{(s)} &= \sigma\big(\mathcal{F}_{\text{conv}}(\mathbf{x}_f^{(s)})\big), \qquad \tilde{\mathbf{x}}_b^{(s)} = g_{fb}^{(s)} \odot \mathcal{F}_{\text{conv}}(\mathbf{x}_b^{(s)}), \end{aligned} \tag{6}$$

其中，$\mathcal{F}_{\text{conv}}$ 表示一个独立的 1×1 卷积层，$\sigma(\cdot)$为Sigmoid 函数，$\odot$表示逐元素乘法。 $\tilde{\mathbf{x}}_a^{(s)}$和 $\tilde{\mathbf{x}}_b^{(s)}$ 为精炼后的单模态特征。

2. 单模态到融合反馈。反之，为了将可靠的模态特定细节融入融合流，每个单模态分支都会根据其原始特征生成反馈门控。这些门控用于调节融合表示的整合：

$$\begin{aligned} g_a^{(s)} &= \sigma\big(\mathcal{F}_{\text{conv}}(\mathbf{x}_a^{(s)})\big), \qquad \hat{\mathbf{x}}_{af}^{(s)} = g_a^{(s)} \odot \mathcal{F}_{\text{conv}}(\mathbf{x}_f^{(s)}), \\ g_b^{(s)} &= \sigma\big(\mathcal{F}_{\text{conv}}(\mathbf{x}_b^{(s)})\big), \qquad \hat{\mathbf{x}}_{bf}^{(s)} = g_b^{(s)} \odot \mathcal{F}_{\text{conv}}(\mathbf{x}_f^{(s)}). \end{aligned} \tag{7}$$

更新后的融合表示 $\tilde{\mathbf{x}}_f^{(s)}$ 通过聚合这些门控贡献得到：

$$\tilde{\mathbf{x}}_f^{(s)} = \hat{\mathbf{x}}_{af}^{(s)} + \hat{\mathbf{x}}_{bf}^{(s)}. \tag{8}$$

与传统的单向融合不同，这种双向耦合建立了闭环信息流。通过允许融合特征对单模态分支进行门控，反之亦然，该机制确保网络能够动态平衡两种模态的贡献。这避免了单一模态占据主导地位，并促进从两条流中选择性整合高保真组件。

3.3.2 早期深度迭代注意力融合

为后续处理提供更干净、更稳定的输入，我们在前端采用深度迭代注意力融合（DIAF）模块。该模块首先进行粗融合，再通过注意力机制逐步细化。在每一步中，通道注意力模块（CAM）都会生成通道级门控，自适应地平衡各模态的贡献，减弱可靠性较低的模态线索，同时突出更具信息量的线索。

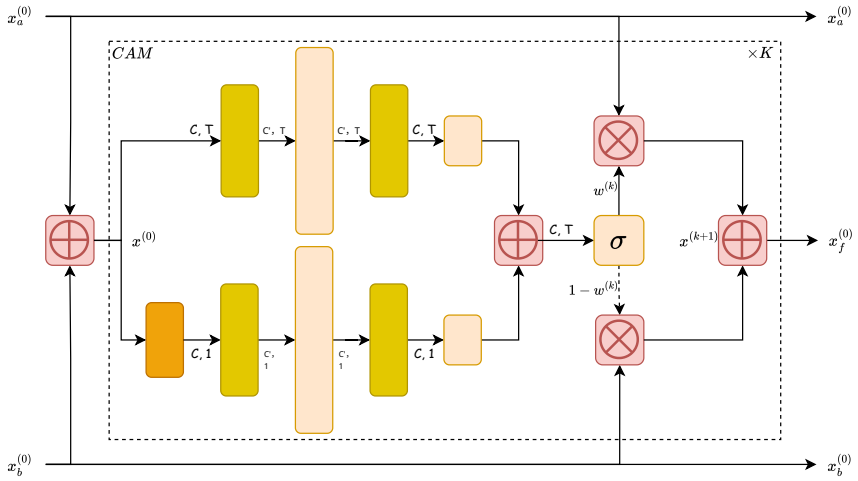


图 3：早期 DIAF 模块示意图。DIAF 通过迭代通道注意力逐步细化空气传导和骨传导特征，以突出可靠的模态线索。

令 $\mathbf{x}_a$ 和$\mathbf{x}_b$ 分别表示空调和骨传导语音对应的输入特征图。如图 3所示，DIAF 以迭代方式运行。首先，为了建立一个中性的起点且不引入额外参数，我们将粗融合 $\mathbf{x}^{(0)}$ 初始化为这些输入的逐元素求和：

$$\mathbf{x}^{(0)} = \mathbf{x}_a^{(0)} + \mathbf{x}_b^{(0)}. \tag{9}$$

随后，对于每次迭代 $k = 1, \ldots\ldots, K$，CAM 根据先前的融合表示生成注意力掩码：

$$\mathbf{w}^{(k)} = \mathrm{CAM}\big(\mathbf{x}^{(k-1)}\big). \tag{10}$$

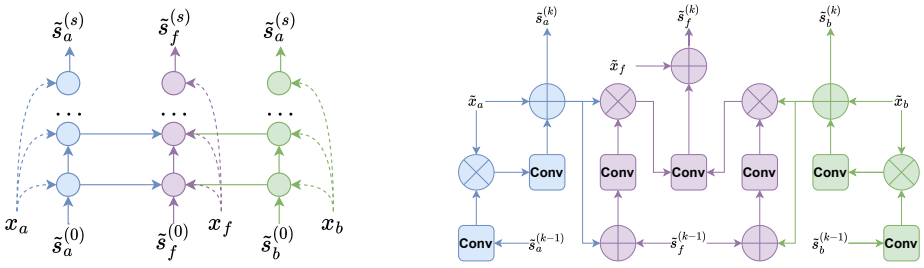
这里，$\mathbf{w}^{(k)} \in \mathbb{R}^C$ 作为逐通道可靠性门控，在全局平均池化后经过 Sigmoid 非线性处理，其取值被限制在$[0, 1]$ 范围内。从数学上看，$\mathbf{w}^{(k)}$ 的每个元素表示对应通道中 AC 模态的置信度，而 $1 - \mathbf{w}^{(k)}$ 则为骨导模态赋权。随后，通过软选择机制应用此掩码来细化融合：

$$\mathbf{x}^{(k)} = \mathbf{w}^{(k)} \odot \mathbf{x}_a^{(0)} + \big(1 - \mathbf{w}^{(k)}\big) \odot \mathbf{x}_b^{(0)}. \tag{11}$$

经过 K 轮细化后，得到最终的质量感知融合嵌入$\mathbf{x}_f^{(0)} = \mathbf{x}^{(K)}$。随着噪声特征变化，这一迭代过程使网络能够动态重新校准模态贡献。在我们的实现中，我们采用 $K = 3$，经实验发现，该方案能够在融合质量与计算成本之间实现稳健的平衡。

3.3.3 深度平衡交互

编码器-解码器瓶颈聚合紧凑且高信息密度的特征，因此对下游处理至关重要。为增强其有效性，我们采用深度平衡交互（DBI）模块。该模块借鉴了深度平衡多模态融合框架 [38]，DBI 通过迭代不动点更新细化模态特定特征。这种平衡机制能够稳定表征并提升泛化能力，同时不会增加网络深度或参数数量。通过在各次迭代间复用参数，DBI 对空调和骨导语音特征进行递归去噪和重新校准，直至其达到平衡状态，从而生成一致的表征。



