视觉信息辅助语音增强：使用基于注意力的波束形成

Chihyun Liu、Jiaxuan Fan、Mingtung Sun、Michael Anthony、Mingsian R. Bai（IEEE 高级会员）和 Yu Tsao（IEEE 高级会员）

摘要—近期研究表明，结合说话人声纹或视觉线索等辅助信息，能够显著提升语音增强（SE）性能。然而，在信噪比较低、混响较强，或涉及动态说话人、重叠语音或非平稳噪声的复杂场景中，单通道方法往往效果不佳。为解决这些问题，我们提出了一种新型视觉信息辅助神经波束形成网络（VI-NBFNet），通过多模态输入特征融合麦克风阵列信号处理与深度神经网络（DNN）。该网络利用预训练的视觉语音识别模型提取唇部运动作为输入特征，用于语音活动检测（VAD）和目标说话人识别。系统采用配备注意力机制的监督式端到端波束形成框架，旨在同时处理静止和移动说话人。实验结果表明，与多种基线方法相比，所提出的视听系统在静止和动态说话人场景下均实现了更佳的语音增强性能和鲁棒性。

索引词—多通道语音增强、视听、波束形成、深度学习、注意力机制

一、引言

语音增强是一种通过减轻背景噪声、干扰和混响来提升语音质量与可懂度的技术。它在多种应用中发挥着重要作用，包括视频会议 [1, 2], 助听器 [3, 4], 语音助手 [5], 以及自动语音识别（ASR）系统 [6–8]。尽管深度学习推动了显著进步，但单耳语音增强方法在复杂声学环境中仍面临诸多挑战，尤其是在存在类语音干扰的情况下，例如竞争说话人，或含有人类对话的背景电视噪声；在这些情况下，分离目标语音本质上十分困难。

为解决这一局限性，研究人员探索了各种辅助模态，以引导语音增强系统聚焦于由说话人嵌入所表征的目标说话人。这类嵌入的示例包括广泛采用的 i-vector[9, 10], 、d-向量[11, 12], 和 X向量[13], 。将这些嵌入纳入语音增强框架[14, 15]后，已展现出显著的性能提升。然而，这些设备通常需要预先注册；当竞争说话人具有相似的声学特征（例如属于同一性别）时，它们可能会失效。

与基于声学的嵌入不同，面部和唇部运动等视觉线索能够提供互补线索，且本质上不易受到声学干扰。借助唇读特征或面部嵌入，语音增强系统能够更有效地增强目标语音信号，尤其是在低信噪比和多说话人环境中 [16, 17]。已有多种方法通过裁剪唇部邻域图像来训练和测试语音增强模型，从而提取视觉嵌入 [18–21]。为降低图像处理的计算复杂度，已有研究报告了采用低分辨率嘴唇区域输入 [22] 和逐位压缩技术 [23, 24] 等策略。为加快视觉特征的提取，研究者引入了特定任务的预训练网络。例如，[26] 采用从预训练人脸识别模型 FaceNet[25], 中获得的面部嵌入，相较于原始图像输入，该嵌入能够更准确地捕捉面部表情特征。这种表征已被证明能够缓解说话人歧义，尤其是在说话人具有相似视觉特征的情况下 [27]。此外，[29, 30] 采用了从视觉语音识别（VSR）模型 [28] 中提取的词级唇读嵌入，将语言信息注入增强或分离流程。正如 [31], 所述，该研究引入了从预训练自动语音识别模型 [32]中提取的唇读嵌入。这些嵌入能够有效检测语音活动区域，并

本研究受台湾国家科学及技术委员会（NSTC）项目 113-2221-E-007-057-MY3 资助。（通讯作者：Mingsian R. Bai）。
Chihyun Liu 任职于台湾新竹国立清华大学动力机械工程系（电子邮件：joanna510119@gmail.com）。Jiaxuan Fan 任职于台湾新竹国立清华大学~~电机工程系（电子邮件：~~ frita7488@gmail.com）。Mingtung Sun 任职于台湾新竹国立清华大学电机工程系（电子邮件：sunmtn0923@gmail.com）。Michael Anthony 任职于台湾新竹国立清华大学电机工程系（电子邮件：michaelzhang220~~@gmail.com）。~~ Mingsian R. Bai 任职于台湾新竹国立清华大学动力机械工程系和电机工程系（电子邮件：msbai@pme.nthu.edu.tw）。Yu Tsao 任职于台湾中央研究院信息技术创新研究中心（电子邮件： minwelltsao@gmail.com）。

静音区域，并在极端嘈杂的条件下提高可懂度。

根据所用特征的不同，语音增强通常分为三类：音频引导语音增强（AI-SE）、视觉引导语音增强（VI-SE）和音视频引导语音增强（AVI-SE）。与不使用任何辅助信息的模型相比，VI-SE 能显著提升性能 [33],，这也是本研究的重点。尽管一些研究表明，AVI-SE 的性能优于 VI-SE [34], ，但相较于其更高的计算复杂度以及说话人声纹注册带来的实际限制，性能提升往往微乎其微。相比之下，VI-SE 在计算开销极小的情况下，显著优于 AI-SE。因此，本文采用 VI-SE 方法，作为适用于现实应用的轻量级解决方案。

尽管已有多种先进的单通道端到端 VI-SE 方法被提出 [35–37], ，但此类系统缺乏空间信息，往往会导致明显的语音失真，尤其是在处理重叠语音信号时。因此，研究者开发了利用空间线索的多通道 VI-SE 框架，例如通道间相位差（IPD）和到达方向（DOA） [21, 34]。一些多通道方法尝试直接根据音视频输入估计波束形成器的参数 [20, 38]。然而，这些方法仅适用于静止说话人，对于非平稳说话人则并非总是适用。

为了解决上述局限性，我们提出了一种新颖的框架——视觉信息辅助神经波束形成网络（VI-NBFNet），该框架将唇读特征作为辅助视觉线索，并结合麦克风阵列波束形成技术。该架构以端到端的方式联合学习视听特征、掩蔽表示和空间信息，并直接以波束形成输出作为监督。与众多两阶段数据驱动系统不同，我们的方法无需使用真实掩膜进行监督，并可对网络中的所有模块进行统一优化。本文介绍了一种端到端系统，该系统结合了基于注意力机制的时变空间协方差矩阵（SCM）估计模型 [39] ，以及作为语音增强后置滤波器的配套视觉信息增强深度滤波器 [40] 。

增强性能通过以下客观测试和性能指标进行评估：语音质量感知评估（PESQ） [41]，短时客观可懂度（STOI） [42]，和深度噪声抑制平均意见分（DNSMOS） [43]。评估中采用一种单通道方法作为基线方法之一，该方法由改进版轻量音视频语音增强（iLAVSE）[24] 改编而来，以下称为视觉引导单通道语音增强（VI-SSE）。此外，还采用基于掩蔽的波束形成器——视觉引导多通道语音增强（VI-MSE） [44], [45], 作为另一种基线方法。在该方法中，时频（T-F）掩码通过神经网络估计，用于计算波束形成权重所需的空间协方差矩阵。最后一种基线方法——视觉引导自注意力波束形成器（VI-SA-BF）——是一种基于注意力机制的网络，它直接根据瞬时空间协方差矩阵（ISCMs）[46]估计时变空间协方差矩阵。

本文的主要贡献总结如下。首先，我们在多模态框架中将唇动特征与音频特征相结合，用于语音增强网络。事实证明，这种融合对语音质量和可懂度有显著影响，尤其是在声学条件不佳的情况下。其次，我们提出了一种基于混合架构的新型 VI-NBFNet，将波束形成与深度学习相结合，以有效抑制噪声和干扰，同时尽量减少语音失真。最后，所提出的系统采用联合训练的注意力网络，以端到端的方式学习空间信息，从而使模型能够为移动目标说话人估计动态波束形成权重

无需额外的头部跟踪器。

本文其余部分安排如下：第二节介绍问题表述和基线系统。第三节介绍所提出的 VI-NBFNet 系统。第四节介绍实验设置和结果。第五节给出结论与未来工作。

II. 问题表述与基线方法

A. 问题表述

考虑这样一种情形：目标语音信号和方向性干扰由混响房间中的 M 个麦克风采集。麦克风接收到的信号可在短时傅里叶变换（STFT）域表示为

$$Y^m(l,f) = A_s^m(f)S(l,f) + \sum_{n=1}^{N} A_n^m(f)I_n(l,f) + V^m(l,f), \quad (1)$$

其中，l 和 f 分别表示时间帧索引和频率单元索引， 表示第 m 个麦克风（$m = 1,2,...,M$），A_s^m 与目标说话人之间的声学传递函数（ATF）， 表示第 m 个麦克风与第 n 个干扰说话人（$n = 1,2,...,N$）之间的 ATF， 表示目标说话人信号， 表示第 n 个干扰说话人信号， 表示第 m 个传感器噪声。目标是在存在多个方向性干扰说话人和加性传感器噪声的情况下，从麦克风信号中提取目标说话人的语音信号。假设摄像头捕获了目标说话人的图像。这些视觉信息有助于模型识别目标说话人。

B. 基线方法

1) 单通道语音增强系统：该基线系统基于 iLAVSE 框架修改而来 [24]，并融入了以下三项关键特性。首先，该系统采用基于掩码的方法，而非直接映射（即网络将带噪信号直接映射为干净语音信号）。其次，采用低维唇部嵌入

