基于掩蔽的空间音频语音增强：信号表示与通道特定掩蔽的影响

Sheli Hendel 和 Boaz Rafaely

内盖夫本-古里安大学电气与计算机工程学院，以色列贝尔谢巴 84105

Dorothea Kolossa

柏林工业大学 以人为中心的信息技术 柏林 10623，德国

arXiv:2611.32910v1 [eess.AS] 24 Nov 2026

摘要

空间音频语音增强需要在提升语音质量和可懂度的同时，保留声学场景的空间特性。时频掩蔽是一种有效的语音增强方法，但其性能和空间保真度可能很大程度上取决于应用掩码时所采用的信号表示。本文系统比较了在模拟混响声学场景中，麦克风域、波束形成器域和一阶 Ambisonics 域中的时频掩蔽。研究同时考察了理想比率掩码（IRM）和轻量级的基于深度学习的掩码估计器，并从语音重建、感知质量、可懂度、混响保留和双耳空间线索保留等方面评估其性能。结果表明，波束形成器域中的通道专属掩蔽可获得最高的语音质量和可懂度，而 Ambisonics 域中的通道专属掩蔽更有利于保留残余干扰的空间特性。将空间表示从三阶 Ambisonics 降至一阶，可显著缩小波束形成器域与 Ambisonics 域之间的性能差距，这是因为由此得到的波束形成器空间分辨率较低。此外，在一阶 Ambisonics 通道间使用共享掩码，能够达到接近通道专属掩蔽的语音增强性能，同时只需一个掩码估计网络。这些结果表明，应联合考虑信号表示和掩码策略；同时也说明，一阶 Ambisonics 结合共享掩蔽能够在语音增强性能、空间保真度和计算复杂度之间实现实用的折中。

索引词—语音增强、时频掩蔽、Ambisonics、波束形成器、双耳线索

1. 引言

语音增强旨在通过抑制干扰声，改善噪声环境中的语音质量和可懂度。时频掩蔽已成为应用最广泛的方法之一，尤其是在使用经训练以将语音与背景噪声分离的深度神经网络来估计掩码时 [1–6]。与经典信号处理技术相比，这些方法已显著提升性能；其优化和评估通常采用与语音质量或可懂度相关的目标指标，例如尺度不变信号失真比（SI-SDR）、语音质量感知评估（PESQ）和短时客观可懂度（STOI） [6–12]。尽管通常会使用多通道录音作为输入，但增强后的输出通常是单通道语音信号，学习目标主要侧重于抑制干扰，同时最大限度地提升语音质量和可懂度。

尽管语音质量和可懂度对语音通信至关重要，但沉浸式音频技术日益普及，包括增强现实和虚拟现实（AR 和 VR）、助听以及三维音频再现，这推动了语音增强方法向多通道空间音频的拓展 [13–17]。在这些应用中，增强算法不仅旨在保持语音质量和可懂度，还要保留准确定位、场景感知、听者舒适度以及空间掩蔽释放（SRM）所需的空间信息 [18–21]。然而，掩蔽可能会扭曲通道间空间信息，从而改变空间感知。尽管先前研究已证明在空间音频系统中应用语音增强的可行性，但对于如何有效兼顾高水平的语音增强与高质量的空间再现，尤其是基于掩蔽的方法，仍缺乏深入理解。

近期的研究已开始弥补这一空白。在 [20],中，对波束形成器域和 Ambisonics 域中的掩蔽进行了比较，结果表明，后者能更好地保留空间

本研究由德国研究基金会（DFG）资助——项目编号 KO3434/9-1。

信息，包括目标语音和残余干扰的信息。最近， [22] 系统地比较了应用于麦克风、波束形成器和 Ambisonics 表示的时频理想比率掩码（IRM），揭示了语音增强性能与空间保真度之间截然不同的权衡。AmbiSep [13, 23] 采用基于 Transformer 的神经网络架构，比较了 Ambisonics 通道和波束形成器通道的性能，以联合执行语音分离和降噪，同时保留空间信息。然而，该分析未涵盖麦克风通道，也未涉及空间度量，例如通过重现的双耳信号计算的混响保留度、双耳时间差（ITD）和双耳声级差（ILD）。尽管取得了这些进展，仍存在一个空白：目前尚无全面的分析，能够同时为如何在实现高语音增强性能的同时保持空间保真度提供深入见解和实用设计指导；这包括在理想和实际的基于神经网络的掩码系统中，保留感知到的混响，以及目标声源和干扰声源的定位。

本研究针对这一空白，对不同信号表示下的理想时频掩码与基于深度学习的时频掩蔽进行系统比较，旨在为基于空间音频的语音增强系统提供理论见解和实用设计指导。具体而言，我们比较了轻量级网络估计的时频掩码与理想IRM，并将其应用于三种信号表示：麦克风信号、波束形成器输出和一阶Ambisonics。聚焦一阶Ambisonics，使我们能够在实际沉浸式音频系统中遇到的现实麦克风阵列约束下开展评估。我们使用模拟混响声学场景评估这些方法，考察其语音增强性能、感知质量、可懂度、混响保留和双耳线索保留。

本研究的主要贡献总结如下：(i) 表明与Ambisonics通道和麦克风通道相比，将通道特定掩蔽应用于波束形成器通道最能保留语音质量和可懂度；(ii) 表明与波束形成器通道和麦克风通道相比，将通道特定掩蔽应用于Ambisonics通道最能保留残余干扰的空间属性；(iii) 表明在Ambisonics通道上应用共享掩码，其语音质量和可懂度接近通道特定掩蔽，仅有轻微的性能下降，因此可实现更高效的方案。