(a) DBI 总体结构　　(b) 一次 DBI 迭代

图 4：DBI 模块示意图。DBI 维护循环模态状态，并通过模态内细化和融合反馈对其进行迭代更新。

如图4a所示，DBI 模块作用于编码器输出的瓶颈特征，记为 $\mathbf{x}_a$、$\mathbf{x}_b$、$\mathbf{x}f$。这些特征作为时不变输入注入。同时，该模块维护三个循环隐藏状态，记为 $\mathbf{z}_a^{(k)}$、$\mathbf{z}_b^{(k)}$ 、$\mathbf{z}_f^{(k)}$ ，对应第 k 次迭代。图 4b 展示了映射函数 $\mathcal{F}$ 的一次更新步骤，该步骤分为两个阶段：*1*）模态内细化：状态 $\mathbf{z}_a$ 和 $\mathbf{z}_b$ 在相应输入 $\mathbf{x}_a$ 和 $\mathbf{x}_b$ 的引导下进行更新，以抑制噪声，同时保留模态特定线索。*2*）融合反馈：细化后的单模态状态与先前的融合状态 $\mathbf{z}_f^{(k-1)}$ 在 $\mathbf{x}f$ 的引导下进行整合，以生成新的融合状态 $\mathbf{z}_f^{(k)}$ 。形式化地，联合更新规则定义如下：

$$\left(\mathbf{z}_a^{(k)}, \mathbf{z}_b^{(k)}, \mathbf{z}_f^{(k)}\right) = \mathcal{F}\left(\mathbf{z}_a^{(k-1)}, \mathbf{z}_b^{(k-1)}, \mathbf{z}_f^{(k-1)} \;\middle|\; \mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_f\right). \tag{12}$$

为简化表示，我们将各状态拼接为一个向量 $\mathbf{z}^{(k)} = \left[\mathbf{z}_a^{(k)}, \mathbf{z}_b^{(k)}, \mathbf{z}_f^{(k)}\right]$。我们的目标是重复应用 $\mathcal{F}$，使序列收敛到满足以下条件的不动点 $\mathbf{z}^\star$：

$$\mathbf{z}^\star = \mathcal{F}(\mathbf{z}^\star). \tag{13}$$

为提高训练稳定性，我们合并批次维度和时间维度，并在平衡过程中将每一帧视为一个独立样本。这使 DBI 能够适应模态主导的瞬时变化。

为确保计算效率，我们将最大迭代次数限制为$K_{\max} = 50$。我们通过计算连续两次迭代之间的相对变化来监控收敛：

$$\delta_k = \left\|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\right\|_F, \qquad k = 2, \ldots, K_{\max}. \tag{14}$$

我们不直接取最后一次迭代的结果，而是选择变化最小的状态 $z^\star$ ，因为这表明它对应最稳定的平衡点，并将其作为最终输出：

$$k^\star = \arg\min_{2 \le k \le K_{\max}} \delta_k, \qquad \mathbf{z}^\star = \mathbf{z}^{(k^\star)}. \tag{15}$$

由于 $\mathcal{F}$ 的参数在所有迭代中共享，DBI 能够以较高的参数效率实现深度特征细化和动态模态平衡。

3.4 判别器架构

为确保生成的波形具有高保真度，我们采用了受 MelGAN [39] 启发的多尺度判别策略。然而，与仅对波形进行处理的传统多尺度判别器不同，我们的设计明确利用子带分解来捕捉精细的频谱细节。具体而言，所提出的判别器由两个不同的部分组成：

1. 全频带波形判别器（$D_{\mathbf{wav}}$）： 该组件以原始采样率处理原始全频带音频，重点捕捉全局

结构一致性和低频一致性。如表1所示，它由一组具有较大感受野的步幅卷积层组成。

2. **多尺度子带判别器（$D_{\mathbf{sub}}$）：**我们采用与生成器相同的多相正交镜像滤波器组分析滤波器组，将真实或生成的语音分解为 4 个子带。这些子带沿通道维度拼接后，输入三个结构相同的判别器子网络，具体如表 2 所示。这些子网络仅在膨胀系数 $d \in \{1, 2, 3\}$上有所不同。这种膨胀卷积设计使模型能够在子带中通过不同的感受野捕捉多种时间结构。

表1：全频带判别器的结构D_{wav}。其输出为按时间索引的逻辑值图，用于全局一致性评估。

层	C_{in}	C_{out}	卷积核	步幅	分组	Act.
1	1	16	15	1	1	Leaky ReLU
2	16	64	41	4	4	Leaky ReLU
3	64	256	41	4	4	Leaky ReLU
4	256	1024	41	4	4	Leaky ReLU
5	1024	1024	41	4	4	Leaky ReLU
6	1024	1024	5	1	1	Leaky ReLU
7	1024	1	3	1	1	None

表2：子带判别器的架构 D_{sub}。实例化三个副本，分别采用膨胀系数 $d \in \{1, 2, 3\}$，以捕捉 PQMF 子带中的多尺度周期性。

层	C_{in}	C_{out}	卷积核	步幅	分组	膨胀率	Act.
1	4	36	3	1	4	d	Leaky ReLU
2	36	72	7	2	4	d	Leaky ReLU
3	72	144	7	2	4	d	Leaky ReLU
4	144	288	7	2	4	d	Leaky ReLU
5	288	576	7	2	4	d	Leaky ReLU
6	576	1152	7	2	4	d	Leaky ReLU
7	1152	1152	5	1	4	d	Leaky ReLU
8	1152	1	3	1	1	d	None

需要注意的是，两个判别器中的所有卷积层均采用权重归一化，并使用带泄漏的线性整流单元激活（$\alpha = 0.1$）；最终层除外，该层输出线性评分图。

3.5 训练目标

所提出框架的优化遵循对抗学习范式，涉及多尺度判别器与生成器之间的极小极大博弈。训练过程由两个彼此不同但相互补充的目标共同驱动：判别器旨在构建自然波形与合成波形之间稳健的分类边界，而生成器则致力于合成感知上真实的语音，从而成功欺骗判别器。在以下小节中，我们将分别详细介绍判别器和生成器的优化目标。

3.5.1 判别器损失

我们采用合页损失形式对判别器进行对抗训练，因为与标准交叉熵目标相比，它能提供更稳健的梯度信号。总判别器损失 L_D 汇总了全频带判别器和三个子带判别器（索引为 $k = 0$、.....、3）的贡献。这种多分辨率评估方案使模型能够同时监督全局频谱包络和细粒度局部纹理。令 D_k(**x**)表示输入**x**时第 k个判别器的输出评分图。判别器通过最小化以下目标，学习区分真实语音$\mathbf{y}$ 与合成语音$\hat{}\mathbf{y}$ ：

$$L_D = \sum_{k=0}^{3} \left[\mathbb{E}_{\mathbf{y}} \left[\max(0, 1 - D_k(\mathbf{y})) \right] + \mathbb{E}_{\hat{\mathbf{y}}} \left[\max(0, 1 + D_k(\hat{\mathbf{y}})) \right] \right]. \tag{16}$$

3.5.2 生成器损失

生成器旨在最小化一个复合目标，以平衡感知真实感与结构保真度。为促使合成出逼真的波形结构，我们首先使用合页损失定义对抗分量 L_G^{adv}。通过鼓励生成样本的评分超过单位间隔，生成器会因判别器检测到的任何伪影而受到惩罚：

$$L_G^{\text{adv}} = \sum_{k=0}^{3} \mathbb{E}_{\hat{\mathbf{y}}} \left[\max(0, 1 - D_k(\hat{\mathbf{y}})) \right]. \tag{17}$$