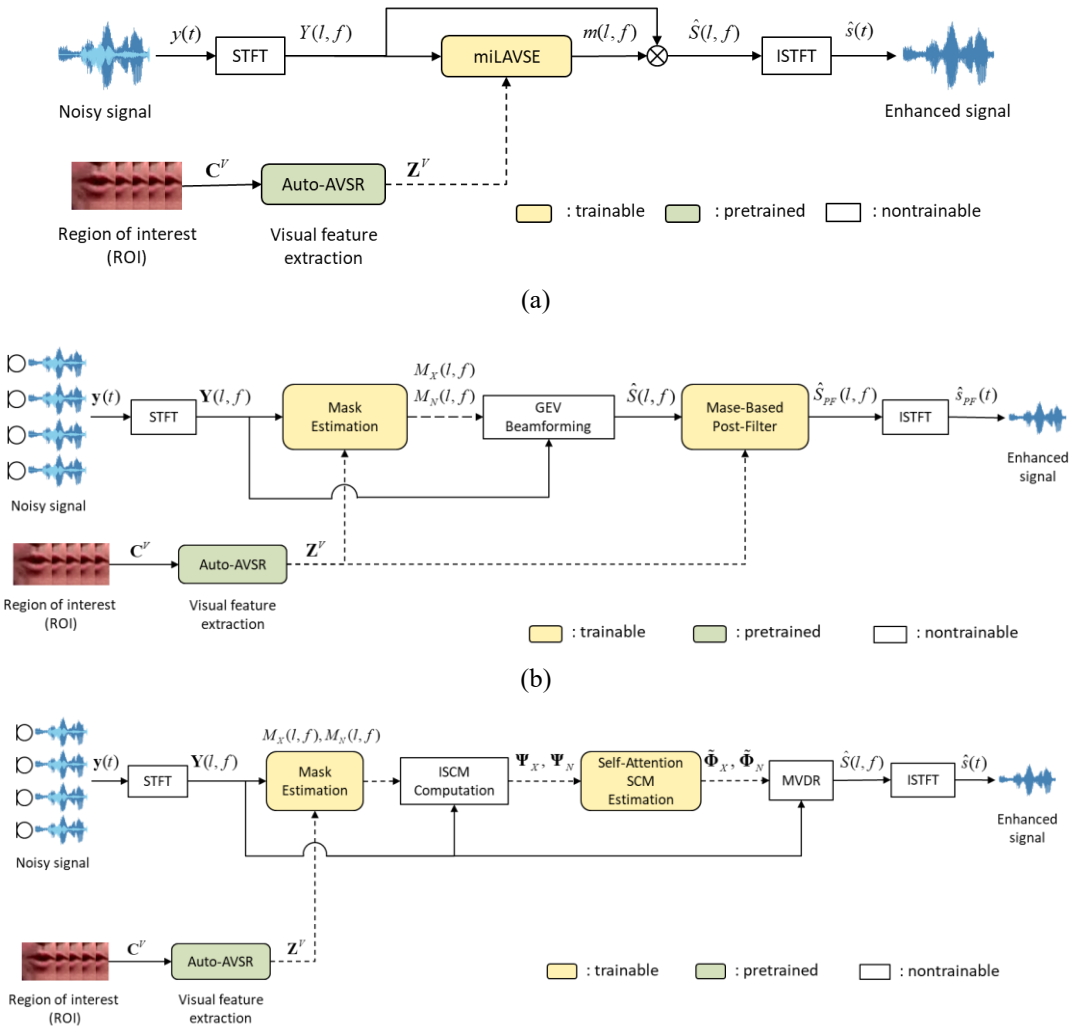


图 1. 基线方法的框图。(a) VI-SSE 系统，(b) VI-MSE 系统，(c) VI-SA-BF系统。黄色和绿色方框分别表示可训练网络和预训练网络，白色方框表示不可训练模块。

这些嵌入由预训练的视觉语音识别模型提取，用以替代唇部图像。在原始 iLAVSE 中，图像处理通过一系列步骤完成：先通过颜色通道缩减、分辨率下采样和量化来缩减图像数据，再使用基于卷积神经网络（CNN）的自编码器压缩视觉输入。相比之下，当前基线采用的嵌入能够隐式编码语音内容和说话人活动，因此比原始模型提供更丰富、更具信息量的视觉线索。接下来，模型的输入从上下文窗口扩展至完整语谱图，使网络能够识别长期时间依赖关系。此外，为了尽量降低实时处理的时延，iLAVSE 中原先采用的双向长短期记忆网络（LSTM）被单向长短期记忆网络取代。总之，如图 1(a) 所示，完整的基线系统 VI-SSE 是一种基于掩码的网络，它将从预训练视觉语音识别（VSR）模型中提取的视觉特征作为辅助信息。在该系统中，网络估计一个时频掩码来处理语谱图。平均训练模型时，采用增强语谱图与干净语谱图之间的均方误差（MSE）作为损失函数。该网络称为改进型 iLAVSE（miLAVSE）。

2）基于神经网络的波束形成器与后置滤波器：基线系统 VI-MSE 如图 1(b) 所示。该系统由三个主要模块组成：掩码估计模块、波束形成模块和后滤波模块。掩码估计模块采用近期神经波束形成器[44]，中使用的常规流程，其中估计语音和噪声的时频掩码。这些掩码用于计算波束形成器所需的空间协方差矩阵。VI-MSE 中的掩码估计网络与 miLAVSE 的类似，不同之处在于它预测两个理想比值掩码（IRM），而非一个。掩码估计网络如图 2 所示。值得注意的是，尽管[44] 采用以纯净语音和噪声作为训练目标的理想二值掩码（IBM），本研究采用类维纳掩码（WLM）[47]，，其提供更平滑且信息更丰富的监督信号。在波束形成模块中，时不变的语音和噪声空间协方差矩阵

利用式（2）和式（3）针对每个频率单元进行计算。如 [44], 中所建议，在最终阶段采用广义特征值（GEV）波束形成器结合盲解析归一化（BAN）后滤波器以获得增强信号。

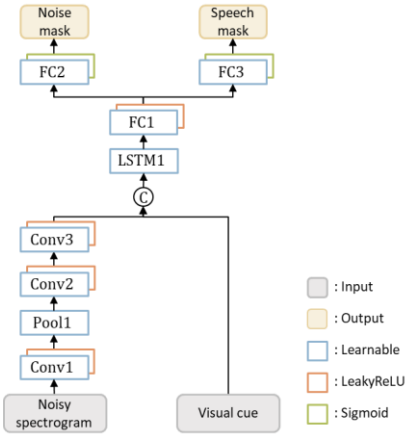


图 2. 掩码估计网络的网络架构。

$$\mathbf{\Phi}_{\mathrm{XX}}(f)=\frac{\sum_{l=1}^{T} M_{\mathrm{X}}(l,f)\mathbf{Y}(l,f)\mathbf{Y}(l,f)^{H}}{\sum_{l=1}^{T} M_{\mathrm{X}}(l,f)} \in \mathbb{C}^{M\times M} \quad (2)$$

$$\mathbf{\Phi}_{\mathrm{NN}}(f)=\frac{\sum_{l=1}^{T} M_{\mathrm{N}}(l,f)\mathbf{Y}(l,f)\mathbf{Y}(l,f)^{H}}{\sum_{l=1}^{T} M_{\mathrm{N}}(l,f)} \in \mathbb{C}^{M\times M} \quad (3)$$

其中 表示频率单元的总数，以及分别表示频率单元 处的时不变语音和噪声空间协方差矩阵， 表示预测的语音掩码，表示预测的噪声掩码。 表示厄米转置。 波束形成器作为预处理器，已证明能够有效抑制在空间上可分离的噪声源。然而，在某些情况下，输出中仍能听到残余噪声。为解决这一问题，我们采用基于学习的后滤波器进一步降低噪声，尤其是针对共址声源。后滤波器网络与 miLAVSE 网络相似，但以波束形成后的频谱图作为输入。在这种情况下，干净语音频谱图用作训练目标。掩码估计网络和后置滤波器采用均方误差损失进行训练。

3) 基于自注意力的神经波束形成器：该基线基于[46]中提出的用于估计时变SCM的框架。原始方法利用 TasNet [48]估计单说话人场景下的时频掩码。通过集成源自VI-MSE 系统的视觉信息引导的掩码估计模块，对该架构进行了修改，以适应多源场景。具体而言，将基于TasNet的掩码估计器替换为基于 miLAVSE 的掩码估计器，以纳入与目标说话人相关的视觉线索；这对于在目标说话人与干扰说话人同时共存时进行准确的掩码估计至关重要。基于自注意力的空间协方差矩阵估计与神经波束形成器的实现保持不变。上述架构称为VI-SA-BF，如图1(c)所示。

III. 提出的方法　方法

在本节中，我们提出VI-NBFNet，以应对具有挑战性的语音增强任务。所提出的系统采用混合架构，将麦克风阵列信号处理（MASP）与深度神经网络相结合。该网络的一个关键特征是对视听信息与空间信息进行端到端联合学习，从而能够预测时变空间协方差矩阵估计所需的权重。所提出系统的示意图如图3所示。该系统由四个主要组件组成：视听编码器、掩码解码器、空间感知解码器以及空间协方差矩阵计算模块。尽管采用模块化设计架构，各组件仍以端到端方式联合优化，使网络能够有效地从空间特征和频谱特征中学习。

A. 视觉特征提取

唇部运动被视为高度可靠的语音活动检测，也是自动语音识别中与词语相关的特征。同样，唇读可作为语音增强任务的有力工具。预训练的视听语音识别（AVSR）模型（auto- AVSR）[49], 被用于为所提出的系统及所有基线方法提取视觉特征。auto- AVSR的前端由自动语音识别编码器和视觉语音识别编码器组成。自动语音识别编码器的输入是带噪语音信号。视觉语音识别编码器的输入是相应的裁剪后的唇部图像 目标说话人的，其中表示图像帧数。音频编码器与视频编码器的输出随后通过多层感知机（MLP）进行整合。视觉语音识别编码器负责生成唇读特征。这些潜在特征不仅与语音活动相关，还与语音发音相关的词嵌入相关，而这在自动语音识别中被广泛使用。多层感知机的输出送入后端，后端由投影层和Transformer解码器组成。在所提出的系统中，上述潜在特征被用于语音增强任务。鉴于视频流的帧率为每秒25帧（fps），而音频流的帧率为100 fps，两种模态之间需要进行时间对齐。因此，每个视频帧被重复四次，以与音频流对齐，从而确保视觉特征与带噪音频频谱图沿时间轴同步。这一步骤对于后续的多模态融合阶段至关重要。

B. 视听编码器

VI-NBFNet采用了与iLAVSE类似的视听融合策略，即将音频和视觉特征拼接后输入长短期记忆网络层，以检测时间依赖关系。然而，这两种模型的音频编码器配置有所不同。尽管iLAVSE采用了一个简单的三层卷积神经网络，专为单