2. 背景

2.1. 信号模型

考虑一个混响声学环境，其中包含目标语音声源、干扰源和稳态背景噪声。使用配备 Q 个麦克风的刚性球形麦克风阵列观测该声学场景。在频域中，麦克风信号建模为

$$\mathbf{x}(f) = \mathbf{x}^{\mathrm{dir}}(f) + \mathbf{x}^{\mathrm{rev}}(f) + \mathbf{n}(f), \tag{1}$$

其中，$\mathbf{x}(f) \in \mathbb{C}^{Q\times 1}$ 表示频率 f处的麦克风信号向量。麦克风位于半径为 r 的球面上，位置为 $\Phi_q = (\theta_q, \phi_q)$、$q = 1$、.....、 Q。$\mathbf{x}^{\mathrm{dir}}(f)$和$\mathbf{x}^{\mathrm{rev}}(f)$分别对应目标语音的直达分量和混响分量，而$\mathbf{n}(f)$表示背景噪声与干扰声源的综合贡献。

同一声场也可以在球谐函数（SH）域（即Ambisonics域）[20] 中表示为

$$\mathbf{a}_{nm}(f) = \mathbf{a}_{nm}^{\mathrm{dir}}(f) + \mathbf{a}_{nm}^{\mathrm{rev}}(f) + \mathbf{a}_{nm}^{\mathrm{noise}}(f), \tag{2}$$

其中，$\mathbf{a}_{nm}(f) = [a_{00}(f)\ a_{1(-1)}(f)\ a_{10}(f)\text{.....},\ a_{NN}(f)]^T \in \mathbb{C}^{(N+1)^2\times 1}$ 是 N [24]阶Ambisonics系数向量。向量$\mathbf{a}_{nm}^{\mathrm{dir}}(f)$、$\mathbf{a}_{nm}^{\mathrm{rev}}(f)$和$\mathbf{a}_{nm}^{\mathrm{noise}}(f)$分别表示直达分量、混响分量和噪声分量。

最后，通过应用一组空间波束形成器，声学场景可以在波束形成器域中表示为：

$$\mathbf{b}(f) = \mathbf{b}^{\mathrm{dir}}(f) + \mathbf{b}^{\mathrm{rev}}(f) + \mathbf{b}^{\mathrm{noise}}(f), \tag{3}$$

其中，$\mathbf{b}(f) \in \mathbb{C}^{B\times 1}$ 包含 B 个波束形成器的输出，这些波束形成器指向由球坐标$\Psi_b = (\theta_b, \phi_b)$表示的方向，$b = 1$、.....、 B。向量 $\mathbf{b}^{\mathrm{dir}}(f)$、$\mathbf{b}^{\mathrm{rev}}(f)$ 和 $\mathbf{b}^{\mathrm{noise}}(f)$ 分别表示相应的直达路径、混响和噪声成分。

2.2. 信号表示之间的变换

在全文中，公式（2）中的 Ambisonics表示作为参考信号表示。随后，根据 [24]中的公式，由 Ambisonics系数推导出麦克风域和波束形成器域中的信号。为了将信号从球谐域转换到波束形成器域或麦克风域，我们使用：

$$\mathbf{b}(f) = \mathbf{Y}_{\mathrm{Beam}}(\mathbf{\Psi})\mathbf{a}_{nm}(f) \tag{4}$$

$$\mathbf{x}(f) = \mathbf{Y}_{\mathrm{Mic}}(\mathbf{\Phi})\mathbf{B}(fr)\mathbf{a}_{nm}(f) \tag{5}$$

其中，$\mathrm{Y}_{\mathrm{Beam}}(\Psi)$ 和 $\mathrm{Y}_{\mathrm{Mic}}(\Phi)$ 分别是 $B\times(N+1)^2$ 和 $Q\times(N+1)^2$ 阶球谐矩阵，如 [20,24] 所示。 $\mathbf{\Psi} = [\Psi_1, \Psi_2, ..., \Psi_B]$ 是 B 个观测方向的向量， $\mathbf{\Phi} = [\Phi_1, \Phi_2, ..., \Phi_Q]$ 是位于半径为 r 的球面上的 Q 个麦克风位置的向量。$\mathrm{B}(fr) = \mathrm{diag}(b_0(fr)\ b_1(fr)\ b_1(fr)\ b_1(fr)\ ..., b_N(fr))$ 是一个 $(N+1)^2 \times (N+1)^2$ 阶对角矩阵，其中 $b_n(fr)$ 表示依赖于频率和

到原点的距离的径向函数。要变换回球谐域，我们将使用：

$$\mathbf{a}_{nm}(f) = \mathbf{Y}^{\dagger}_{\text{Beam}}(\mathbf{\Psi})\mathbf{b}(f) \tag{6}$$

$$\mathbf{a}_{nm}(f) = (\mathbf{Y}_{\text{Mic}}(\mathbf{\Phi})\mathbf{B}(fr))^{\dagger}\mathbf{x}(f) \tag{7}$$

其中 $(.)^{\dagger}$ 为伪逆，假设 $Q, B \geq (N+1)^2$。由于我们使用伪逆，式 (7) 可能会引入误差；如果 B 和 Q 方向对球面的采样近乎均匀，则该误差可能较小 [24]。

2.3. 时频掩码

频域信号先变换到时域，再通过短时傅里叶变换（STFT）变换到时频域，从而得到表示 $\mathbf{x}(\tau,\nu)$、$\mathbf{a}_{nm}(\tau,\nu)$ 和 $\mathbf{b}(\tau,\nu)$。随后，对每种表示分别应用时频掩码，以估计所需的信号分量：

$$\hat{\mathbf{x}}(\tau,\nu) = \mathbf{M}_{\text{Mic}}(\tau,\nu)\mathbf{x}(\tau,\nu) \tag{8}$$

$$\hat{\mathbf{a}}_{nm}(\tau,\nu) = \mathbf{M}_{\text{Amb}}(\tau,\nu)\mathbf{a}_{nm}(\tau,\nu) \tag{9}$$

$$\hat{\mathbf{b}}(\tau,\nu) = \mathbf{M}_{\text{Beam}}(\tau,\nu)\mathbf{b}(\tau,\nu) \tag{10}$$