为补充对抗目标并缓解基于GAN的合成中常见的过度平滑问题，我们引入特征匹配损失 L_G^{fm}。该项通过最小化真实输入与生成输入在判别器中间特征图之间的 L_1 距离来稳定训练，并有效地充当一种学习式感知度量：

$$L_G^{\text{fm}} = \sum_{k=0}^{3} \sum_{i=1}^{N_k - 1} \frac{1}{M_i} \left\| D_k^{(i)}(\mathbf{y}) - D_k^{(i)}(\hat{\mathbf{y}}) \right\|_1, \tag{18}$$

其中， $D_k^{(i)}$ 表示从第 i 层的第 k 个判别器中提取的特征图，而 M_i 则作为归一化因子，表示该层中的元素数量。

因此，生成器的总训练目标被表述为这两个项的加权组合：

$$L_G = L_G^{\text{adv}} + \lambda L_G^{\text{fm}}, \tag{19}$$

其中， λ 是一个超参数，用于控制特征匹配监督的相对重要性。

4 实验设置

4.1 数据集

我们使用空气传导与骨传导同步语音语料库 [40]作为数据集。该语料库包含 100 位说话人的时间对齐空气传导和骨传导语音录音，共计 47,182 条语句，总时长为 42 小时，单条时长为 1 至 5 秒。所有音频录音均以 16 kHz 采样。

训练期间，我们采用在线切片与噪声混合流程，将每条语句切分为 1 秒的窗口。我们仅向空气传导语音通道注入噪声，噪声样本取自 DNS 挑战赛 [41] 和 UTD 环境噪声 [42]数据集。这些噪声源涵盖机械噪声、车内环境噪声和多人谈话噪声等多种类别。信噪比从 $[-15, 5]$ dB 的范围内均匀采样。注入噪声后，我们重新缩放混合信号，使其均方根能量与原始干净语音相匹配。最后，我们对片段边界应用 50 ms 线性淡入淡出，以减轻切片造成的不连续性。

为进行评估，我们使用三个未见噪声语料库，以确保多样化的声学条件。这些语料库包括用于模拟咖啡馆和汽车等真实场景的 QUT-NOISE [43] 、用于标准化工业和军事基准测试的 NOISEX-92 [44]，以及用于日常环境噪声的 Non-speech [45]。与训练阶段不同，我们对测试语句以完整长度进行处理，不进行分段。每条测试语句都在固定条件下与这些噪声源进行混合，其中信噪比从集合 (-15, -10, -5, 0, 5) dB 中选取。

4.2 对比基线

为全面评估DBMIF的有效性，我们以表3中汇总的代表性基线为参照，对其进行基准评测。这些基线主要由多模态方法构成，包括FCN等时域模型 [15] 以及MMINet [46]，还有DenGCAN等时频域模型 [17] 和Dc–Crn [16]。此外，为验证多模态融合的优势，我们还纳入了单模态基线，具体为GaGNet [20] （用于气传导语音增强）以及EBEN [11] （用于骨传导语音增强）。

表3：比较模型的详细信息。

模型	模态	领域	基于GAN
GaGNet [20]	AC	时频	No
EBEN [11]	BC	Time	Yes
FCN [15]	AC&BC	Time	No
MMINet [46]	AC&BC	Time	No
DenGCAN [17]	AC&BC	时频	No
Dc–Crn [16]	AC&BC	时频	No
DBMIF（本文）	AC&BC	Time	**Yes**

4.3 配置与实现细节

模型使用同步的1秒空气传导语音与骨传导语音对进行训练，批大小为32，共训练100个轮次。优化采用Adam优化器，学习率为 3×10^{-4}，动量参数为$(\beta_1, \beta_2) = (0.5,0.9)$。训练过程中采用余弦退火调度，使学习率逐渐衰减。在对抗学习期间，生成器与判别器交替更新，特征匹配损失的权重为 $\lambda = 1000$。

4.4 评估指标

性能采用五项广泛使用的客观指标进行评估，包括尺度不变信号失真比（SI-SDR）、语音质量感知评价（PESQ） [47], 短时客观可懂度（STOI） [48], 扩展短时客观可懂度（ESTOI） [49], 以及 DNSMOS [41]。与前四项需要干净参考信号的侵入式指标不同，DNSMOS 是一种非侵入式深度神经网络模型，旨在直接根据含噪或增强信号估计语音的感知质量。研究表明，它与人类平均意见得分具有良好相关性，因而适合在无需主观听音测试的情况下进行大规模评估。

为进一步评估增强语音的实际收益，我们开展了下游自动语音识别实验。字符错误率基于两种具有代表性的端到端自动语音识别框架进行报告，即 Whisper [50] 与 WeNet [51]。Whisper 具备强大的多语言及噪声鲁棒识别能力，而 WeNet 是一种轻量级开源流式自动语音识别框架，针对特定语言场景下的效率进行了优化。

5 实验结果

我们对所提出的 DBMIF 框架进行了全面评估，以验证其有效性和鲁棒性。首先，我们将 DBMIF 与具有代表性的单模态和多模态基线进行比较，并通过客观指标、下游自动语音识别准确率和主观听音测试分析其性能。为评估其泛化能力，我们进一步考察了模型在不同信噪比条件和多种噪声类别下的性能一致性，并辅以详细的频谱分析。最后，我们进行了系统的消融实验，以验证关键架构组件的贡献，并可视化模态融合机制的自适应行为。

5.1 与基线的比较

表 4 汇总了所有模型的定量性能。未经处理的空气传导语音受到严重的噪声退化影响，信号完整性较差，尺度不变信号失真比为 -4.96 dB，Whisper 自动语音识别错误率高达 47.79%。单模态基线表现出各自的局限性。具体而言，基于空气传导语音的 GaGNet 在语音质量感知评估方面仅取得微弱提升，而基于骨导的 EBEN 受限于带宽，尺度不变信号失真比极低。

5.1.1 总体定量比较

表4：比较　　客观指标在原始信号与　　增强语音。

模型	尺度不变信噪比失真比	PESQ	短时客观可懂度	扩展短时客观可懂度	DNSMOS	Whisper	WeNet
Origin							
空气传导语音	-4.96	1.13	0.68	0.50	2.39	47.79	34.62
骨传导语音	-7.53	1.43	0.69	0.58	2.48	29.30	25.44
AC-only							
GaGNet [20]	7.84	1.55	0.77	0.63	2.69	40.17	26.17
BC-only							
EBEN [11]	0.80	1.84	0.79	0.70	3.17	43.86	34.72
Joint Model							
FCN [15]	8.16	1.43	0.77	0.66	2.39	50.83	29.65
MMINet [46]	10.45	1.77	0.83	0.73	2.69	32.58	24.65
DenGCAN [17]	10.75	1.98	0.88	0.81	2.94	20.65	14.80
Dc–Crn [16]	10.18	2.62	0.86	0.79	2.66	19.46	14.50
DBMIF	12.42	2.70	0.90	0.85	3.35	15.84	12.07