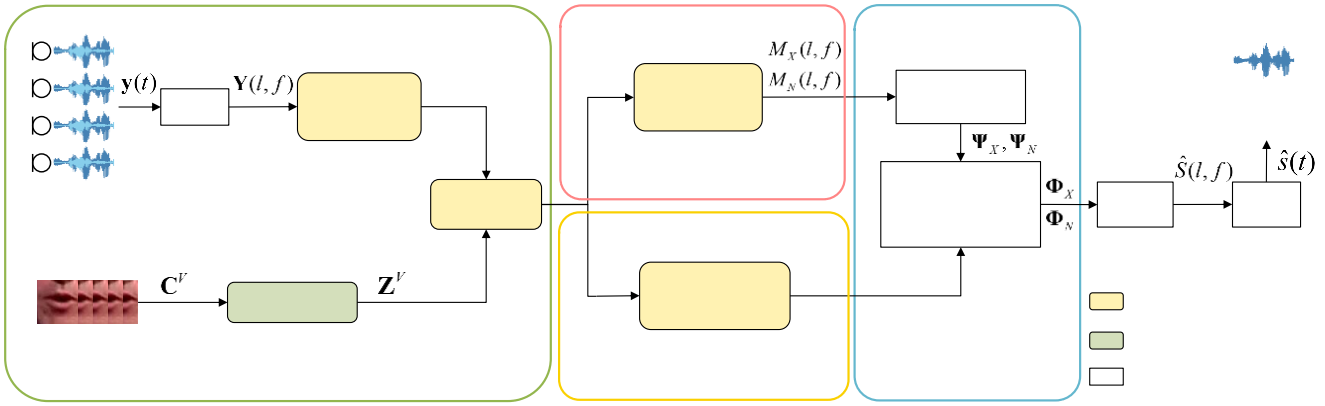


图3. VI-NBFNet的示意图。黄色和绿色模块分别表示可训练网络和预训练网络，白色模块表示不可训练模块。

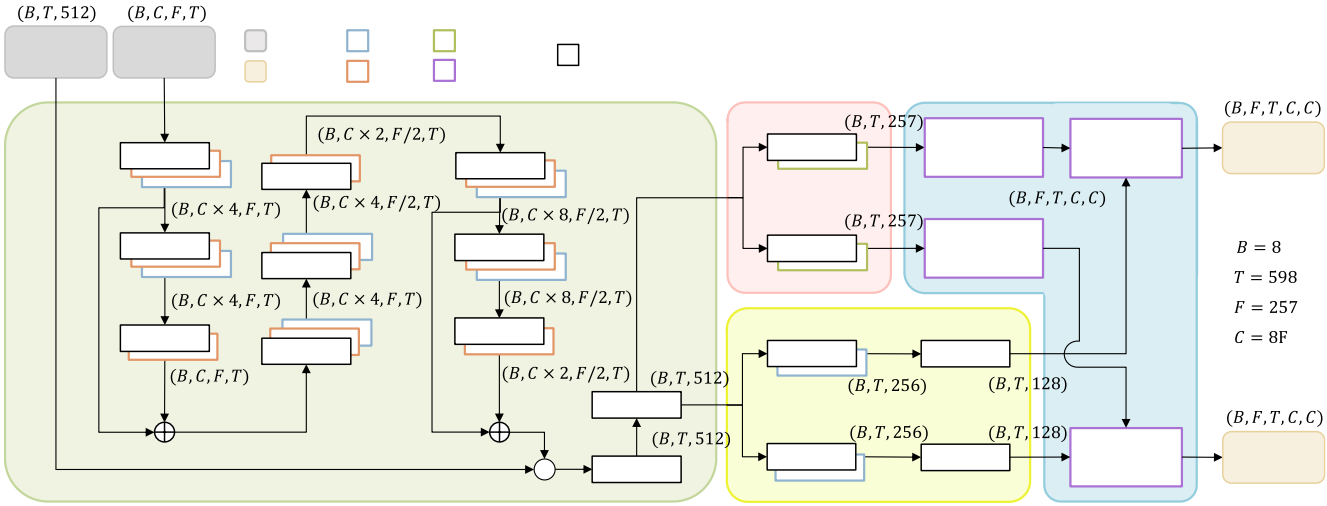


图4. VI-NBFNet的网络架构和张量 参数。

通道掩码估计而设计，VI-NBFNet则采用轻量版MobileNetV2 [50] ，以更有效地处理多通道输入并生成表达力更强的表征。这种方法旨在从麦克风阵列信号中提取更显著的特征，这些特征对于估计波束形成所需的空间协方差矩阵计算权重至关重要。尽管MobileNetV2最初并非为空间建模而设计，但其采用的逐点和逐深度可分离卷积在计算复杂度与表征效果之间实现了权衡。如图5所示，该架构作为我们音频编码器的主干网络。具体而言，初始的逐点（1×1）卷积对堆叠的输入通道进行线性组合，从而使网络能够学习麦克风间相关性。随后，采用深度（3×3）卷积提取每个通道中的局部T-F模式。尽管该网络并未显式编码音频信号的空间幅度和相位差，但通过沿通道轴堆叠多通道信号的实部和虚部，实现了对空间信息的隐式编码。研究表明，所学习的表征有助于空间感知解码器和注意力等下游模块——基于SCM的估计模块能够更有效地建模空间特征。

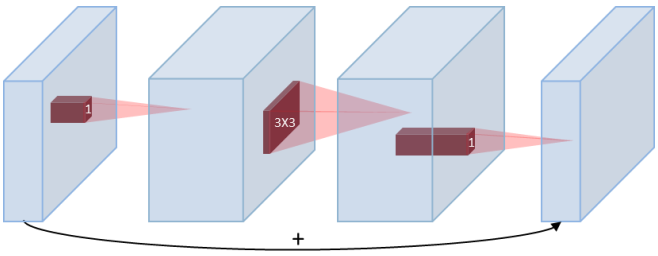


图5. MobileNetV2的倒置残差结构。

C. 掩码解码器和空间感知解码器

当前网络与 VI-SA-BF系统主要有两点不同。首先，整个框架采用端到端训练，而非采用将掩码估计与空间协方差矩阵计算分开的两阶段流程。其次，我们没有将瞬时空间协方差矩阵（ISCM）同时用作注意力机制的查询和键输入，而是引入了专门的空间感知解码器，用于生成独立的特征表征，以进行注意力权重估计。在这两种

然而，在这些方法中，瞬时空间协方差矩阵被用作注意力层的值输入，这与 [46] 中建议的 SCM 公式一致，如下所示：

$$\tilde{\mathbf{\Phi}}_X(t,f)=\sum_{t'=1}^{T}c(t,t')\mathbf{\Psi}_X(t',f)\in\mathbb{C}^{M\times M},\tag{4}$$

$$\tilde{\mathbf{\Phi}}_N(t,f)=\sum_{t'=1}^{T}c(t,t')\mathbf{\Psi}_N(t',f)\in\mathbb{C}^{M\times M},\tag{5}$$

其中 和 表示时频单元处随时间变化的语音和噪声空间协方差矩阵。 表示与以下帧的贡献相关联的加权系数，用于估计当前帧处的空间协方差矩阵。同一频率单元处第帧的瞬时空间协方差矩阵、以及，通过以下方式计算：

$$\mathbf{\Psi}_X(t,f)=M_X(t,f)\mathbf{Y}(t,f)\mathbf{Y}^H(t,f)\in\mathbb{C}^{M\times M}\tag{6}$$

$$\mathbf{\Psi}_N(t,f)=M_N(t,f)\mathbf{Y}(t,f)\mathbf{Y}^H(t,f)\in\mathbb{C}^{M\times M}\tag{7}$$

前述瞬时空间协方差矩阵是通过将语音掩码和噪声掩码应用于多通道输入而获得的。此过程需要一个掩码解码器来生成相应的掩码。该掩码解码器基于 miLAVSE 系统中使用的掩码估计网络。具体而言，通过长短期记忆网络层与共享线性投影相结合来实现视听融合。随后，将输出数据送入两个并行线性层，并通过 sigmoid 激活预测语音掩码和噪声掩码。为避免注意力权重估计受到掩码学习过程的影响，采用独立模块形式的空间感知解码器。该解码器由一个轻量级多层感知机实现，包含两个线性层，中间使用 ReLU 激活，将融合后的视听特征从 512 维投影至 128 维。也就是说，(B, T, 512) → (B, T, 256) → (B, T, 128)。通过将共享表征转换至紧凑的特征空间，空间感知解码器提供独立的空间线索，引导注意力机制捕捉时间依赖关系。此过程对于精确的空间协方差估计至关重要。

D. 基于注意力的SCM计算

基于注意力的SCM的计算可能与 [46]中概述的方法相似。然而，经过进一步分析可以发现，这些方法在复杂度和可学习性方面存在显著差异。这种差异体现在表I所汇总的参数规模和乘加运算次数上。时变SCM使用堆叠式自

- 注意力网络，如 [46] 所述。该网络由类似于Transformer编码器 的层组成，不仅包含注意力机制，还包含额外的前馈子层。相比之下，VI-NBFNet 中的SCM计算模块 不包含任何可学习单元。然而，其具有物理可解释性的结构，结合以波束形成输出为监督信号的端到端训练，使模型能够有效学习空间表征。该计算基于一种注意力机制，其中查询和键之间进行缩放点积运算，随后通过Softmax激活 生成注意力权重。瞬时ISCM 用作值输入，注意力权重则应用于跨时间帧的求和，如下所示：

$$\mathbf{A}=\mathrm{Weight}(\mathbf{Q},\mathbf{K})=\mathrm{Softmax}(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d_k}}),\tag{8}$$

$$\mathbf{Z}=\mathrm{Att}(\mathbf{A},\mathbf{V})=\mathbf{A}\mathbf{V},\tag{9}$$

其中 和 分别表示查询矩阵和键矩阵，它们由空间感知解码器生成， 是查询和键的特征维度，上标 表示转置，而 表示与瞬时空间协方差矩阵对应的值。这一过程得到时变SCM，如图6所示。注意力权重的计算依据公式（8）进行，最终输出Z则根据公式（9）得到。当V表示ISCMs时，**所得的Z即为时变SCM。上述公式在数学上等价于公式（4）和（5）中的公式，因此注意力机制具有可解释性。**

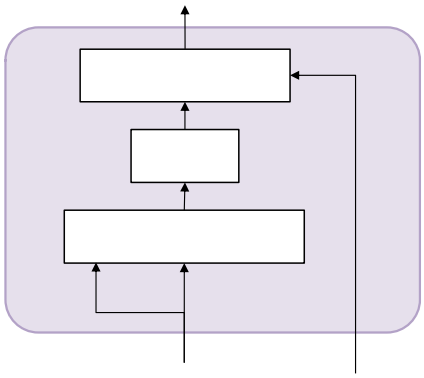


图6. 用于时变SCM估计的注意力层。

表I VI-NBFNET和VI-SA-BF的SCM估计网络的参数规模和乘加运算次数。

	VI-NBFNet	VI-SA-BF
参数（M）	7.15	14.7
乘加运算次数（G/s）	0.4	0.6

E. 最小方差无失真响应（MVDR）波束形成器

获得时变空间协方差矩阵后，将实现MVDR波束形成器 [49]。

$$\mathbf{w}_{MVDR}(t,f)=\frac{\tilde{\mathbf{\Phi}}_N(t,f)^{-1}\tilde{\mathbf{\Phi}}_X(t,f)}{\mathrm{Tr}(\tilde{\mathbf{\Phi}}_N(t,f)^{-1}\tilde{\mathbf{\Phi}}_X(t,f))}\mathbf{u},\tag{10}$$