其中 $\mathbf{M}_{\text{Mic}}(\tau,\nu)$、$\mathbf{M}_{\text{Amb}}(\tau,\nu)$ 和 $\mathbf{M}_{\text{Beam}}(\tau,\nu)$ 为对角矩阵。这里，所需信号定义为混响目标信号。此过程结束时，信号将被变换回时域。本文将重点讨论两种掩码：

1）理想比率掩码（IRM）：我们将使用 [20,22] 中的理想比率掩码：

$$M(\text{ch},\tau,\nu) = \frac{SNR(\text{ch},\tau,\nu)}{SNR(\text{ch},\tau,\nu)+1} \tag{11}$$

$$\mathbf{M}(\tau,\nu) = \text{diag}(M(0,\tau,\nu), M(1,\tau,\nu), ...) \tag{12}$$

其中，$SNR(\text{ch}\ \tau,\nu)$ 表示每个通道中期望信号功率与非期望信号功率之比。

2）FT-JNF估计掩码：如 [25] 所示，FT-JNF（频率–时间联合非线性滤波器）是一种数据驱动模型，可直接根据含噪多通道输入估计掩码。

2.4. 双耳信号表示

增强后的信号通过在频域中应用基于头相关传递函数的全景声解码器，渲染为双耳信号。双耳左/右耳信号可表示为

$$p^{l/r}(f) = \tilde{\mathbf{h}}^{l/r}_{nm}(f)^T\mathbf{a}_{nm}(f), \tag{13}$$

其中，$\tilde{\mathbf{h}}^{l/r}_{nm}(f)$ 表示经过修改的球谐域头相关传递函数表示。它是根据 $\tilde{h}^{l/r}_{nm}(f) = (-1)^m h^{l/r}_{n\ (-m)}(f)$ [26]，由原始球谐域头相关传递函数系数 $\mathbf{h}^{l/r}_{nm}(f)$ 推导得到的。当增强后的信号在麦克风域或波束形成器域中获得时，首先使用公式 (6) 和 (7) 将其转换回球谐域，然后进行双耳解码。

3. 所提出的分析

本研究的核心目标是系统分析信号表示与通道特定掩蔽应用在空间音频语音增强中的相互作用。所提出的分析并非将语音增强与空间再现分开评估，而是考察抑制非期望信号与保留声学场景空间特性之间的权衡。

3.1. 研究问题

本研究目标通过以下研究问题加以探讨：

1. 信号表示如何影响语音增强与空间保真度之间的权衡?

2. 与计算效率较高的共享掩码相比，通道特定掩蔽在多大程度上能够提升增强效果？这种效果是否取决于信号表示?

3.2. 分析维度：信号表示与掩码策略

为回答这些问题，分析围绕两个维度展开。第一个维度是信号表示。本文考察三种表示：麦克风通道、一阶Ambisonics通道和波束形成器输出。核心假设是，掩蔽对这些表示的影响可能各不相同，因为它们的通道以不同方式编码空间信息。此外，分析还比较一阶和三阶Ambisonics表示，以考察Ambisonics阶数对这三种表示的影响。第二个维度是掩码策略。共享掩码是指将单个掩码应用于所有通道。通道特定掩蔽则是为每个通道分别估计独立掩码。通道特定掩码可以通过适应各通道的特性，提升语音增强效果，使其优于共享掩码；但也可能引入依赖通道的修改，从而扭曲空间线索。

3.3. 理想掩蔽与估计掩蔽

为将掩蔽策略的影响与掩码估计本身的影响区分开来，本文同时考察理想比率掩码（IRM）和轻量级 FT-JNF 神经估计器。IRM 为可实现性能提供理想参考，而 FT-JNF 则作为实际的掩码估计器，用于确定当必须从含噪多通道信号中估计掩码时，通过理想掩蔽观察到的关系是否仍然成立。

3.4. 评估框架

最后，分析综合考察目标信号和残余干扰的语音质量与可懂度、混响保留以及双耳线索。这样便可以跨不同信号表示和掩蔽策略，评估增强性能与空间保真度之间的权衡。随后将在仿真实验中受控声学条件下对该框架进行评估。

4. 性能客观指标

性能通过客观指标进行评估；如上文所述，这些指标用于评估语音可懂度、语音质量、双耳线索和混响保留情况。具体如下。

- SI-SDR：使用 SI-SDR [8]量化信号重建准确度。该指标分别针对左、右双耳通道独立计算，并取平均值以获得总体性能指标。混响目标语音信号用作参考信号。

- **ΔITD**、 **ΔILD**：通过 ITD 和 ILD [18,19]评估双耳空间信息的保留情况。ITD 是约 1.5 kHz 以下的主要定位线索，而 ILD 在更高频率下占主导地位 [18, 19]。这些指标使用目标信号的直达路径分量计算。增强过程引入的空间失真，表示为估计值与参考值之间 ITD 和 ILD 的绝对差异，分别记为 ΔITD 和 ΔILD。

- **ΔC_{50}**：使用清晰度指数 C_{50} [27]评估增强过程对混响的影响。该指标量化前 50 ms 内的早期到达能量与随后后期混响能量之间的比值。通过计算与参考混响目标信号之间的偏差，得到绝对 C_{50} 误差，并对双耳取平均，记为 ΔC_{50}。

- DBSTOI：使用确定性双耳 STOI（DBSTOI） [28], 测量语音可懂度，该指标考虑了双耳效应和非线性听觉处理。直达路径目标语音作为此次评估的参考。