总体而言，与原始含噪输入相比，基于融合的基线方法取得了更好的信号质量指标，这验证了多模态方法的有效性。然而，不同架构之间存在明显的性能差距。FCN 和 MMINet 等时域模型尽管在其他指标上有所提升，但在自动语音识别任务上的表现不佳。这种差异可能是因为全频带时域卷积会引入结构伪影，或过度平滑高频细节，从而破坏机器识别所需的声学线索。相比之下，DenGCAN 和 Dc-CRN 等时频方法在有效性和可懂度方面表现得更加均衡，但与干净语音相比仍存在不可忽视的差距。最后，DBMIF 在所有指标上均取得了最先进的结果。这一优异表现可能归功于基于PQMF的子带策略，以及空调和骨传导语音分支之间精巧的交互机制；二者共同减轻了传统全频带时域建模中常见的伪影。

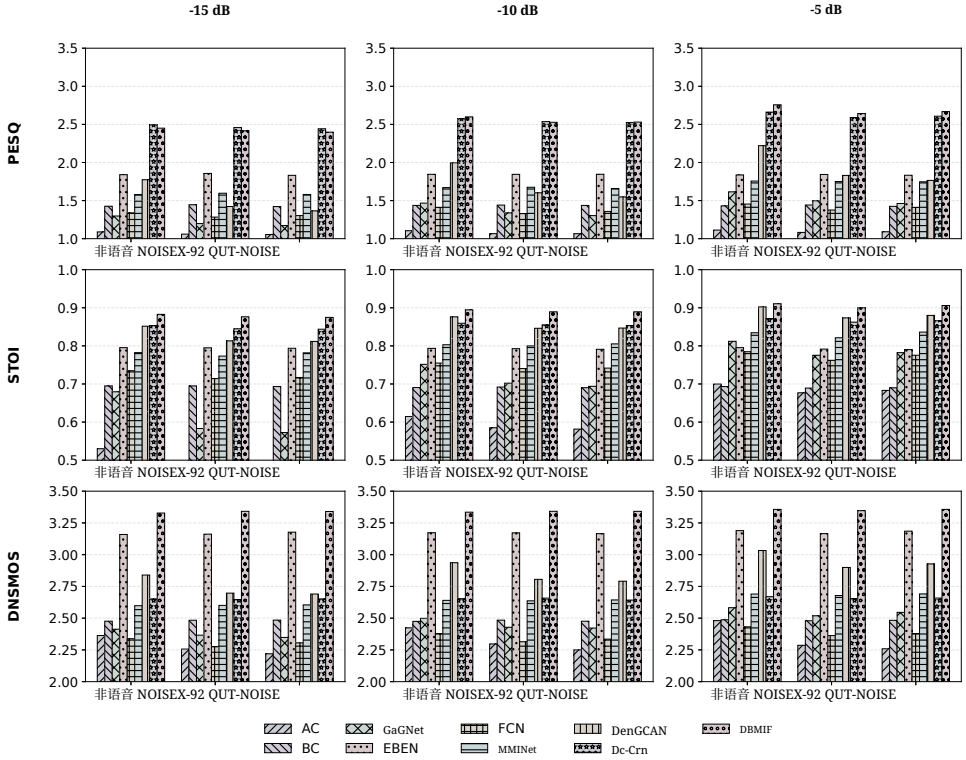


图 5：不同方法在低信噪比下的结果。

5.1.2 不同信噪比区间下的性能一致性

我们在 -15、-10 和 -5 dB 的低信噪比条件下，针对多种噪声语料库比较了各方法。如图 5 所示，GaGNet 在低信噪比下性能严重退化，且在不同噪声类型下波动较大。DenGCAN 的表现也不理想，这可能是因为其处理复杂非平稳噪声的能力有限。相比之下，其余多模态方法在不同数据集上表现得更加稳健，表明引入骨传导语音能够减轻在未见噪声下的性能退化。此外，EBEN 和 DBMIF 均采用生成对抗网络训练，并取得了较高的 DNSMOS 分数，这说明对抗学习有助于提升感知质量和自然度。总体而言，DBMIF 在低信噪比条件下的各项指标均取得稳定提升，且通常在不同噪声条件下位居表现最佳的方法之列。我们还观察到少数极端情况，在某些感知指标上，DBMIF 的优势有所缩小。例如，在 -15 dB 的强非平稳 babble噪声下，仅使用骨传导的基线方法 EBEN 取得了与 DBMIF 相当的 DNSMOS 分数。这归因于BC传感器固有的物理特性：它们产生的信号没有噪声，但带宽有限。相比之下，DBMIF 优先恢复高频细节。因此，即使需要保留少量残余噪声，DBMIF 仍能生成自然度显著更高的语音。

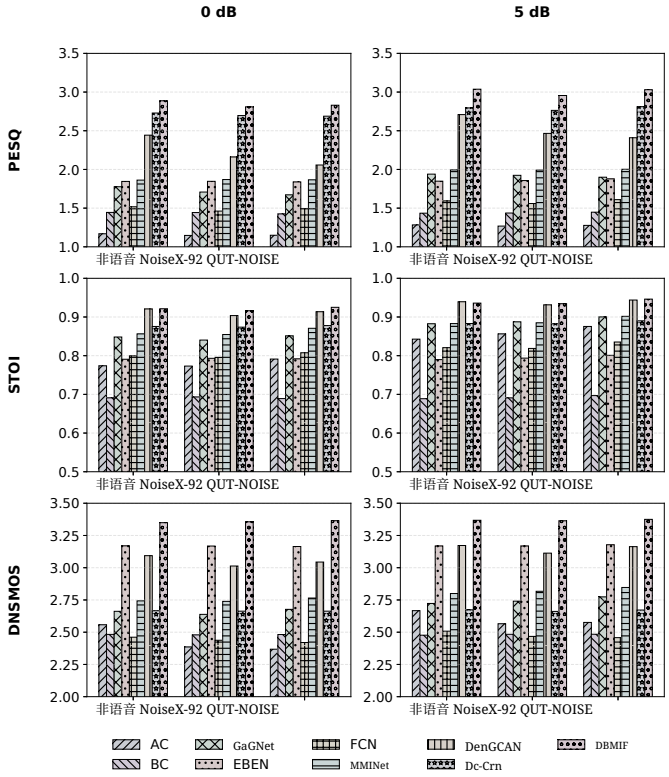


图6：不同方法在高信噪比下的结果。

图 6 展示了在 0 和 5 dB 高信噪比条件下的性能比较。随着信噪比提高、空气传导语音质量改善，融合空气传导语音模态的方法性能持续提升。值得注意的是，在信噪比为 5 dB 时，这些方法在语音质量感知评估和短时客观可懂度方面均优于仅依赖骨传导语音的方法。这一趋势凸显了在可获得空气传导语音的纯净声学特征时，骨传导语音模态固有的带宽限制。

在评估的基线方法中，MMINet 在多种噪声环境下始终保持稳定的基线性能；而在低信噪比条件下似乎较为脆弱的 DenGCAN，则随着声学环境变得清晰而显著恢复。尤为重要的是，DBMIF 在所有测试场景中始终名列前茅。