其中 表示矩阵迹算子。 是与参考麦克风关联的独热向量。因此，波束形成输出为

$$\hat{S}(t,f)=\mathbf{w}_{MVDR}^{H}(t,f)\mathbf{Y}(t,f), \tag{11}$$

其中 表示短时傅里叶变换域中的波束形成信号，作为模型的最终输出。尽管我们的方法与 VI-SA-BF 均将 MVDR 波束成形结果用作网络输出，但所采用的损失函数有所不同。在我们的框架中，目标函数由在时频域计算的均方误差损失和在时域计算的 SNR损失组成。其公式如下：

$$loss=\frac{100}{FT}\sum_{f=1}^{F}\sum_{l=1}^{T}\left(|S(l,f)|-|\hat{S}(l,f)|\right)^{2}+\frac{1}{\mathrm{T}}\sum_{n=1}^{\mathrm{T}}\frac{|s(n)|^{2}}{|s(n)-\hat{s}(n)|^{2}}, \tag{12}$$

其中 和 分别表示参考麦克风在时频域和时域中的真实值（干净）信号。类似地， 和 表示通过波束形成估计得到的相应干净信号。是时频域中的时间帧数， 表示时域中的时间采样点数。 T T

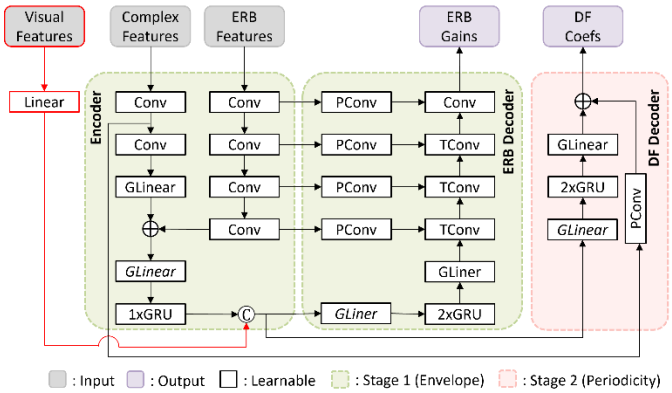


图 7. 视觉信息引导的 DeepFilter 后滤波器的网络架构。

F. 视觉信息增强 DeepFilterNet

除 VI-NBFNet 之外，本文还提出了一种可选的后置滤波器，称为视觉信息增强深度滤波器（VIDF），以进一步提升噪声抑制性能，但代价是计算复杂度略有增加。VIDF 的架构设计如图 7 所示。该后置滤波器的设计受到 DeepFilterNet2 [40]， 的启发，这是一种轻量级深度神经网络，已被证明能够抑制深度神经网络输出端存在的残余噪声。原始 DeepFilterNet2 由一个编码器和一个两阶段解码器组成，其中包括等效矩形带宽（ERB）增益预测阶段和深度滤波阶段。在本研究中，我们利用视觉线索增强输入特征，以提升语音增强性能，尤其是在涉及竞争说话人的场景中。我们采用了与 miLAVSE 类似的融合策略，即在音频编码器之后将音频特征与视觉特征拼接，并通过长短期记忆网络层来捕获时间依赖关系。编码器输出随后与唇读视觉特征拼接，融合后的表征被送入 ERB 解码器和深度滤波解码器。因此，VIDF 的两个解码器均包含门控循环单元层，其工作方式类似于 miLAVSE 中的长短期记忆网络，旨在对由视觉线索引导的语音时间动态进行建模。

IV. 实验

A. 数据集与设置

本研究采用唇读句子 3 -TED（LRS3- TED）数据集 [52], 生成训练和测试数据，该数据集包含 400 小时的 TED 和 TEDx 演讲。在训练和测试阶段，从该语料库中共选取 800 名说话人，其中 50 名作为测试集。为构建多样化的干扰条件，我们纳入了多个数据集。我们使用 LibriSpeech 语料库 [53],中的素材表示竞争说话人，该语料库包含 1000 小时来自公有领域有声读物的英语朗读语音。此外，我们还使用自由音乐档案馆（FMA）数据集 [54] 中的样本引入音乐干扰。对于办公环境中常见的稳态噪声，我们选取了微软可扩展噪声语音数据集（MS-SNSD）[55], 中的若干类别，包括吸尘器、空调和复印机。所有音频信号均以 16 千赫采样。

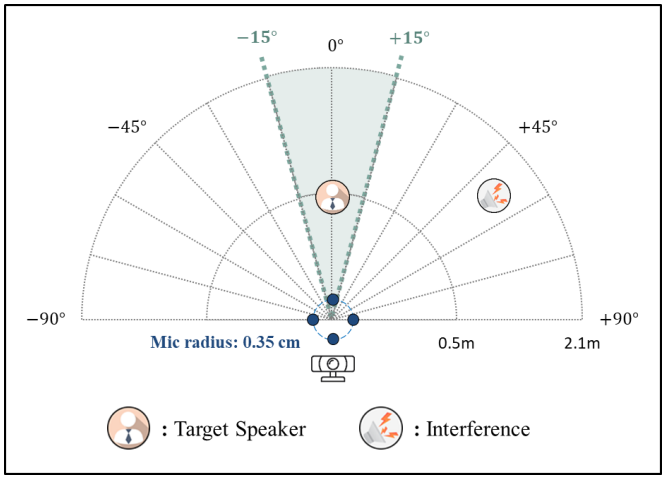


图 8. 训练阶段实验设置示意图。

在训练阶段，采用了半径为 3.5 cm 的四麦克风均匀圆形阵列 ReSpeaker® Mic Array v2.0。为模拟真实的多声源环境，共生成了 30,000 个训练样本和 9,000 个验证样本。每个样本由一段 6 秒的纯净语音与各种类型的干扰混合而成。目标语音与所有干扰按照 –10、–5、0、5、10 和 15 分贝的信干比（SIR）进行混合。此外，还在各通道分别加入信噪比在 30 至 40 分贝之间随机选取的高斯噪声，以模拟真实麦克风录音中的传感器噪声。包括目标说话人和干扰说话人在内的所有声源，均位于阵列前方的一段半圆弧上，距阵列中心 0.5 至 2.1 m。目标说话人的

方位角被限制在 –15° 至 15° 的范围内，以确保目标位于摄像头有效视角之内。在静态条件下，目标比干扰更靠近阵列中心。然而，在动态条件下，目标说话人会在相同的角度范围内沿直线移动。尽管不同样本中的移动速度有所变化，每个样本的 6 秒话语期间均保持恒定速度，以真实地模拟说话人运动。实验设置如图 8 所示。使用镜像源法生成混响时间（RT60）为 200 至 500 ms 的房间脉冲响应（RIR）[56]。随后，将这些 RIR 与纯净声源和干扰信号进行卷积，以生成多通道混响混合信号。

使用训练所用麦克风阵列进行的真实数据录音，是在台湾国立清华大学Engineering Building I的一间会议室内完成的，如图9所示。该房间尺寸为7.0 m × 5.3 m × 2.7 m，RT60约为430 ms。视频使用Apple® iPad Air 5录制。该配置便于在真实混响环境中同步采集视听数据。目标语音由真人说话者录制，或通过扬声器回放录制。干扰信号包括非目标语音的扬声器播放，以及单独的空调系统环境噪声。

模拟测试数据采用与训练设置相同的配置，不同之处在于所有话语均来自未见说话人。共生成1,800个测试样本，其中包括900个含目标说话人的样本和900个含移动目标说话人的样本。这两个数据集分别用于评估系统在静态条件和动态条件下的性能。



图 9. 视听录音实验环境。

所提出的模型和基线方法均采用 25 ms 的帧长和 10 ms 的时间步长，并使用 Adam优化器进行训练。采用了 512 点快速傅里叶变换（FFT）。

B. 评估指标

为评估所提出系统的增强性能，采用了三种客观指标。PESQ 和短时客观可懂度指标用于评估语音质量和有参考评估中的可懂度。相反，在无参考评估中，采用基于 P.835人工评分训练的 DNSMOS 指标 [43], 作为语音质量指标。DNSMOS 包含三个性能指标：语音质量（SIG）、背景噪声质量（BAK）和总体质量（OVRL）。通常，SIG 和 BAK 用于表示语音失真与噪声抑制之间的权衡。

C. 仿真结果

为验证所提出的系统，我们进行了全面评估。根据架构特征和输出格式，将基线系统和所提出的方法分为两组。由于所提出的 VI-NBFNet 会生成波束形成信号，我们选择的第一组基线方法包括 VI-MSE (BF) 和 VI-SA-BF，这两种方法也会生成波束形成输出。第二组基线方法 VI-SSE 和 VI-MSE (BF+PF) 则额外包含一个后置滤波器，用于在所提出的网络与 DF 级联并将其用作后置滤波器时进行基准比较。在确定这两组基线方法后，我们进一步开展了一系列分析，以全面评估所提出的 VI-NBFNet。评估涵盖五个方面：(1) 对采用波束形成技术的多通道视听系统与其单通道对应系统进行比较分析；(2) 考察使用后置滤波器的影响；(3) 对第 (12) 式中定义的所提出的联合损失函数进行消融研究，以验证其贡献；(4) 可视化学习到的注意力权重，以解读空间感知解码器（SAD）的行为；(5) 在各种视觉退化条件下进行鲁棒性评估。这些实验共同帮助我们全面了解所提出的视听波束形成框架的性能、可解释性和鲁棒性。

表 II 所提方法与基于波束形成的基线方法在信干比为 -10、-5、0、5、10 和 15 分贝时的平均 PESQ、短时客观可懂度与 DNSMOS。

	静止目标说话人				
	PESQ	短时客观可懂度	BAK	SIG	OVRL
带噪	1.342	0.659	1.391	1.713	1.308
VI-MSE(BF)	1.588	0.768	1.749	2.144	1.573
VI-SA-BF	1.559	0.768	3.013	2.254	1.732
VI-NBFNet	**1.954**	**0.835**	**3.735**	**2.570**	**2.177**
	移动目标说话人				
	PESQ	短时客观可懂度	BAK	SIG	OVRL
带噪	1.334	0.656	1.383	1.699	1.303
VI-MSE(BF)	1.483	0.736	1.658	1.946	1.484
VI-SA-BF	1.554	0.764	3.027	2.212	1.701
VI-NBFNet	**2.026**	**0.849**	**3.749**	**2.569**	**2.183**