- PESQ：使用 PESQ [9],评估语音质量，其中混响目标信号作为参考。分别计算每只耳朵的 PESQ 分数，并取平均值以获得最终分数。

5. 仿真实验

本节介绍用于评估所考虑语音增强方法的仿真框架。实验旨在相同声学条件下，对麦克风、Ambisonics 和波束形成器表示进行受控比较。仿真场景包括混响、定向干扰和背景噪声，从而能够同时评估语音增强性能以及空间信息的保留情况。首先介绍仿真设置、信号表示和训练过程，随后给出实验结果。

5.1. 设置

我们使用 shroom [29], 生成模拟数据集，该工具提供了源图像模型 [30]的实现。每个场景包含一个目标说话人、一个干扰说话人和一个平稳背景噪声源，并由虚拟接收器进行采集。该数据集总共包含 600 个训练集样本、200 个验证集样本和 100 个测试集样本，每个样本的时长为 6 秒。

对于每个场景，随机生成一个具有不同尺寸和墙体吸声的矩形房间，同时将接收器放置在距离墙壁至少 1 m 的随机位置。目标说话人位于接收器正前方的随机距离处，干扰说话人则位于相同距离处，但方向随机。平稳嘈杂噪声源被放置在随机选定的房间角落，以模拟高度混响的噪声场。根据指定的信噪比条件对干扰声和噪声源信号进行缩放，非期望信号则为干扰说话人与噪声之和。表 1 详细说明了声源位置及其他场景参数的选择方式。

每个场景均以 $N = 10$ 的高球谐阶数进行仿真，并由此导出所有信号表示，使三个域描述相同的底层声场。在主要实验中，随后将 Ambisonics 信号截断至一阶。麦克风信号由高阶声场通过半径为$r = 0.042$ m、包含 $Q = 4$ 个麦克风的刚性球形阵列渲染得到，而波束形成器信号则通过对一阶声场进行与信号无关的平面波分解，映射至 $B = 4$ 个观测方向 [24]而获得。麦克风位置和波束形成器观测方向均匀分布在球面上，其中第一个方向与目标信号对齐。在三阶Ambisonics通道比较中，相应的表示则在 $N = 3$ 阶下构建，使用 $Q = 24$ 个麦克风和 $B = 24$ 个波束形成器观测方向，后者采用近似均匀的斐波那契网格进行分布。通道类型表示如下：

- TFM：时频麦克风
- TFA：时频Ambisonics

• TFB：时频波束形成

训练集和验证数据：为了让网络接触到各种各样的空间配置，干扰源方位角和高度根据表 1 中指定的分布随机采样。

测试数据：对于评估集，相对于目标信号，干扰声被限制在两个固定的方位角方向上，从而使目标信号与干扰之间具有较大的空间分离度。这使得保留干扰的空间线索成为一项具有感知意义的任务。对于每个测试场景，还存储了额外的参考信号：目标信号的直达声（a$_{nm}^{\mathrm{dir}}$）、干扰声的直达声，以及与房间ARIR卷积后截断至直达声到达后50 ms的目标信号，用于计算C$_{50}$。

5.2. 方法

如第3节所述，计算两种类型的语音增强掩码。理想比率掩码直接根据已知的目标信号和非目标成分计算，如[20, 22]所示。对于数据驱动方法，我们按照 [25] 训练一个双向LSTM（FT-JNF），使其从含噪信号中估计复数时频掩码。网络输入为给定域中含噪多通道混合信号的单边短时傅里叶变换，将所有通道的实部和虚部拼接，形成一个 $[T \times F \times 2C]$ 张量，其中 $C = 4$ 是该域中的通道数。将估计掩蔽应用于选定参考通道的短时傅里叶变换，再将掩蔽后的短时傅里叶变换转换回时域。与 [25], 不同，此处训练网络时，优化的是重建参考信号与目标参考信号之间的信号失真比（SDR），而非尺度不变信号失真比。我们选择使用SDR，以保留通道间的尺度关系，这对后续空间处理十分重要。将混响目标信号用作目标参考信号，使网络保留目标信号的混响 [22]。此处将FT-JNF网络用作实用的掩码估计器，以研究不同信号表示和掩蔽策略的表现，而非将其作为架构比较的对象。

使用 FT-JNF 掩码估计器考虑了两种掩码应用策略。在第一种策略中，以通道0作为参考通道，训练单个 FT-JNF 网络。随后，将生成的掩码应用于多通道信号的所有通道，从而在各通道之间得到共享掩码。该方法仅使用单个网络，因此会在所有通道上施加相同的时频衰减模式。

在第二种策略中，每个参考通道都单独训练一个 FT-JNF 网络。每个网络接收完整的多通道输入，但优化目标是重建其对应的参考通道。生成的通道特定掩码分别独立应用于各自的通道，从而得到式（8）、（9）和

表1：仿真参数详情

变量	描述
语音	WSJ0 语音片段 [32]，$f_s = 16$ kHz
Room	矩形；宽度 $\in [10, 20]$，深度 $\in [6, 10]$，高度 $\in [3.2, 4]$ m；吸声 $\in [0.15, 0.5]$
接收器	随机位置，至少距墙壁 1.5 m，高度
目标信号	前方（0°），距离 $\in [1, 3]$ 米
干扰声	与目标信号的距离相同；高度$\sim \mathcal{N}$（1.6, 0.08）米；方位角偏移：训练集 $\sim \mathcal{U}[30°, 330°]$，测试集 $\in \{+60°, -100°\}$
SNR	目标源–干扰源 $\sim \mathcal{U}[-20, 10]$ 分贝；目标源–噪声 $= 20$ 分贝
噪声	随机房间角落中的多人语音噪声源

（10）所示的对角掩码矩阵。该策略使估计掩蔽能够适应每个通道的特性。针对全部三种信号表示，对这两种掩蔽策略均进行了评估。

在原始 FT-JNF 架构中，第一个 BLSTM层使用 256 个隐藏单元，第二个使用 128 个隐藏单元，因此共有约 1.2M 个可训练参数。在本研究中，我们将网络规模缩小约 10 倍，第一个和第二个 BLSTM层分别使用 80 个和 40 个隐藏单元，因此共有约 122K 个参数，以模拟计算资源有限且更贴近实际的边缘设备应用。