总体而言，DBMIF 在整个信噪比范围内表现出稳健的稳定性，在所有评估指标中始终位居表现最佳的方法之列。同时，它在多种噪声条件下展现出很强的泛化能力。

5.1.3 声谱图分析

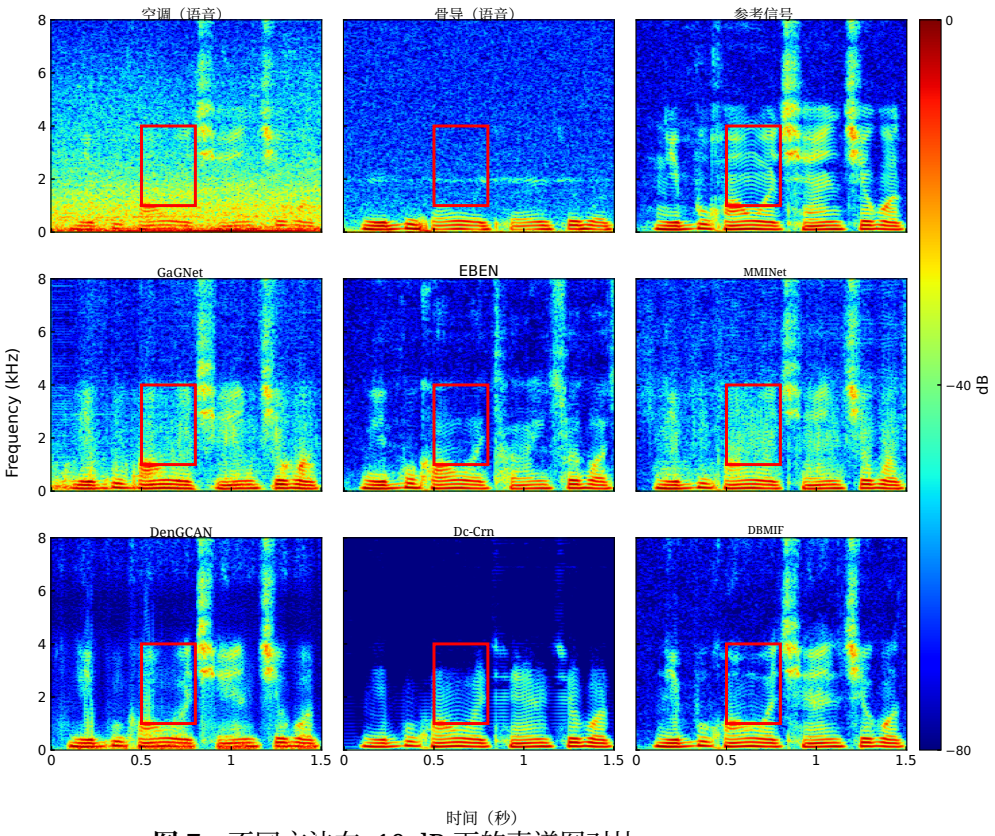


图 7：不同方法在 -10 dB 下的声谱图对比。

图 7 展示了在信噪比为 -10 dB 的街道噪声环境下的声谱图对比。在如此严峻的噪声条件下，空气传导语音信号受到背景噪声的严重掩蔽，而骨传导语音主要保留了 1 kHz 以下低频区域中的关键成分。尽管 GaGNet 实现了部分噪声抑制，却引入了明显的伪影。同时，EBEN 的输出更为干净，但明显缺乏足够的高频成分。在多模态方法中，Dc–Crn 保留了低频能量，却未能有效重建中高频细节，导致可感知的失真。类似地，尽管 MMINet 和 DenGCAN 都致力于恢复高频成分，MMINet 产生的频谱过于平滑，而 DenGCAN 的频谱改善微乎其微，呈现模糊效果。相比之下，DBMIF 的重建保真度最高，其频谱模式在高亮区域与干净参考信号高度吻合。它有效消除了 Dc–Crn 中出现的结构伪影，并恢复了其他模型遗漏的精细频谱细节。这些发现表明

这表明，DBMIF成功利用了骨导语音中的低频引导，同时保留了空气传导语音的高频和细粒度特征，从而确保在恶劣声学场景下具有更优的自然度和可懂度。

5.1.4 主观评价

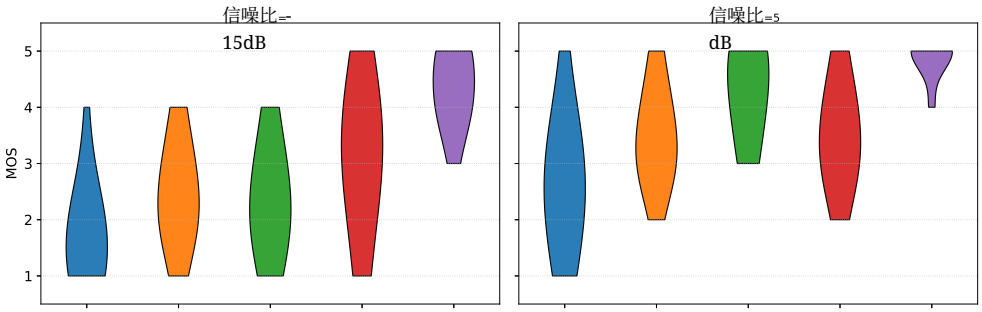


图 8：主观听音测试获得的平均意见分。较宽的部分表示更高的评分密度，位置越高表示感知质量越好。

为严格评估感知质量，我们采用绝对类别评分协议开展了主观听音测试。共有10名汉语母语者参与。干净AC语音和未经处理的噪声信号作为锚点，用于校准评分量表。图 8 展示了极端噪声条件下MOS的分布。在严重信噪比 -15 dB 条件下，对比方法表现出较大的方差，评分经常降至1，表明信号发生了显著退化。相比之下，DBMIF展现出更优的鲁棒性，其MOS评分密集分布在4到5之间，且没有样本得分低于3。这表明，即使空气传导信号受到严重破坏，DBMIF也能有效利用骨导信号，防止可懂度崩溃。在5 dB条件下，DBMIF通过较高的一致性进一步脱颖而出。不同于因残余伪影而仍保留低分尾部的基线方法，DBMIF在接近最大MOS的位置保持紧凑分布，证实了其能够保留语音的自然度和细节，同时不会引入可听失真。

5.2 消融实验

5.2.1 架构组件的影响

表5 展示了累积消融实验的结果，用于评估各组件的贡献。基线定义为不包含 CBGI、DIAF 和 DBI 模块的标准三分支模型，并逐步添加各组件以评估其

表 5：三分支架构组件的消融实验。

组件			字符错误率 (%) ↓[a]	
CBGI模块	DIAF	DBI	Whisper（自动语音识别）	WeNet（自动语音识别）
空气传导语音			47.79	34.62
骨传导语音			29.30	25.44
三分支基线			26.90	22.00
✓	–	–	19.80	15.25
✓	✓	–	18.50	14.68
✓	✓	✓	15.84	12.77

[a]向下箭头（↓）表示数值越低，性能越好。

单独影响。所提出的架构在 Whisper 和 WeNet 两种后端上均显著提升了下游自动语音识别性能。值得注意的是，CBGI模块带来了最显著的改进，使每种自动语音识别系统的字符错误率降低约 6%。这凸显了该模块在实现门控信息交换的同时保留模态特定表示方面的关键作用。在初步实验中，我们实现了标准的门控多模态单元（GMU），但发现其性能不如所提出的 CBGI。此外，DIAF 和 DBI 模块通过优化融合过程并加强分支间协作，带来了互补的收益，从而进一步提升了性能。总的来说，这些结果表明，完整的模块组合能够最有效地整合空调语音和骨传导语音中的互补线索，从而生成稳健的多模态表示。