1) VI-NBFNet、VI-MSE (BF) 和 VI-SA- BF 的比较：结果见表 II。对于移动说话人场景，VI-SA-BF 和 VI-NBFNet 在所有评估指标（PESQ、短时客观可懂度和 DNSMOS）上的平均得分均高于 VI-MSE (BF)。这一结果表明，在波束形成中使用时变空间协方差矩阵优于使用时不变空间协方差矩阵，尤其是在

DNSMOS 指标方面。然而，对于静态说话人场景，VI-SA-BF 的 PESQ 和短时客观可懂度得分略低于 VI-MSE (BF)。这种差异可能是由于在低信干比水平下时变空间协方差矩阵估计不准确所致，与纯净语音相比，这可能会引入轻微失真。

值得注意的是，如表 II 所示，在静止和移动说话人场景下，所提出的 VI-NBFNet 在所有评估指标和信干比水平上均优于所有基于波束形成的基线方法。此外，DNSMOS 的提升，尤其是 SIG 和 BAK 的提升，表明所提出的模型在提升目标语音质量和抑制干扰方面效果更佳。值得注意的是，在低信干比水平（如 –10 和 –5 分贝）下，PESQ 和短时客观可懂度均有显著提升。这一发现表明，与依赖两阶段掩码估计和基于自注意力的波束成形的 VI-SA-BF 相比，VI-NBFNet 所采用的联合学习策略在不利声学条件下表现出令人满意的泛化能力和鲁棒性。

2）VI-NBFNet、VI-NBFNet-DF、VI-SSE 和 VI-MSE (BF+PF) 的比较：结果如表 III 所示。通过查看汇总结果可以发现，与单通道基线方法（VI-SSE）相比，所有多通道方法在各项性能指标上均表现更优，尤其是在 PESQ、短时客观可懂度和 SIG 方面。这一观察凸显了空间线索的重要性；单通道系统缺少这些线索，因此更容易受到语音失真的影响。

表 III 所提方法、VI-SSE 和 VI-MSE 在信干比为 -10、-5、0、5、10 和 15 分贝时的平均 PESQ、短时客观可懂度和 DNSMOS。

	静止目标说话人				
	PESQ 短时客观可懂度		BAK	SIG	OVRL
带噪	1.342	0.659	1.391	1.713	1.308
VI-SSE	1.558	0.750	3.129	2.448	2.014
VI-MSE (BF+PF)	1.798	0.811	3.355	2.748	2.301
VI-NBFNet	1.954	0.835	**3.735**	2.570	2.177
VI-NBFNet-DF	**2.015**	**0.841**	3.474	**2.813**	**2.399**
	移动目标说话人				
	PESQ	STOI	BAK	SIG	OVRL
带噪	1.334	0.656	1.383	1.699	1.303
VI-SSE	1.553	0.746	3.146	2.439	2.010
VI-MSE (BF+PF)	1.651	0.783	3.251	2.571	2.134
VI-NBFNet	2.026	0.849	**3.749**	2.569	2.183
VI-NBFNet-DF	**2.088**	**0.855**	3.597	**2.838**	**2.438**

在多通道方法中，VI-MSE（BF+PF）已被证明能够取得高于 VI-NBFNet 的 SIG 分数，这表明波束形成与后置滤波器的结合对于改善语音质量非常有效。该优势在涉及静止说话人的场景中尤为明显，因为固定波束形成非常适用于此类场景。然而，在不使用后置滤波器的情况下，VI-NBFNet 模型在大多数评估指标上均表现优于 VI-MSE（BF+PF）模型。VI-NBFNet 取得了最高的 BAK 分数，展现出强大的噪声抑制能力。尽管在静止说话人条件下，VI-NBFNet 的 SIG 略低，但其增强性能稳定，并且对移动目标说话人具有良好的鲁棒性。

此外，我们评估了在 VI-NBFNet 后应用基于视觉信息的 DeepFilter 后置滤波器的效果。如表 III 所示，SIG 分数显著提高，表明感知语音质量得到改善。然而，BAK 略有下降，这可能是后置滤波器引入残留伪影所致。在这一特定情况下，尽管大多数性能指标均有所改善，但后置滤波器会带来更大的内存存储需求和计算开销。因此，后续实验未采用这一配置。

3）损失函数消融研究：为进一步验证式（12）中提出的联合损失函数的有效性，该函数结合了时频域均方误差损失和时域 SNR损失。我们开展了消融研究，将其与两种替代训练目标进行比较：仅使用均方误差损失和仅使用 SNR损失。

如表 IV 所示，所提出的联合损失在所有指标上始终取得最佳的整体性能。仅使用均方误差损失的模型倾向于保留频谱结构，但会产生可闻的时间伪影；而仅使用 SNR 损失的模型则以轻微频谱失真为代价保留时间保真度。通过联合优化这两个目标，网络在频谱准确度与感知质量之间取得了更好的平衡，从而生成更清晰、听感更自然的增强语音。此外，结合两个域的互补线索，可在动态声学条件下提升鲁棒性与泛化能力。

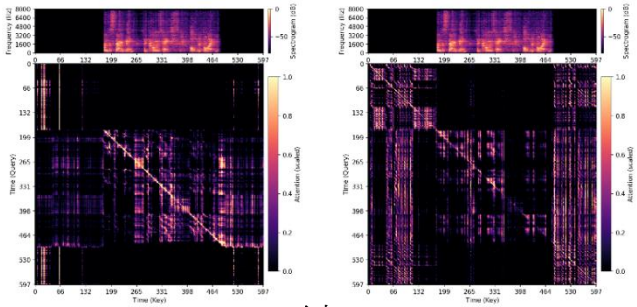
表 IV 所提出的 VI-NBFNET 在信干比为 -10、-5、0、5、10 和 15 分贝条件下，分别使用所提出的联合损失、仅均方误差损失和仅 SNR 损失进行训练时的平均 PESQ、短时客观可懂度与 DNSMOS。移动目标说话人

	PESQ 短时客观可懂度	背景噪声质量		SIG	OVRL
带噪	1.334	0.656	1.383 1.699	1.303	
所提出的联合方法 loss	2.026	0.849	3.749 2.569		**2.183**
仅使用均方误差损失	2.022	0.849	3.714 2.390	1.987	
仅使用SNR损失	1.975	0.847	3.944 2.492		2.122

4) 注意力权重的可视化：图10可视化了在移动说话人场景中估计时变空间协方差矩阵时SAD所学到的注意力权重，直观展示了注意力机制如何运作，并动态响应目标说话人位置的变化。

如图10(a)所示，用于语音空间协方差矩阵计算的注意力权重呈现出与语音活动片段相对应的强烈对角能量集中。这表明该模型有效地强调了语音活动区域周围的帧，从而在估计的语音

空间协方差矩阵中保持时间一致性与连续性。相比之下，对于图10(b)所示的噪声空间协方差矩阵计算，注意力模块主要关注语音非活动区域，例如话语起始与结束。 这一观察表明，所提出的注意力机制会在语音活动片段与语音非活动片段之间动态调整其关注焦点，从而能够在动态条件下准确估计语音和噪声空间协方差矩阵。

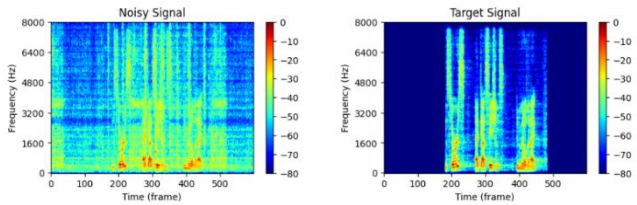


(a) (b)

图10. 移动说话人场景中SAD注意力权重的可视化。（a）用于语音空间协方差矩阵计算的注意力权重，（b）用于噪声空间协方差矩阵计算的注意力权重。

5）视觉退化鲁棒性评估：通过模拟实际视听场景中常见的三种视觉退化类型——部分遮挡、马赛克遮挡和低分辨率退化，对所提出的VI-NBFNet在具有挑战性的视觉条件下的鲁棒性进行了检验。在部分遮挡条件下，在嘴部区域覆盖黑色矩形块，以模拟佩戴口罩或部分遮挡。在马赛克遮挡条件下，将嘴部区域进行像素化处理，以表示视觉失真输入。在低分辨率条件下，将每个视频帧下采样至原始宽度和高度的30%，随后再上采样至原始尺寸，从而产生模糊的嘴唇区域，用于表示低质量视觉输入。

图11展示了具有代表性的示例及其对应的增强频谱图，定量结果则汇总于表 V中。尽管视觉质量有所下降，VI-NBFNet仍保持了稳定的性能，在所有测试条件下，PESQ/短时客观可懂度的差异均低于0.05，OVRL的差异均低于0.1。这些结果表明，所提出的视听融合策略对中等程度的视觉失真具有很强的韧性，因为该模型主要利用时间上保持一致的唇动线索，而非细粒度的像素级细节。此外，如图11所示，与存在遮挡、马赛克和低分辨率视频对应的增强频谱图仍保留了目标语音的谐波结构和共振峰轨迹，证实VI-NBFNet能够有效泛化至视觉退化输入，同时不会造成显著的性能损失。



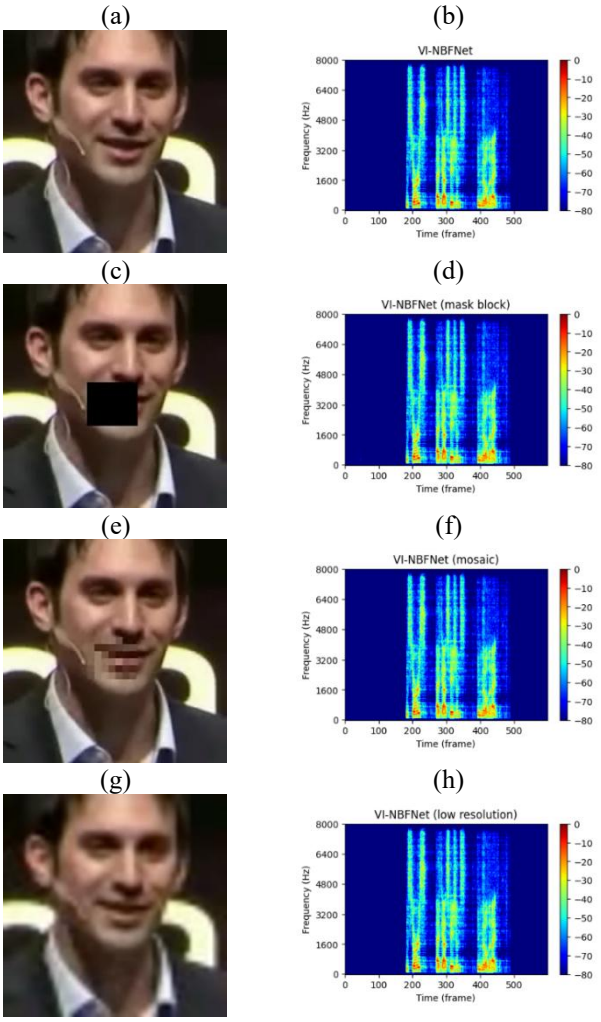


图 11. VI-NBFNet 在视觉退化条件下的鲁棒性评估。(a) 带噪输入频谱图，(b) 干净目标语音频谱图，(c) 原始视频帧，以及 (d) VI-NBFNet 生成的增强频谱图；(e) 部分嘴部遮挡的视频，以及 (f) 其增强频谱图；(g) 马赛克遮挡的视频，以及 (h) 其增强频谱图；(i) 低分辨率视频输入，以及 (j) 其增强频谱图。