对于波束形成器和麦克风域，在进行双耳解码之前，使用公式 (6) 和 (7) 将增强后的信号变换回球谐域。最后，使用科隆HRTF汇编中的 NeumannKU-100 [31]，通过公式 (13) 计算双耳信号。

5.3. 结果

本节将介绍仿真实验的结果。我们首先考察基于一阶Ambisonics表示的通道特定掩蔽，随后讨论共享掩蔽、高阶表示以及逐通道 SI-SDR 分析。

5.3.1. 一阶通道特定掩蔽：理想掩蔽与估计掩蔽

我们首先考虑基本配置，即基于一阶Ambisonics表示并采用通道特定掩蔽。对比两种掩码：oracleIRM，它根据已知的期望成分和非期望成分，针对每个通道独立计算；以及 FT-JNF 掩码，它根据含噪多通道输入，针对每个通道分别估计。

表2：通道特定IRM以及应用于麦克风（TFM）、Ambisonics（TFA）和波束形成器（TFB）域的一阶Ambisonics（$N = 1$）FT-JNF估计掩码的性能。估计掩码分别采用每个通道一个独立掩码和跨通道共享掩码进行评估。

Mask	方法	尺度不变信号失真比（dB）↑	DBSTOI（-）↑	**PESQ**（-）↑	Δ双耳时间差$_{\text{tar}}$（μs）↓	Δ双耳声级差$_{\text{tar}}$（dB）↓	Δ双耳时间差$_{\text{int}}$（μ秒）↓	Δ双耳声级差$_{\text{int}}$（分贝）↓	**ΔC50**（分贝）↓
–	含噪	-9.37	0.478	1.191	–	–	–	–	–
IRM（通道特定）	TFM	3.02	0.797	2.367	0.0	0.87	**59.4**	5.32	1.75
	TFA	8.04	0.838	2.701	0.6	**0.07**	86.9	**0.84**	**1.29**
	TFB	**10.34**	**0.852**	**2.969**	0.0	0.80	117.5	1.46	1.65
FT-JNF（通道特定）	TFM	1.84	0.748	1.805	0.0	0.68	83.8	3.68	2.05
	TFA	6.08	0.786	1.966	0.0	**0.07**	**69.4**	**1.16**	**1.57**
	TFB	**7.21**	**0.791**	**2.039**	0.0	0.75	103.8	1.34	1.81
FT-JNF（共享）	TFM	1.47	0.719	1.717	0.0	0.35	82.5	3.56	2.05
	TFA	**5.20**	**0.773**	**1.859**	0.0	0.07	53.8	0.40	1.55
	TFB	3.17	0.757	1.668	0.0	**0.05**	**43.1**	**0.36**	**1.51**

表3：理想比率掩码应用于麦克风（TFM）、Ambisonics（TFA）和波束形成器（TFB）域时，在一阶（$N = 1$）和三阶（$N = 3$）球谐函数表示下的性能。为方便起见，$N = 1$ 行摘自表2。

阶数	方法	尺度不变信号失真比（dB）↑	DBSTOI（-）↑	**PESQ**（-）↑	Δ双耳时间差$_{\text{tar}}$（μ秒）↓	Δ双耳声级差$_{\text{tar}}$（分贝）↓	Δ双耳时间差$_{\text{int}}$（μ秒）↓	Δ双耳声级差$_{\text{int}}$（dB）↓	**ΔC50**（dB）↓
$N = 1$	含噪	-9.37	0.478	1.191	–	–	–	–	–
	TFM	3.02	0.797	2.367	0.0	0.87	59.4	5.32	1.75
	TFA	8.04	0.838	2.701	0.6	0.07	86.9	**0.84**	1.29
	TFB	**10.34**	**0.852**	**2.969**	0.0	0.80	117.5	1.46	1.65
$N = 3$	含噪	-7.60	0.513	1.182	–	–	–	–	–
	TFM	2.06	0.803	2.295	0.0	0.89	321.9	15.47	1.30
	TFA	9.27	0.844	2.770	0.0	0.08	**85.0**	**2.72**	1.31
	TFB	**13.18**	**0.884**	**3.360**	0.0	0.35	294.4	6.89	1.45

在后续章节考察掩码策略和空间分辨率的影响之前，该比较确立了oracle与估计掩蔽之间的关系。

如表2（上面两部分）所示，对于两种掩码，波束形成器域在语音质量和可懂度方面均表现最佳，取得了最高的 SI-SDR、DBSTOI 和 PESQ 得分。Ambisonics域的性能略低，而麦克风域的语音增强性能最低。因此，在这一基础配置下，无论掩码是估计得到的还是理想掩码，波束形成器域中的通道特定掩蔽都能提供最强的语音增强效果。这可以通过波束形成预处理的方向选择性来解释，因为它能够对目标信号进行初步增强。对于估计掩码，在三个域中的语音增强性能均低于理想比率掩码（IRM），SI-SDR、PESQ 和 DBSTOI 均较低。尽管先前的研究表明，估计掩码有能力超越 IRM [10], 但此处的网络是 FT-JNF 的缩小版本，参数数量约为原始网络的十分之一，这或许可以解释 IRM 的优越性。不过需要注意的是，此处的比较旨在考察实际掩码估计条件下不同表示的相对表现，而非追求最佳可实现性能。