5.2.2 分支策略的影响

我们评估了单分支、双分支和三分支变体，以考察不同分支配置在各种噪声条件下对性能的影响。在单分支变体中，AC 和 BC 语音特征通过 DIAF 进行早期融合，并由共享的编码器-解码器处理，这代表了一种简化的单阶段设计，且未显式进行模态分离。双分支变体通过交叉门控注意力机制独立处理各模态，但缺少专用的最终融合流。三分支配置则代表我们提出的完整模型，其中额外的分支专门用于整合两种模态的互补线索。

表 6中的结果表明，不同策略之间的性能差异显著。单分支基线的表现最差，尤其是在低信噪比条件下，这表明简单的早期融合难以克服模态主导问题，从强噪声中分离出干净语音。双分支配置始终优于单分支设置，证明了在相互引导下保留独立表征的优势。三分支模型在所有指标上均取得最佳结果。专门的融合分支有效增强了跨模态交互并抑制伪影，从而使语音更加清晰。

表6：不同信噪比下分支策略的性能。

	尺度不变信噪比失真比			STOI			DNSMOS		
模型	−15	−5	5	−15	−5	5	−15	−5	5
空气传导语音	-14.71	-4.73	5.03	0.52	0.69	0.85	2.29	2.48	2.60
单分支	0.86	6.04	10.96	0.67	0.75	0.81	2.57	2.66	2.74
双分支	7.15	9.98	11.96	0.82	0.86	0.91	3.19	3.21	3.23
三分支	9.27	12.86	15.89	0.88	0.91	0.94	3.34	3.35	3.36

有趣的是，由于指标趋于饱和，DNSMOS 在不同信噪比下始终保持较高水平。即使在 -15 dB 下，模型也能利用 BC 线索抑制主要伪影，从而确保基础感知质量。虽然在较高信噪比下恢复 AC 语音细节能显著提升保真度，但其带来的感知增益会逐渐减小。

表7：分支策略在各噪声类别（脉冲噪声、平稳噪声和多人谈话噪声）下的性能

	尺度不变信噪比失真比			STOI			DNSMOS		
模型	Imp.	Sta.	Bab.	Imp.	Sta.	Bab.	Imp.	Sta.	Bab.
空气传导语音	-4.87	-4.71	-4.82	0.60	0.66	0.60	2.16	2.23	2.33
单分支	4.62	7.39	4.59	0.74	0.76	0.75	2.67	2.62	2.68
双分支	8.92	9.53	9.04	0.84	0.83	0.85	3.22	3.20	3.21
三分支	11.26	12.41	11.24	0.89	0.89	0.90	3.34	3.30	3.38

表 7进一步比较了不同噪声类别下的性能。单分支模型在脉冲噪声和 babble 噪声下性能显著下降，表明其难以处理突发瞬态噪声或复杂的类语音干扰。相比之下，双分支和三分支配置表现出稳定的抗噪能力，证实当空气传导语音严重退化时，骨传导语音仍能提供可靠线索。三分支设计通过门控融合机制进一步增强了这一优势，该机制可选择性地抑制不可靠的信息流。因此，随着空气传导语音在严峻噪声环境中的可靠性降低，三分支集成的结构优势也愈发显著。

5.2.3 DIAF 中的自适应模态融合行为

图9 展示了融合权重 $w^{(3)}$的变化，其中数值越高表示对 AC 模态的依赖程度越高。如左图所示，均值权重与信噪比呈正相关，证实 DIAF 具有信噪比感知能力：在低信噪比条件下优先采用抗噪的骨传导信号，随着条件改善，则逐渐转向 AC，以恢复频谱保真度。右图进一步揭示了一种频率依赖型自适应策略。对于低频（通道1），模型始终保持较低权重，以依靠稳健的骨传导

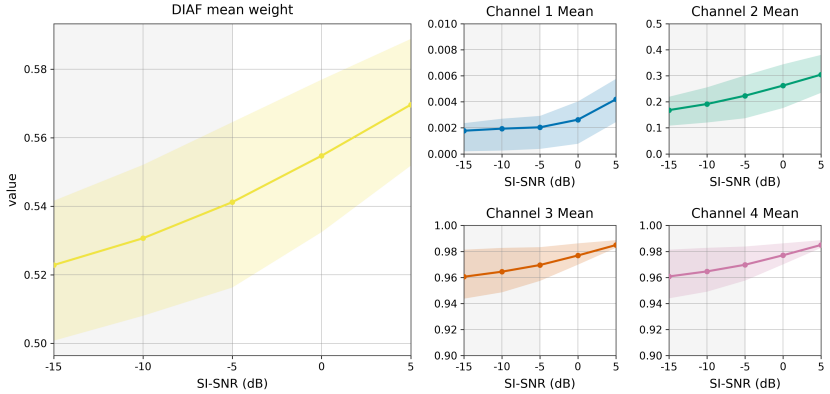


图9：DIAF融合权重 $w^{(3)}$ （$w \to 1$ 表示气导依赖）。左：全局均值随信噪比的变化，阴影区域为四分位距。右：跨子带的频率相关自适应。

模态锚定基频信息。相反，高频（通道3和通道4）则保持较高权重，依靠 AC 特征来弥补 BC传感器带宽有限的问题。值得注意的是，中频带（通道2）表现出最强的动态自适应能力，作为权衡区，根据瞬时噪声水平主动平衡骨导稳定性与气导细节。

5.2.4 DBI求解器的收敛行为

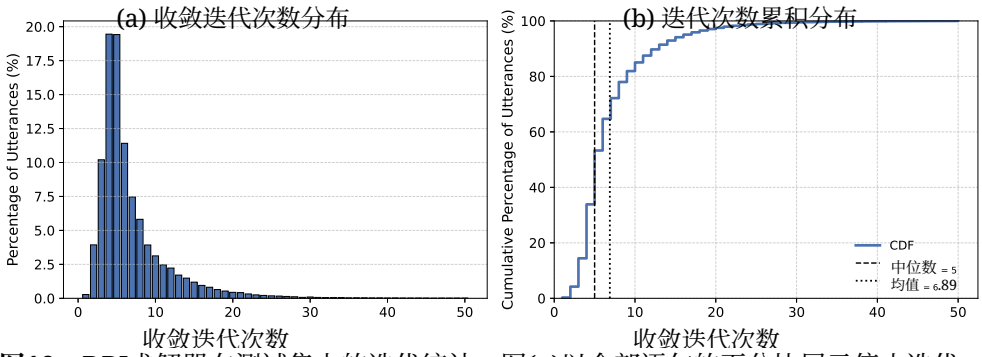


图10：DBI求解器在测试集上的迭代统计。图(a)以全部语句的百分比展示停止迭代 k^* 的分布。图(b)绘制 k^* 的累积分布函数，并标出中位数与均值。

为了量化实际推理代价，我们分析停止迭代 k^* 在测试集上的分布。此处， k^* 表示使以下指标达到最小值的迭代索引

迭代间变化量 δ_k ，且不超过给定预算 $K_{\max}$。图10 总结了 k^* 的经验分布及其累积分布函数（CDF）。

如图10(b)所示，DBI求解器通常只需几步细化即可收敛。具体而言，停止迭代的中位数为 $k^* = 5$，均值为 6.89。尽管某些较具挑战性的语句呈现轻微的长尾特征，但实际迭代次数仍远低于预设上限 $K_{\max} = 50$ 。这些结果表明， $K_{\max}$ 主要起到保守保护措施的作用。因此，对于绝大多数语句，不动点细化都能高效终止。这确保了推理过程中的计算开销保持在较低且可控的水平。