表 V VI-NBFNet 模型在视觉退化数据条件下的性能比较，包括掩膜遮挡、马赛克和低分辨率。PESQ 短时客观可懂度 背景噪声质量 SIG OVRL

带噪	1.198	0.771	1.424	2.137	1.414
无损	**2.184**	**0.957**	**4.007**	**2.850**	**2.624**
Mask 遮挡	2.167	0.957	3.928	2.807	2.549
马赛克	2.213	0.958	3.974	2.843	2.603
Low 分辨率	2.167	0.958	3.974	2.843	2.603

D. 实验结果

除前述模拟实验外，我们还进行了真实数据实验。如第 IV.A 节所述，录制评估中使用的目标说话人信号

通过两种方式生成：通过扬声器回放，以及由真人说话者进行现场语音录制。扬声器回放了来自模拟测试集的预录干净语句，这对于音视频信号之间的同步至关重要。现场说话人则用于更真实的场景，使我们能够通过人工后期编辑和轻微的头部运动确保视听同步。采用扬声器重放和现场说话人录音得到的频谱图分别如图 12 和图 13 所示。所有实验的评估指标均值汇总于表 VI。请注意，由于现场说话人场景缺少真实干净语音，因此评估过程中仅使用了非侵入式DNSMOS指标。

表 VI 使用 DNSMOS 对真实世界录音进行评估。

	BAK	SIG	OVRL
带噪	1.904	1.382	1.378
VI-SSE	2.592	2.513	1.972
VI-MSE (BF)	1.761	2.254	1.732
VI-MSE（BF+PF）	2.567	2.679	1.952
VI-SA-BF	2.109	1.972	1.672
VI-NBFNet	**2.743**	**2.829**	**2.193**

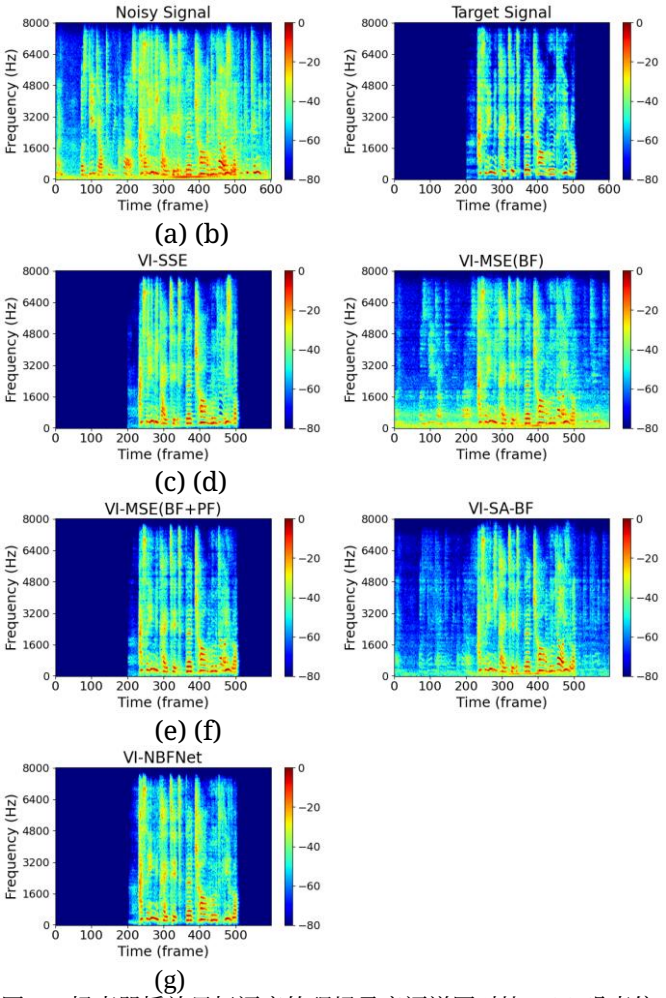


图 12. 扬声器播放目标语音的现场录音语谱图对比。(a) 噪声信号，(b) 干净目标语音信号，(c) VI-SSE 增强信号，(d) VI-MSE 经波束形成增强的信号，(e) VI-MSE 经波束形成和后滤波处理的增强信号，(f) VI-SA-BF 增强信号，以及 (g) VI-NBFNet 增强信号。

如表 VI 所示，在所有评估方法中，VI-NBFNet 的 DNSMOS 最高。具体而言，其 OVRL 分数显著较高，这表明与其他方法相比，VI-NBFNet 在语音质量与噪声抑制之间实现了更优的平衡。此外，VI-SSE 的 SIG 分数低于 VI-MSE (BF+PF) 和 VI-NBFNet，这与仿真结果的发现一致。这一发现进一步支持了单通道方法往往难以保留语音质量这一论断。在实际录音环境中，与 VI-MSE (BF+PF) 相比，VI-NBFNet 的 SIG 分数更高。这一结果与仿真结果中观察到的趋势相反。这表明所提出的方法在真实条件下具有更优的增强性能和鲁棒性。

表 VII 使用 Whisper 基础版和 Turbo 自动语音识别模型对真实世界录音进行评估，指标为词错误率（%）。基础版 Turbo

VI-SSE	93%	27%
VI-MSE (BF)	113%	33%
VI-MSE (BF+PF)	58%	13%
VI-SA-BF	75%	35%
VI-NBFNet	50%	**8%**

为了在缺少真实干净语音的情况下进一步评估语音可懂度，我们还使用 Whisper 系列中的两种代表性自动语音识别（ASR）模型 [57] 评估了词错误率（WER）：Whisper-base 和 Whisper-turbo，分别对应中等容量和大容量配置。Turbo 模型是一种经过优化的大规模变体，旨在提高推理效率。如表 VII 所示，所提出的 VI-NBFNet 在所有对比系统中取得了最低的词错误率，使用 Whisper-base 时为 50%，使用 Whisper-turbo 时为 8%。从基础版模型到 Turbo 模型的显著提升表明，VI-NBFNet 生成的增强语音保留了清晰的语音内容，使更大容量的自动语音识别模型能够更好地利用其声学特征。结合表 VI 中的 DNSMOS 结果（VI-NBFNet 在所有指标上均取得最高分），词错误率分析进一步证实，所提出的模型在真实录音环境中能够生成可懂度高且听感自然的语音。

正如预期，图 13 显示，与扬声器播放时使用的模拟唇部视频（图 12）相比，使用现场录制视频时性能较差，这是由于视觉训练数据存在差异。训练期间，从 LRS3 视频集中选取唇部图像，方法是从每段话语中截取固定时长的片段。值得注意的是，在处理语音停顿期间的面部图像时，采用了专门的处理方法。具体而言，反复使用初始和末尾的视频帧，填补没有语音的时段。这种方法确保了视听连续性，并

在无语音时段保持同步。这对于视听网络准确地将唇部运动与语音相关信息联系起来，以进行模型训练和推理至关重要。相反，现场录音即使在语音静默时段也包含连续的面部动作，而模型并未针对这种情况进行训练。如图 13 所示，录制的人脸图像比训练集中的图像小。这导致裁剪后的唇部图像分辨率较低，并使预训练视听语音识别模型的性能下降。造成性能下降的另一个因素是录制期间存在无法控制的环境干扰，例如过往车辆噪声。不过，该模型适用于涉及两个定向声源的场景。尽管存在这些挑战，VI-NBFNet 仍是图 13 中唯一能够有效抑制非目标语音的方法。这表明它在真实环境中具有更强的鲁棒性和泛化能力。

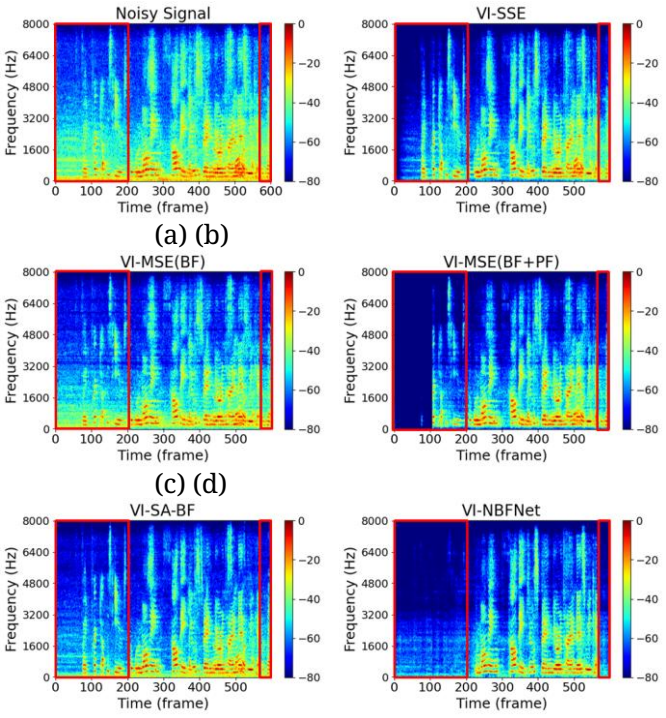
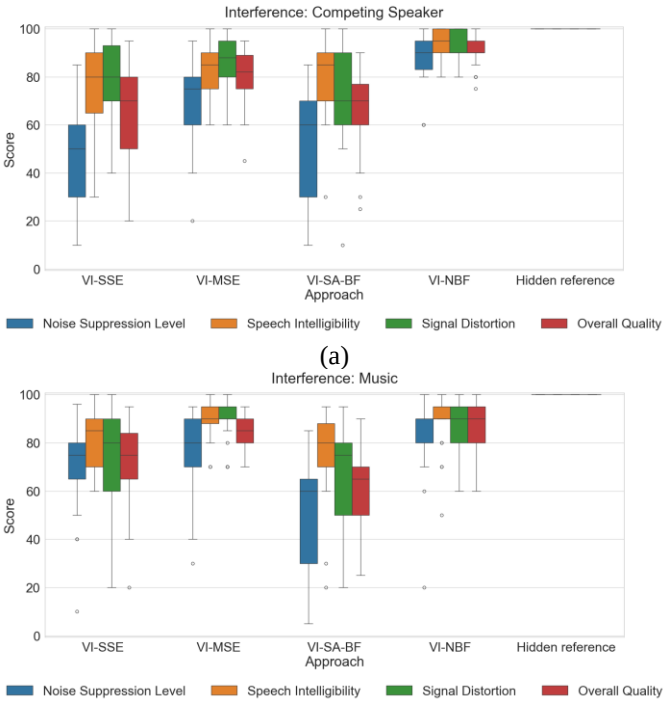