对于空间线索，两种掩码表现相似。对于目标说话人，ΔITD 和 ΔILD 始终低于最小可觉差阈值；ITD 的阈值约为 100 μs，ILD 的阈值为 1 dB [18, 33]。对于干扰声，在两种掩码下，麦克风和 Ambisonics 表示的 ΔITD 值均低于最小可觉差阈值，而波束形成器表示的值则略高于该阈值。在三种表示中，Ambisonics 对干扰声的 ΔILD 误差最低：使用 IRM 时低于最小可觉差阈值，使用估计掩码时则略高于阈值；而麦克风表示的误差最高。

表4：含噪信号以及应用 FT-JNF 掩码后的逐通道尺度不变信号失真比（dB），在应用掩码的域中进行评估。“全部”表示四个通道的平均值。

		尺度不变信号失真比（dB）				
域	掩蔽	**ch0**	**ch1**	**ch2**	**ch3**	**All**
Ambisonics	含噪	−5.86	−6.03	−5.98	−6.03	−5.97
	共享	8.13	6.49	6.65	6.49	6.94
	通道特定	8.13	7.99	7.43	8.05	7.90
波束形成器	含噪	1.80	−9.48	−12.83	−8.43	−7.24
	共享	10.87	2.57	−0.68	3.58	4.08
	通道特定	10.87	4.90	4.45	5.50	6.43
麦克风	含噪	−4.36	−6.37	−7.19	−6.21	−6.03
	共享	7.61	6.21	5.57	6.32	6.43
	通道特定	7.61	6.76	5.93	6.66	6.74

最高误差。

以 ΔC_{50} 衡量的混响保留程度，在三种表示之间相对相近，并且两种掩码下的排序相同。Ambisonics 的保留效果最好，其次是波束形成器域和麦克风域。与理想比率掩码相比，掩码估计使这三个域中的 ΔC_{50} 均增加约 0.15–0.30 dB。对于所有表示和两种掩码，ΔC_{50} 略高于可觉差（约为 1 dB） [27], ，表明目标信号的混响特性发生了可感知但总体有限的变化。

总体而言，基本的一阶配置揭示了语音增强与空间保持之间的权衡。波束形成器域中的通道特定掩蔽能提供最强的语音增强，而 Ambisonics 则能更好地保留残余干扰和混响的空间特性。重要的是，从理想IRM 转为估计的 FT-JNF 掩码后，这种权衡以及三种表示的相对排序仍然保持不变；掩码估计会降低所有域的绝对性能，但不会改变它们之间的关系。这表明，在理想掩蔽下观察到的信号表示之间的主要相对差异，在实际掩码估计下基本得以保留。因此，第 5.3.3 节采用理想比率掩码来考察三阶 Ambisonics。使用理想比率掩码无需针对高阶表示所引入的更多通道训练独立的掩码估计网络，同时也为比较空间分辨率的影响提供了一致的依据。

5.3.2. 共享掩蔽的影响

接下来，我们研究在所有通道上使用同一个掩码，与通道特定掩蔽相比，会如何影响增强效果和空间保持。表2（下半部分，标记为“共享”）中的结果表明，共享掩码对麦克风域和 Ambisonics 域中的语音增强影响较小。相比之下，波束形成器域中的语音质量和可懂度下降更为明显。这表明波束形成器通道对使用共享掩码更为敏感，因为较强的方向选择性会导致通道间差异更大。

在三种表示中，共享掩蔽始终能够减少残余干扰的空间线索误差。对于 Ambisonics 和波束形成，这种减少尤为明显，其中 Δ双耳时间差和 Δ双耳声级差相较于通道特定掩蔽均大幅降低，表明残余干扰的空间特征得到了更好的保留。一种可能的解释是，对所有通道应用相同的时频衰减会引入更少的通道相关频谱修改，从而减少对通道间相对关系的失真。因此，通道之间的相对空间关系得到了更好的保留。

对于混响保留，也观察到类似但幅度较小的改善。对于 Ambisonics 和波束形成器表示，共享掩蔽降低了 ΔC_{50} 误差，而对于麦克风表示，该误差基本保持不变。

总体而言，共享掩蔽以降低语音增强为代价，实现了对空间信息更好的保留；这种权衡在波束形成器域中最为明显。相比之下，Ambisonics表示的语音增强性能与通道特定掩蔽接近，同时受益于更好的空间线索保留。

5.3.3. Ambisonics阶数的影响

接下来，我们使用理想IRM考察将空间表示从一阶提高到三阶的影响，结果见表3。提高阶数会显著增

大Ambisonics域与波束形成器域在语音质量和可懂度方面的差异。这归因于空间分辨率的提高，它产生了指向性更强的波束形成器；同时，更密集的空间采样使目标信号与干扰声源能够在掩蔽前实现更强的空间分离。相比之下，一阶表示提供了更宽的波束形成器响应，因此限制了这种空间选择性。

Ambisonics 阶数的影响也体现在对残余干扰空间特性的保留上。使用三阶表示时，Ambisonics 域与波束形成器域之间的差异更加明显，Ambisonics 对干扰双耳线索的保留效果更好。当表示降至一阶时，波束形成器的空间线索误差与 Ambisonics 的误差更加接近。麦克风表示在一阶时对干扰线索的保留效果也有所改善，不过其双耳声级差误差仍大于 Ambisonics 和波束形成器表示。

总体而言，提高 Ambisonics 阶数可提升波束形成器表示的空间分辨率，进而增强其在语音增强方面的优势，同时加大其在保留干扰声空间线索方面的劣势。在一阶时，波束形成器指向性减弱，使波束形成与 Ambisonics 之间的性能差距缩小。

5.3.4. 按通道性能分析

表4对共享和通道特定 FT-JNF 掩码架构所得的 SI-SDR 进行了逐通道分析。请注意，此处采用一阶 Ambisonics，因此仅展示 FT-JNF 掩码，不展示理想比率掩码，因为前者代表更实际的取值。结果显示，三种信号表示之间存在显著差异，凸显了此类分析的重要性。