6 结论

本文提出了 DBMIF，这是一种深度平衡多模态迭代融合框架，旨在利用互补的空调和骨导信号，稳健地增强语音。所提出的架构采用由粗到细的交互策略：首先通过早期 DIAF 模块，根据瞬时可靠性自适应地为各模态分配权重；接着使用配备 CBGI 的多尺度编码器–解码器，实现受控的双向信息交换；最后通过参数共享的 DBI 瓶颈逐步细化融合表示，同时不增加模型规模。实验结果证实，在多种噪声环境下，所提出的框架始终优于最先进的单模态和多模态基线。

尽管其具备泛化能力，当前的非因果架构和高计算复杂度给实时部署带来了挑战。此外，尽管DBMIF在极端噪声条件下取得了显著改进，增强语音与干净语音之间的可懂度差距仍然明显，表明在超低信噪比场景中仍需进一步创新。未来工作将聚焦于优化模型，以在嵌入式平台上实现因果流式推理，并探索自监督预训练以进一步提升增强性能。

致谢。 本工作部分得到国家自然科学基金委员会（批准号62332019）、国家重点研发计划（2023YFF1203900、2023YFF1203903）以及北京科技新星计划（20240484513）的资助。

数据可用性。 本文所用数据均来自公开数据集。

声明

利益冲突 作者声明，他们不存在已知的竞争性经济利益或个人关系，可能会对本文所报告的研究产生影响。

参考信号

[1] Wang, D., Chen, J.：基于深度学习的监督式语音分离：综述。IEEE/ACM音频、语音与语言处理汇刊 26(10)，1702–1726（2018年）https://doi.org/10.1109/TASLP.2018.2853117[2] Healy, E.W., Johnson, E.M., Pandey, A., Wang, D.：基于深度学习的降噪在效能与可行性方面取得的进展。《美国声学学会杂志》153(5)，2751–2751（2023）https://doi.org/10.1121/10.0019341[3] Zhou, D., Zhang, Y., Wu, J., Zhang, X., Xie, L., Yin, E.：Ave speech：用于语音识别的综合多模态数据集，融合音频、视觉和肌电信号。IEEE人机系统汇刊（2025年）https://doi.org/10.1109/THMS.2025.3585165[4] Dubey, H., Aazami, A., Gopal, V., Naderi, B., Braun, S., Cutler, R., Ju, A., Zohourian, M., Tang, M., Golestaneh, M., 等：ICASSP 2023年深度降噪挑战赛。IEEE开放信号处理期刊 5，725–737（2024年）https://doi.org/10.1109/OJSP.2024.3378602[5] Zhou, Y., Chen, Y., Ma, Y., Liu, H.：一种由骨传导传感器辅助的实时双麦克风语音增强算法。《传感器》20(18)，5050（2020年）https://doi.org/10.3390/s20185050[6]Huang, B., Gong, Y., Sun, J., Shen, Y.：一种适用于强背景噪声环境的可穿戴式骨传导语音增强系统。载于：2017年第18届国际电子封装技术会议（ICEPT），第1682–1684页（2017年）。https://doi.org/10.1109/ICEPT.2017.8046759 。IEEE[7] Ito, T., R¨o¨osli, C., Kim, C., Sim, J., Huber, A., Probst, R.：在人体头部不同部位施加刺激时测得的牙齿处骨导听阈与颅骨振动。《听力学与神经耳科学》16(1)，12–22（2010）https://doi.org/10.1159/000314282[8] Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N.：使用贴附于皮肤的听诊式麦克风进行非可听低语识别的输入接口。载于：2003年IEEE国际声学、语音与信号处理会议论文集（ICASSP'03），第5卷，第708页（2003年）。https://doi.org/10.1109/ ICASSP.2003.1200069 。IEEE[9] Kondo, K., Fujita, T., Nakagawa, K.：关于骨传导语音均衡以提升语音质量。载于：2006年IEEE国际信号处理与信息技术研讨会，第426–431页（2006年）。https://doi.org/10. 1109/ISSPIT.2006.270839 。IEEE

[10] Pan, Q., Pan, Y., Zhou, J., Wang, H., Tao, L., Kwan, H.K.：带有双重的CycleGAN

对抗损失的骨传导语音增强。载于：TENCON 2022-2022年IEEE 第十区会议（TENCON），第1–4页（2022年）。https://doi.org/10.1109/TENCON55691.2022.9977942。IEEE[11]Hauret, J., Joubaud, T., Zimpfer, V., Bavu, ´E.：EBEN：应用于抗噪声体传导麦克风所采集语音信号的极限带宽扩展网络。载于：ICASSP 2023-2023年IEEE国际声学、语音与信号处理会议（ICASSP），第1–5页（2023年）。https://doi.org/10.1109/ICASSP49357.2023.10096301。IEEE[12] Hao, B., Zhou, D., Li, X., Zhang, X., Xie, L., Wu, J., Yin, E.：Lipgen：用于增强视觉语音识别的视素引导唇部视频生成。载于：IEEE国际声学、语音与信号处理会议，第1–5页（2025年）。 https://doi.org/10.1109/ICASSP49660.2025.10889163[13]Dusan, S.V., Lindahl, A., Andersen, E.B.：在移动设备中混合加速度计与麦克风信号以改善语音质量的系统与方法。Google专利。美国专利 9,363,596（2016年）[14] Lee, C.-H., Rao, B.D., Garudadri, H.：骨传导传感器辅助噪声估计以改进语音增强。载于：Interspeech，第2018年卷，第1180页（2018年）[15] Yu, C., Hung, K.-H., Wang, S.-S., Tsao, Y., Hung, J.-w.：时间域多模态骨/气传导语音增强。IEEE信号处理快报**27**，1035–1039（2020年） https://doi.org/10.1109/LSP.2020.3000968[16] Wang, H., Zhang, X., Wang, D.：融合骨传导与气传导传感器的复数域语音增强。IEEE/ACM音频、语音与语言处理汇刊 **30**，3134–3143（2022年）https://doi.org/10.1109/TASLP.2022.3209943[17] Kuang, K., Yang, F., Yang, J.：融合骨传导与气传导语音的轻量级语音增强网络。美国声学学会期刊**156**(2)，1355–1366（2024年）https://doi.org/10.1121/10.0028339[18] Tan, K., Wang, D.：用于实时语音增强的卷积循环神经网络。载于：Interspeech，第2018年卷，第3229–3233页（2018年）。 https://doi.org/10.21437/Interspeech.2018-1405[19]Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., Xie, L.：DCCRN：用于相位感知语音增强的深度复数卷积循环网络。Interspeech（2020年）https://doi.org/10.21437/interspeech.2020-2537[20] Li, A., Zheng, C., Zhang, L., Li, X.：Glance and gaze：用于单通道语音增强的协同学习框架。应用声学 **187**，108499