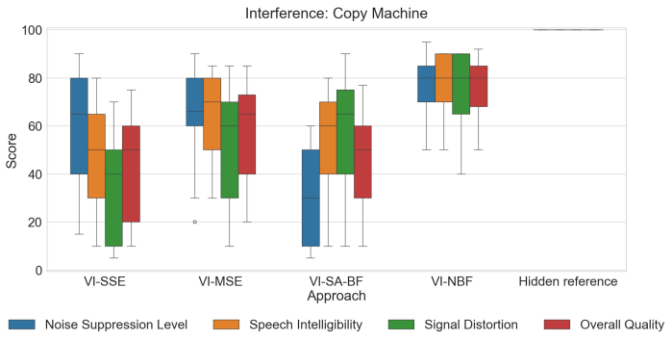
图 13. 现场录制目标语音的现场录音频谱图对比。(a) 噪声信号，(b) VI-SSE 增强信号，(c) 采用波束形成的 VI-MSE 增强信号，(d) 采用波束形成和后滤波的 VI-MSE 增强信号，(e) VI-SA-BF 增强信号，(f) VI-NBF 增强信号。

E. 听力测试

除上述客观评价外，我们还采用基于带隐藏参考和锚点的多重刺激（MUSHRA）流程 [58] 的主观听力测试，对所提出的 VI-NBFNet 与基线方法进行了主观评价。共有 21 名参与者参加了主观评价。参与者通过评定四项感知属性来评价语音增强（SE）的感知性能。考察的属性包括：噪声抑制水平、语音可懂度，信号失真和总体质量。噪声抑制水平用于评估干扰的减少程度，而语音可懂度指数则用于评估增强语音的清晰度和可理解性。通过信号失真指数评估增强过程中引入的音频伪影。参与者对总体质量的感知偏好与参考信号进行了比较。各项属性均采用 1 至 100 的连续量表进行评分。为确保结果可靠，若参与者对隐藏参考的评分低于 90，或对锚点的评分高于 10，则排除其回答。

在每个测试案例中，参与者都会听到四种处理后的信号，包括所提出的方法 VI-NBFNet 和三种基线方法：VI-SSE、VI-MSE (BF+PF) 以及 VI-SA-BF。此外，还包含一段与真实干净信号相对应的隐藏参考信号。我们还将截止频率为 2 kHz 和 4 kHz 的低通滤波参考信号用作锚点。为确保评估全面，听力测试采用了三种真实场景中的干扰信号：竞争说话人、音乐和复印机噪声。这些信号也用于模型训练。选择这些案例是为了代表三种常见的噪声类型：类语音干扰、音乐信号和背景机器噪声。





(c)

图 14. (a) 竞争说话人、(b) 音乐和 (c) 复印机的听力测试结果箱线图。每个箱体中的中线表示中位数，箱体的下界和上界分别表示第 25 个百分位数（Q1）和第 75 个百分位数（Q3）。须线延伸至非离群值范围内的最大值和最小值，任何离群值则表示为超出这些界限的单个点。

为评估听力测试中各增强方法结果差异的统计显著性，我们进行了一系列统计分析。首先，使用 Shapiro–Wilk检验 [59]评估各方法在每种干扰条件（竞争说话人、音乐和复印机）下评分的正态性。分析结果表明，所有 p值均大于 0.05，说明所有组均满足正态性假设。接下来，使用 Levene检验 [60] 检验方差齐性。在竞争说话人和音乐条件下，p值均小于 0.05，表明方差相等的假设不成立。因此，我们进行了 Welch方差分析 [61]。结果显示，这两种条件下的增强方法之间均存在显著差异：竞争说话人条件下，$F(3, 40.35) = 23.11$，$p < 0.001$；音乐条件下，$F(3, 42.34) = 12.55$，$p < 0.001$。相比之下，在复印机条件下，Levene检验得出的 p值大于 0.05，证实方差齐性成立。因此，我们进行了标准的单因素方差分析 [62]，结果显示各方法之间存在显著差异：$F(4, 100) = 41.45$，$p < 0.001$。鉴于三种条件下均存在显著的统计差异，我们随后进行了事后两两比较，结果汇总于表 VIII。

表 VIII 竞争说话人 (A)、音乐 (B) 和复印机噪声 (C) 的总体质量 p 值。p 值小于 0.05 表示具有较大的统计差异。

	VI-SSE	VI-MSE (BF+PF)	VI-SA-BF	VI-NBFNet
VI-SSE		0.017	1.000	0.000
VI-MSE (BF+PF)			0.036	0.005
VI-SA-BF				0.000
VI-NBFNet				

(a)

	VI-SSE	VI-MSE (BF+PF)	VI-SA-BF	VI-NBFNet
VI-SSE		0.024	0.799	0.009
VI-MSE (BF+PF)			0.000	1.000
VI-SA-BF				0.000
VI-NBFNet				

(b)

	VI-SSE	VI-MSE (BF+PF)	VI-SA-BF	VI-NBFNet
VI-SSE		0.022	0829	0.000
VI-MSE (BF+PF)			0.169	0.024
VI-SA-BF				0.000
VI-NBFNet				

(c)

听力测试结果以箱线图的形式呈现在图14中。在所有三种干扰类型下，与基线方法相比，所提出的 VI-NBFNet 在大多数主观属性上表现更优。值得注意的是，在所有条件下，VI-NBFNet 的总体质量中位数评分均为最高，且箱体较窄，表明听众的评分更一致。尤其是在竞争说话人和复印机噪声条件下，VI-NBFNet 明显优于所有基线方法，包括 VI-SSE、VI-MSE 和 VI-SA-BF，所有两两比较的 p 值均小于 0.05。在音乐干扰情况下，VI-NBFNet 的表现优于 VI-SA-BF 和 VI-SSE（p 值 < 0.05）。然而，与 VI-MSE 相比，结果未达到统计显著性（p 值 = 1）。这可能是因为该场景中的信干比较高，使 VI-MSE 在感知上与 VI-NBFNet 相当。在所有噪声条件下，单通道方法 VI-SSE 的评分始终最低，这可能是由于其感知失真更大。这些发现凸显了在语音增强中，将阵列波束形成所提供的空间线索与时频线索相结合，在不影响音质的情况下能够有效提升性能。

五、结论

在本研究中，我们提出了一种新颖的视觉信息辅助语音增强网络 VI-NBFNet，将基于注意力机制的波束形成整合到端到端学习架构中。与分别估计掩码和波束形成权重的传统两阶段流程不同，VI-NBFNet 在统一框架内同时学习这两个组件。这种方法能够有效抑制干扰。所提出的系统将空间模态与视听模态相结合，并利用注意力机制进行空间协方差矩阵估计，即使在动态声学环境中也能实现卓越的波束形成性能。对 VI-NBFNet 和单通道端到端系统 VI-SSE 的比较表明，前者在语音失真极小的情况下，噪声抑制效果更佳。总之，所提出的方法提供了

一种视听语音增强解决方案，在静态和移动说话人场景中均展现出出色的性能和鲁棒性。未来，我们将考虑采用以感知为导向的损失函数，例如尺度不变信号失真比或基于DNSMOS的损失，以使训练目标更好地契合人类听觉。我们正在扩展训练数据，纳入更多样且更贴近真实条件的场景，例如包含两名以上说话人或移动干扰源的混合语音，以期进一步提升所提系统的泛化能力。

参考文献

[1] W. Rao、Y. Fu、Y. Hu、X. Xu、Y. Jv、J. Han、Z. Jiang、L. Xie、Y. Wang、S. Watanabe、Z.-H. Tan、H. Bu、T. Yu 和 S. Shang，“会议语音挑战赛：面向视频会议的远场多通道语音增强”，载于 IEEE 自动语音识别与理解研讨会（ASRU）论文集，哥伦比亚卡塔赫纳，2021年12月，第 679–686 页。[2] H. K. Maganti、D. Gatica-Perez 和 I. McCowan，“利用视听传感器阵列进行会议语音增强与识别”，《IEEE 音频、语音与语言处理汇刊》，第 15 卷，第 8 期，第 2257–2269 页，2007年11月。[3] I. Fedorov、M. Stamenovic、C. Jensen、L.-C. Yang、A. Mandell、Y. Gan、M. Mattina 和 P. N. Whatmough，“TinyLSTMs：面向助听器的高效神经网络语音增强”，arXiv 预印本 arXiv:2005 11138，2020年。。[4] G. Park、W. Cho、K. S. Kim 和 S. Lee，“基于深度学习的环境噪声助听器语音增强”，《应用科学》，第 10 卷，第 17 期，第 6077 页，2020年。[5] R. Haeb-Umbach、S. Watanabe、T. Nakatani、M. Bacchiani、B. Hoffmeister、M. L. Seltzer、H. Zen 和 M. Souden，“数字家庭助理的语音处理：结合信号处理与深度学习技术”，《IEEE 信号处理杂志》，第 36 卷，第 6 期，第 111–124 页，2019年11月。[6] H. Sato、T. Ochiai、M. Delcroix、K. Kinoshita、N. Kamo 和 T. Moriya，“学习增强与否：基于神经网络的增强信号和观测信号切换，用于重叠语音识别”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，新加坡，2022年5月，第 6287–6291 页。[7] Y. Masuyama、X. Chang、S. Cornell、S. Watanabe 和 N. Ono，“语音识别、去混响、波束形成与自监督学习表征的端到端集成”，载于 IEEE 口语技术研讨会（SLT）论文集，卡塔尔多哈，2023年1月，第 260–265 页。[8] Z. Q. Wang、P. Wang 和 D. Wang，“用于单通道和多通道语音增强及鲁棒自动语音识别的复数频谱映射”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 28 卷，第 1778–1787 页，2020年。[9] N. Dehak、P. J. Kenny、R. Dehak、P. Dumouchel 和 P. Ouellet，“用于说话人验证的前端因子分析”，《IEEE 音频、语音与语言处理汇刊》，第 19 卷，第 4 期，第 788–798 页，2011年5月。[10] Y. Lei、N. Scheffer、L. Ferrer 和 M. McLaren，“一种使用音素感知深度神经网络进行说话人识别的新方案”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，意大利佛罗伦萨，2014年5月，第 1695–1699 页。[11] E. Variani、X. Lei、E. McDermott、I. L. Moreno 和 J. Gonzalez-Dominguez，“用于小型化文本相关说话人验证的深度神经网络”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，意大利佛罗伦萨，2014年5月，第 4052–4056 页。[12] L. Wan、Q. Wang、A. Papir 和 I. L. Moreno，“ 用于说话人验证的广义端到端损失 ”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，加拿大阿尔伯塔省卡尔加里，2018年4月，第 4879–4883 页。[13] D. Snyder、D. Garcia-Romero、G. Sell、D. Povey 和 S. Khudanpur，“X 向量：用于说话人识别的鲁棒深度神经网络嵌入”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，加拿大阿尔伯塔省卡尔加里，2018年4月，第 5329–5333 页。[14] H. Taherian、Z. Q. Wang、J. Chang 和 D. Wang，“基于单通道和多通道语音增强的鲁棒说话人识别”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 28 卷，第 1293–1302 页，2020年。