对于波束形成器表示，掩码类型的选择会显著影响性能。将针对通道0训练的共享掩码应用于其他通道时，SI-SDR 会大幅下降。相比之下，对于每个通道分别训练独立网络的通道特定掩码，各通道间的差异没有那么大。这种现象可以用波束形成器通道的方向性来解释。由于每个波束形成器都指向不同方向，各通道的目标干扰比可能存在显著差异。

对于麦克风表示，可以观察到不同的表现。在四个通道中，使用共享掩码和通道特定掩码获得的尺度不变信号失真比（SI-SDR）值总体上较为相近，但仍可观察到细微差异。这些相对较小的差异表明，麦克风通道具有相似但并不完全相同的信号特性；含噪通道的SI-SDR值相近也反映了这一点。这可能归因于该紧凑球形麦克风阵列中麦克风之间相对较小的间距，其半径为0.042 m。

对于 Ambisonics表示，观察到不同的表现。尽管四个通道的指向性模式差异显著，但使用共享掩码获得的尺度不变信号失真比仅略低于使用通道特定掩码获得的结果。含噪通道的尺度不变信号失真比也反映了这一点：四个 Ambisonics通道的数值相近。对此的一种可能解释是，零阶Ambisonics通道表示声场的全向分量，因此其估计掩蔽较少受特定空间方向的影响；而其他所有通道虽然具有方向性响应，但由于球谐函数的复指数项，其响应沿方位角方向仅与相位有关，而与增益无关 [24]。

6. 结论

本研究探讨了空间音频语音增强中的时频掩蔽，并比较了麦克风、Ambisonics和波束形成器信号表示。研究考察了理想比率掩码（oracle）以及使用轻量级 FT-JNF网络估计的掩码，并从语音增强、可懂度、感知质量、混响保留和双耳空间线索保留等方面评估了性能。

结果表明，将Ambisonics表示从三阶降至一阶，会显著缩小Ambisonics域与波束形成器域之间的性能差距。空间分辨率的降低限制了波束形成器的指向性，从而削弱了其在空间上分离目标信号与干扰的能力。然而，借助通道特定的神经掩码，波束形成器域仍能提供最高的语音质量和可懂度；与此同时，Ambisonics域在保持残余干扰的空间特性和混响水平方面表现更佳。

对掩码生成策略的比较进一步揭示了增强性能与空间保真度之间的重要权衡。通道特定掩码通常能够提升语音增强效果，尤其是在波束形成器域中。相比之下，共享掩码能更好地保持通道间关系，并且在某些情况下还能更好地保留目标信号的混响特性。值得注意的是，在一阶Ambisonics域中，共享掩码策略仅使用一个网络，却能达到接近通道特定掩蔽的语音增强性能。因此，Ambisonics域中的共享掩蔽在计算复杂度、语音增强和空间保持之间实现了颇具吸引力的折中。

逐通道分析进一步表明，共享掩码的适用性在很大程度上取决于信号表示。波束形成器通道呈现出显著不同的方向性特征，因此更适合使用通道特定的掩码。

麦克风通道之间更为相似，因此两种掩蔽策略之间的差异仅较为有限。Ambisonics通道可能具有显著不同的空间特性，但针对全向零阶通道优化的掩码仍能有效泛化到各通道。这些结果强调，在设计空间语音增强系统时，应当联合考虑信号表示与掩码策略的选择。

总体而言，研究结果表明，不存在一种能够同时优化所有性能指标的单一表示。相反，适当的掩码策略取决于语音增强、空间保真度和计算复杂度之间所需的平衡。具体而言，采用共享掩码的一阶Ambisonics能够在保持较强语音增强性能的同时保留空间信息，是一种实用且高效的解决方案；而当最大化语音质量和可懂度是首要目标时，则更适合采用通道特定波束形成器掩码。

7. 参考文献

[1] A. Narayanan 和 D. Wang，《使用深度神经网络进行理想比率掩码估计以实现鲁棒语音识别》，载于 2013 年 IEEE 国际声学、语音与信号处理会议。IEEE，2013 年，第 7092–7096 页。[2] D. Wang 和 J. Chen，《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，第 26 卷，第 10 期，第 1702–1726 页，2018 年。[3] C. Zheng、H. Zhang、W. Liu、X. Luo、A. Li、X. Li 和 B. C. J. Moore，《频域单耳语音增强的六十年：从传统方法到深度学习方法》，Trends in Hearing，第 27 卷，第 23312165231209913 页，2023 年。[4] Y. Wang、A. Narayanan 和 D. Wang，《关于监督式语音分离的训练目标》，IEEE/ACM音频、语音与语言处理汇刊，第 22 卷，第 12 期，第 1849–1858 页，2014 年。[5] P.-S. Huang、M. Kim、M. Hasegawa-Johnson 和 P. Smaragdis，《用于单耳语音分离的深度学习》，载于 2014 IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2014 年，第 1562–1566 页。[6] P.-S. Huang、M. Kim、M. Hasegawa-Johnson 和 P. Smaragdis，《用于单耳源分离的掩码与深度循环神经网络联合优化》，IEEE/ACM音频、语音与语言处理汇刊，第 23 卷，第 12 期，第 2136–2147 页，2015 年。[7] C. H. Taal、R. C. Hendriks、R. Heusdens 和 J. Jensen，《时频加权噪声语音可懂度预测算法》，IEEE Transactions on Audio, Speech, and Language Processing，第 19 卷，第 7 期，第 2125–2136 页，2011 年。[8] J. Le Roux、S. Wisdom、H. Erdogan 和 J. R. Hershey，《SDR——半生不熟还是恰到好处？》，载于 ICASSP 2019-2019 IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，2019 年，第 626–630 页。

[9] A. W. Rix、J. G. Beerends、M. P. Hollier 和 A. P. Hekstra，《语音质量感知评估（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 IEEE 国际声学、语音与信号处理会议。论文集（编目号：01CH37221）。IEEE，2001 年，第 2 卷，第 749–752 页。