(2022年) https://doi.org/10.1016/j.apacoust.2021.108499

[21]Pascual, S., Bonafonte, A., Serr`a, J.：《SEGAN：语音增强生成对抗网络》。Interspeech，3642（2017）https://doi.org/10.21437/ Interspeech.2017-1428[22] D´efossez, A., Synnaeve, G., Adi, Y.：《波形域实时语音增强》。Interspeech（2020）https://doi.org/10.21437/Interspeech. 2020-2773[23]Sato, H., Ochiai, T., Delcroix, M., Moriya, T., Ashihara, T., Masumura, R.：《采用自监督表征空间损失的通用语音增强》。《信号处理前沿》5，1587969（2025）[24] Shimamura, T., Tamiya, T.：《骨传导语音重建滤波器》。载于：第48届中西部电路与系统研讨会，2005年，第1847–1850页（2005）。https://doi.org/10.1109/MWSCAS.2005.1594483。IEEE[25] Toda, T., Nakagiri, M., Shikano, K.：《用于体传导清音语音增强的统计语音转换技术》。IEEE音频、语音与语言处理汇刊20(9)，2505–2517（2012）https://doi.org/10.1109/TASL.2012.2205241[26] Shin, H.S., Kang, H.-G., Fingscheidt, T.：《骨传导麦克风辅助语音增强综述》。载于：《语音通信》；第10届ITG研讨会，第1–4页（2012）。VDE[27] Liu, H.-P., Tsao, Y., Fuh, C.-S.：《采用深度去噪自编码器的骨传导语音增强》。《语音通信》104，106–112（2018）https://doi.org/10.1016/j.specom.2018.06.002[28] Erzin, E.：《通过联合分析喉部麦克风与声学麦克风录音，提升喉部麦克风语音识别性能》。IEEE音频、语音与语言处理汇刊17(7)，1316–1324（2009）https://doi.org/10.1109/TASL. 2009.2016733[29] Zheng, C., Zhang, X., Sun, M., Yang, J., Xing, Y.：《基于深度双向长短期记忆循环神经网络的新型喉部麦克风语音增强框架》。载于：2018年IEEE第4届计算机与通信国际会议（ICCC），第1258–1262页（2018）。https://doi.org/10.1109/CompComm.2018. 8780872。IEEE[30]Zheng, C., Cao, T., Yang, J., Zhang, X., Sun, M.：《利用基于注意力的上下文信息和时频结构约束恢复骨传导语音频谱》。IEICE电子、通信与计算机科学基础汇刊102(12)，2001–2007（2019）https://doi.org/10.1587/transfun.E102.A.2001

[31] Pan, Q., Zhou, J., Gao, T., Tao, L.：《基于循环一致对抗网络的骨传导语音到空气传导语音转换》。载于：2020年IEEE第3届信息通信与信号处理国际会议（ICICSP），第168–172页（2020）。https://doi.org/10.1109/ICICSP50920.2020.9232121。IEEE[32] Cheng, L., Dou, Y., Zhou, J., Wang, H., Tao, L.：《面向骨传导语音的说话人无关频谱增强》。《Algorithms》16(3)，153（2023）https://doi.org/10.3390/a16030153[33] Li, C., Yang, F., Yang, J.：《一种恢复骨传导语音质量的两阶段方法》。IEEE/ACM音频、语音与语言处理汇刊32，818–829（2023）https://doi.org/10.1109/TASLP.2023.3337988[34] Dekens, T., Verhelst, W.：《通过均衡与信号融合增强体传导语音》。IEEE音频、语音与语言处理汇刊21(12)，2481–2492（2013）https://doi.org/10.1109/TASL.2013.2274696[35] Lai, R.L., Hou, J.-C., Chern, I.-C., Hung, K.-H., Chen, Y.-T., Gogate, M., Arslan, T., Hussain, A., Lin, C.-W., Tsao, Y.：《利用自监督音视频预训练模型提高人工耳蜗模拟中的声码器语音可懂度》。IEEE《生物医学工程汇刊》（2025）[36] Kim, Y., Chung, Y.：《面向声学麦克风和体传导麦克风框架的模态专属语音增强与噪声自适应融合》（2025）[37] Nguyen,T.Q.：《近乎完美重建的伪QM滤波器组》。IEEE《信号处理汇刊》42(1)，65–76（2002）https://doi.org/10.1109/78.277808[38] Ni, J., Bai, Y., Zhang, W., Yao, T., Mei, T.：《深度平衡多模态融合》。arXiv预印本arXiv:2306.16645 （ 2023 ） https://doi.org/10.48550/arXiv. 2306.16645[39] Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W.Z., Sotelo, J., De Brebisson, A., Bengio, Y., Courville, A.C.：《MelGAN：用于条件波形合成的生成对抗网络》。神经信息处理系统进展32（2019）[40] Wang, M., Chen, J., Zhang, X.-L., Rahardja, S.：《在空气传导与骨传导语音语料库上进行端到端多模态语音识别》。IEEE/ACM音频、语音与语言处理汇刊31，513–524（2022）https://doi.org/10.1109/TASLP.2022.3224305[41] Reddy, C.K., Gopal, V., Cutler, R.：《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：ICASSP 2021—2021年IEEE国际声学、语音与信号处理会议（ICASSP），第6493–6497页（2021）。https://doi.org/10.1109/ICASSP39728.2021.

9414878 . IEEE

[42] Saki, F., Sehgal, A., Panahi, I., Kehtarnavaz, N.: 使用子带特征和随机森林分类器对噪声信号进行基于智能手机的实时分类。载于：2016年IEEE国际声学、语音与信号处理会议（ICASSP），第2204–2208页（2016年）。https://doi.org/10.1109/ICASSP.2016.7472068。IEEE[43] Dean, D., Sridharan, S., Vogt, R., Mason, M.: 用于评估语音活动检测算法的QUT-NOISE-TIMIT语料库。载于：国际语音通信协会第11届年会论文集，第3110–3113页（2010年）。https://doi.org/10.21437/Interspeech.2010-774 。国际语音通信协会[44]Varga, A., Steeneken, H.J.: 自动语音识别评估：II. NOISEX-92：用于研究加性噪声对语音识别系统影响的数据库与实验。《语音通信》**12**(3), 247–251（1993年）https://doi.org/10.1016/0167-6393(93)90095-3[45] Hu,G., Wang, D.: 一种用于基音估计和浊音语音分离的串联算法。IEEE音频、语音与语言处理汇刊**18**(8), 2067–2079（2010年）[46] Wang, M., Chen, J., Zhang, X., Huang, Z., Rahardja, S.: 时域中利用骨传导语音进行多模态语音增强。《应用声学》**200**，109058（2022年）https://doi.org/10.1016/j.apacoust.2022.109058[47] Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: 语音质量感知评价（PESQ）——一种用于电话网络和编解码器语音质量评估的新方法。载于：2001年IEEE国际声学、语音与信号处理会议论文集（目录号01CH37221），第2卷，第749–752页（2001年）。https://doi.org/10.1109/ICASSP.2001.941023 。IEEE[48] Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: 一种用于预测时频加权含噪语音可懂度的算法。IEEE音频、语音与语言处理汇刊**19**(7), 2125–2136（2011年）https://doi.org/10.1109/TASL.2011.2114881[49] Jensen, J., Taal, C.H.: 一种用于预测调制噪声掩蔽下语音可懂度的算法。IEEE/ACM音频、语音与语言处理汇刊**24**(11), 2009–2022（2016年）https://doi.org/10.1109/TASLP.2016.2585878[50] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: 通过大规模弱监督实现鲁棒语音识别。载于：国际机器学习会议，第28492–28518页（2023年）。PMLR

[51] Yao, Z., Wu, D., Wang, X., Zhang, B., Yu, F., Yang, C., Peng, Z., Chen, X., Xie, L., Lei, X.：Wenet：面向生产的流式和非流式端到端语音识别工具包。载于：Interspeech，第 2021 卷，第 4054–4058 页（2021）。https://doi.org/10.21437/Interspeech.2021-1983