[15] S. E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang，“个性化语音增强：新模型与全面评估”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，新加坡，2022年5月，第356–360页。[16] J. Xiong, Y. Zhou, P. Zhang, L. Xie, W. Huang, and Y. Zha，“Look& listen：用于主动说话人检测和语音增强的多模态相关性学习”，《IEEE 多媒体汇刊》，第25卷，第5800–5812页，2022年。[17] D. Stewart, R. Seymour, A. Pass, and J. Ming，“噪声音视频条件下的鲁棒音视频语音识别”，《IEEE 控制论汇刊》，第44卷，第2期，第175–184页，2014年2月。[18] A. Gabbay, A. Shamir, and S. Peleg，“视觉语音增强”，arXiv预印本 arXiv:1711.08789，2017年。[19] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida，“MMTM：用于CNN融合的多模态迁移模块”，载于 IEEE/CVF 计算机视觉与模式识别会议（CVPR）论文集，2020年6月，第13289–13299页。[20] Y. Xu, M. Yu, S. X. Zhang, L. Chen, C. Weng, J. Liu, and D. Yu，“用于目标语音分离的神经时空波束形成器”，arXiv预印本 arXiv:2005.03889，2020年。[21] K. Tan, Y. Xu, S. X. Zhang, M. Yu, and D. Yu，“采用两阶段多模态网络的音视频语音分离与去混响”，《IEEE 信号处理专题期刊》，第14卷，第3期，第542–553页，2020年4月。[22] J. C. Hou, S. S. Wang, Y. H. Lai, Y. Tsao, H. W. Chang, and H. M. Wang，“采用多模态深度卷积神经网络的音视频语音增强”，《IEEE 新兴计算智能专题汇刊》，第2卷，第2期，第117–128页，2018年4月。[23] S. Y. Chuang, Y. Tsao, C. C. Lo, and H. M. Wang，“轻量音视频语音增强”，arXiv预印本 arXiv:2005.11769，2020年。[24] S. Y. Chuang, H. M. Wang, and Y. Tsao，“改进的轻量音视频语音增强”，《IEEE/ACM 音频、语音与语言处理汇刊》，第30卷，第1345–1359页，2022年。[25] F. Schroff, D. Kalenichenko, and J. Philbin，“FaceNet：用于人脸识别和聚类的统一嵌入”，载于 IEEE 计算机视觉与模式识别会议（CVPR）论文集，美国马萨诸塞州波士顿，2015年6月，第815–823页。[26] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein，“在鸡尾酒会上看着听：用于语音分离的说话人无关音视频模型”，arXiv预印本 arXiv:1804.03619，2018年。[27] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani，“多模态SpeakerBeam：利用音视频说话人线索进行单通道目标语音提取”，载于 Interspeech 论文集，奥地利格拉茨，2019年9月，第2718–2722页。[28] T. Stafylakis and G. Tzimiropoulos，“结合残差网络与长短期记忆网络进行唇读”，载于 Interspeech 论文集，瑞典斯德哥尔摩，2017年8月，第3652–3656页。[29] T. Afouras, J. S. Chung, and A. Zisserman，“The Conversation：深度音视频语音增强”，载于 Interspeech 论文集，印度海得拉巴，2018年9月，第3244–3248页。[30] C. Li and Y. Qian，“利用注意力机制进行深度音视频语音分离”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，西班牙巴塞罗那，2020年5月，第7314–7318页。[31] M. Sadeghi, S. Leglaive, X. Alameda-Pineda, L. Girin, and R. Horaud，“使用条件变分自编码器进行音视频语音增强”，《IEEE/ACM 音频、语音与语言处理汇刊》，第28卷，第1788–1800页，2020年。[32] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic，“端到端音视频语音识别”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，加拿大阿尔伯塔省卡尔加里，2018年4月，第6548–6552页。[33] Z. Aldeneh, A. P. Kumar, B. J. Theobald, E. Marchi, S. Kajarekar, D. Naik, and A. H. Abdelaziz，“论视觉线索在视听语音增强中的作用”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，加拿大安大略省多伦多，2021年6月，第8423–8427页。[34] R. Gu, S. X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu，“多模态多通道目标语音分离”，《IEEE 信号处理专题期刊》，第14卷，第3期，第530–541页，2020年4月。[35] T. Afouras, J. S. Chung, and A. Zisserman，“我的嘴唇被遮挡：通过遮挡进行音视频语音增强”，arXiv预印本 arXiv:1907.04975，2019年。[36] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani，“多模态SpeakerBeam：利用音视频说话人线索进行单通道目标语音提取”，载于 Interspeech 论文集，奥地利格拉茨，2019年9月，第2718–2722页。

[37] A. Adeel, M. Gogate, and A. Hussain，“基于上下文深度学习的视听切换：用于真实环境中的语音增强”，《Information Fusion》，第59卷，第163–170页，2020年。[38] G. Li, J. Deng, M. Geng, Z. Jin, T. Wang, S. Hu, M. Cui, H. Meng, and X. Liu，“音视频端到端多通道语音分离、去混响与识别”，《IEEE/ACM 音频、语音与语言处理汇刊》，第31卷，第2707–2723页，2023年。[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin，“Attention Is All You Need”，载于《神经信息处理系统进展》（NeurIPS），第30卷，2017年。[40] H. Schröter, A. Maier, A. N. Escalante-B. and T. Rosenkranz，“DeepFilterNet2：面向嵌入式设备全频带音频实时语音增强”，载于国际声学信号增强研讨会（IWAENC）论文集，德国班贝格，2022年9月，第1–5页。[41] I.-T. Recommendation，“语音质量感知评估（PESQ）：用于窄带电话网络和语音编解码器端到端语音质量评估的客观方法”，ITU-T P. 862号建议书，2001年。[42] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen，“一种用于预测时频加权带噪语音可懂度的算法”，《IEEE/ACM 音频、语音与语言处理汇刊》，第19卷，第7期，第2125–2136页，2011年。[43] C. K. A. Reddy, V. Gopal, and R. Cutler，“DNSMOS P.835：一种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 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，新加坡，2022年5月，第886–890页。[44] J. Heymann, L. Drude, and R. Haeb-Umbach，“基于神经网络的声学波束成形频谱掩码估计”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，中国上海，2016年3月，第196–200页。[45] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux，“使用单通道掩码预测网络改进MVDR波束成形”，载于 Interspeech 论文集，美国加利福尼亚州旧金山，2016年9月，第1981–1985页。[46] T. Ochiai, M. Delcroix, T. Nakatani, and S. Araki，“针对移动说话人的基于掩码的神经波束成形及基于自注意力的跟踪”，《IEEE/ACM 音频、语音与语言处理汇刊》，第31卷，第835–848页，2023年。[47] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic，“Auto-AVSR：采用自动标签的视听语音识别”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，希腊罗得岛，2023年6月，第1–5页。[48] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux，“利用深度循环神经网络进行相位敏感且识别增强的语音分离”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，澳大利亚昆士兰州布里斯班，2015年4月，第708–712页。[49] Y. Luo and N. Mesgarani，“Conv-TasNet：超越理想时频幅度掩蔽的语音分离”，《IEEE/ACM 音频、语音与语言处理汇刊》，第27卷，第8期，第1256–1266页，2019年8月。[50] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen，“MobileNetV2：倒置残差与线性瓶颈”，载于 IEEE 计算机视觉与模式识别会议（CVPR）论文集，美国犹他州盐湖城，2018年6月，第4510–4520页。[51] M. Souden, J. Benesty, and S. Affes，“关于用于噪声抑制的最优频域多通道线性滤波”，《IEEE 音频、语音与语言处理汇刊》，第18卷，第2期，第260–276页，2010年2月。[52] T. Afouras, J. S. Chung, and A. Zisserman，“LRS3-TED：用于视觉语音识别的大规模数据集”，arXiv预印本 arXiv:1809 00496，2018年。。[53] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur，“LibriSpeech：基于公有领域有声读物的自动语音识别语料库”，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，澳大利亚布里斯班，2015年4月，第5206–5210页。[54] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson，“FMA：用于音乐分析的数据集”，arXiv预印本 arXiv:1612.01840，2016年。[55] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke，“可扩展噪声语音数据集与在线主观测试框架”，arXiv预印本 arXiv:1909.08050，2019年。[56] E. A. Lehmann and A. M. Johansson，“采用镜像声源模型模拟的房间脉冲响应中的能量衰减预测”，《美国声学学会期刊》，第124卷，第1期，第269–277页，2008年。[57] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever，“通过大规模弱监督实现鲁棒语音识别”，载于 ICML 论文集，2023年，第28492–28518页。

[58] ITU-R 建议书，《中间音质主观评价方法（MUSHRA）》，国际电信联盟，BS. 1534-1，日内瓦，瑞士，2001年。[59] S. S. Shapiro 和 M. B. Wilk，《正态性方差分析检验（完整样本）》，Biometrika，第 52 卷，第 3–4 期，第 591–611 页，1965 年。[60] J. L. Gastwirth、Y. R. Gel 和 W. Miao，《Levene 方差齐性检验对统计理论与实践的影响》，Statistical Science，第 24 卷，第 3 期，第 343–360 页，2009 年。[61] B. L. Welch，《多个均值的比较：一种替代方法》，Biometrika，第 38 卷，第 3–4 期，第 330–336 页，1951 年。[62] L. Ståhle 和 S. Wold，《方差分析（ANOVA）》，Chemometrics and Intelligent Laboratory Systems，第 6 卷，272，1989 年 259 页。–。