[10] Yi Luo 和 Nima Mesgarani，“Conv-tasnet：超越用于语音分离的理想时频幅度掩蔽”，《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。

[11] M. Kolbæk、Z.-H. Tan、S. H. Jensen 和 J. Jensen，《监督式单通道时域语音增强的损失函数》，IEEE/ACM音频、语音与语言处理汇刊，第 28 卷，第 825–838 页，2020 年。[12] S.-W. Fu、T.-W. Wang、Y. Tsao、X. Lu 和 H. Kawai，《利用全卷积神经网络，通过端到端波形语句增强优化直接评估指标》，IEEE/ACM音频、语音与语言处理汇刊，第 26 卷，第 9 期，第 1570–1584 页，2018 年。[13] A. Herzog、S. R. Chetupalli 和 E. A. P. Habets，《Ambisep：联合全景声到全景声语音分离与降噪》，IEEE/ACM音频、语音与语言处理汇刊，第 31 卷，第 3081–3094 页，2023 年。[14] J. Wang、J. Zhang、S. Chen 和 M. Sun，《一种轻量级实时双耳语音增强模型及其空间线索保留能力》，载于 ICASSP 2025-2025 IEEE国际声学、语音与信号处理会议 (ICASSP)。IEEE，2025 年，第 1–5 页。[15] X. Leng、J. Chen 和 J. Benesty，《双耳语音增强中降噪与语音/噪声空间信息保留之间的折中》，美国声学学会期刊，第 149 卷，3162，2021 5，第 3151no. 页 – 。[16] X. Sun、R. Xia、J. Li 和 Y. Yan，《一种基于深度学习并保留空间线索的双耳语音增强方法》，载于 ICASSP 2019-2019 IEEE国际声学、语音与信号处理会议 (ICASSP)。IEEE，2019 年，第 5766–5770 页。[17] M. Kumar、A. Bastine、L. Birnie、S. A. Holzinger、P. N. Sa marasinghe 和 T. D. Abhayapala，《基于相对传递矩阵的头戴式麦克风阵列录音双耳信号去噪》，载于第 11 届欧洲声学协会大会。欧洲声学协会，EAA，2025 年，第 4299–4306 页。[18] J. Blauert，《双耳听音技术》，Springer，2013 年。[19] C. F. Hauth、S. C. Berning、B. Kollmeier 和 T. Brand，《利用盲双耳处理阶段对语音双耳去掩蔽建模》，Trends in Hearing，第 24 卷，第 1–16 页，2020 年。[20] M. Lugasi 和 B. Rafaely，《利用掩蔽增强 Ambisonics信号双耳重放的语音》，IEEE/ACM音频、语音与语言处理汇刊，第 28 卷，第 1767–1777 页，2020 年。

[21] B. Rafaely、V. Tourbabin、E. Habets、Z. Ben-Hur、H. Lee、H. Gamper、L. Arbel、L. Birnie、T. Abhayapala 和 P. Samarasinghe，《录制声学场景双耳重放的空间音频信号处理——综述与挑战》，Acta Acustica，第 6 卷，第 47 页，2022 年。[22] Sheli Hendel、Boaz Rafaely 和 Dorothea Kolossa，《空间音频的基于掩蔽的语音增强：Ambisonics、波束形成和麦克风通道的比较》，2026 年。[23] A. Herzog、S. R. Chetupalli 和 E. A. P. Habets，《Ambisep：利用Transformer网络进行全景声到全景声混响语音分离》，载于 2022 年国际声学信号增强研讨会 (IWAENC)，2022 年，第 1–5 页。[24] B. Rafaely，《球形阵列处理基础》，第 8 卷，Springer，2015 年。[25]K. Tesch 和 T. Gerkmann，《深入了解用于改善多通道语音增强的深度非线性滤波器》，IEEE/ACM音频、语音与语言处理汇刊，第 31 卷，第 563–575 页，2022 年。[26] Y. Gayer、V. Tourbabin、Z. Ben-Hur、D. Alon 和 B. Rafaely，《适用于可穿戴阵列并改善双耳重放效果的全景声编码器》，载于 Forum Acusticum 论文集，2025 年，第 1–6 页。[27] C. L. Christensen、G. Koutsouris 和 J. H. Rindel，《ISO 3382 参数：我们能否模拟？能否测量？》，载于国际房间声学研讨会，2013 年，第 9–11 页。

[28] A. H. Andersen、J. M. De Haan、Z.-H. Tan 和 J. Jensen，《一种预测含噪及非线性增强双耳语音可懂度的方法》，载于《2016年IEEE国际声学、语音与信号处理会议（ICASSP）》。IEEE，2016年，第4995–4999页。[29] Y. Gayer，《Shroom：用于Ambisonics房间声学仿真和双耳渲染的Python框架》，arXiv预印本 arXiv:2603.27342，2026年。[30] J. B. Allen 和 D. A. Berkley，《用于高效仿真小房间声学的镜像法》，《美国声学学会期刊》，第65卷，第4期，第943–950页，1979年。[31] B. Bernsch¨utz，《Neumann KU 100球形远场HRIR/HRTF汇编》，载于第40届意大利声学年会（AIA）暨第39届德国声学年会（DAGA）会议论文集。德国声学学会（DEGA），柏林，2013年，第29卷。[32] D. B. Paul 和 J. Baker，《基于〈华尔街日报〉的CSR语料库设计》，载于《语音与自然语言：1992年2月23日至26日在纽约州哈里曼举行的研讨会论文集》，1992年。[33] N. J. Spencer、M. L. Hawley 和 H. S. Colburn，《正常听力与听力受损听者在若干参数上的双耳间差异敏感度之间的关系》，《美国声学学会期刊》，第140卷，第3期，第1783–1799页，2016年。