arXiv:2304.08707v1 [eess.AS] 18 Apr 2023

通过集成全频带与子频带建模实现极低算法延迟和复杂度的神经语音增强

Zhong-Qiu Wang[1], *Samuele Cornell*[2], *Shukjae Choi*[3],*Younglo Lee*[3],*Byeong-Yeol Kim*[3],*Shinji Watanabe*[1]

[1]语言技术研究所，卡内基梅隆大学，美国匹兹堡[2]意大利马尔凯理工大学 [3]现代汽车集团和42dot Inc.，韩国首尔

wang.zhongqiu41@gmail.com

摘要

我们提出了 FSB-LSTM，这是一种新颖的、基于长短期记忆网络（LSTM）的架构，将全频带与子频带（FSB）建模相结合，用于短时傅里叶变换（STFT）域中的单通道与多通道语音增强。该模型通过多个 FSB-LSTM 模块传递过完备输入表示，并维持一条信息高速通道。每个 FSB-LSTM 模块由一个全频带块和一个子频带块组成，前者用于对所有频率上的时频模式进行建模，后者用于对各子频带内部的模式进行建模；两个块均以降采样后的表示作为输入，并返回上采样后的判别性表示，再通过残差连接将其加到块输入上。该模型旨在实现较低的算法复杂度、较小的运行时缓冲区和极低的算法延迟，同时，即使帧移低至 2 毫秒，在噪声混响语音增强任务中也能取得出色的增强效果。

索引词— 低复杂度语音增强、逐帧在线语音增强、深度学习、助听器设计。

1. 引言

在过去十年中，深度学习极大地推动了语音增强的发展 [1]。然而，目前报告了优异性能的增强模型通常包含许多卷积、循环或自注意力块。这些模型往往计算量大、资源需求高，且处理延迟较长，不适用于在线实时增强；低延迟、低复杂度增强的研究则仍很不足。这些问题阻碍了现代神经语音增强模型在助听器等现实产品中的部署，因为这类产品的计算能力通常十分有限，也极大限制了基于深度神经网络（DNN）的增强技术的潜在应用范围。正如 [2, 3], 所指出的，理想的神经语音增强系统需要模型规模小，并在训练和推理时仅消耗少量内存、计算量和能耗，同时实现优异的增强性能和极低的处理延迟[1]。

许多近期的神经语音增强研究 [4–9]着重于通过缩小模型大小来实现更强的增强效果。尽管模型大小越小当然越理想，但在大多数现代边缘设备上，低于 20 兆字节（MB）的模型大小通常已足够，因为其存储空间和随机存取存储器（RAM）通常都大得多。我们认为，更迫切的问题是在实时执行单帧输入、单帧输出的语音增强时的运行时内存开销、算法复杂度和计算需求。解决这些问题需要对许多现有深度神经网络架构进行重大改动。例如，

- 注意力机制 [10–14], 尤其是其原始形式，会关注过去帧以捕获长程上下文，因此并不适合低复杂度的在线增强，因为它需要缓存许多过去帧，因而会占用大量内存；
- 尽管二维卷积（二维卷积）具有较少的参数，并且广泛应用于基于 U-Net 的语音增强中，包括幅度 [15–17], 复数时频（T-F） [18–29]和时域 [7], ，但其计算量通常很大。
- DPRNN [6] 和 TF-GridNet [9] 等最先进的双路径模型并不适合实时增强，因为它们的计算开销很大。例如，TF-GridNet 在每一层、每个频率上都运行一个长短期记忆网络，而且在每一帧中，它不会并行处理序列中的所有时间步。

基于这些认识，我们认为，LSTM [30] 等循环神经网络更适用于在线低复杂度语音增强，因为运行时只需缓冲一帧过去帧，内存开销和系统复杂度都可以较低。此外，小型全连接块（或一维卷积）的开销通常低于采用大型卷积核且输入、输出通道数较多的二维卷积块。

在此背景下，我们研究将堆叠长短期记忆网络用作深度神经网络骨干网络，以实现算法延迟和复杂度极低的帧在线语音增强。尽管已有研究探索了这一方向 [31–35], ，但通常是在单耳条件下以及远程会议场景中进行研究；在这些场景中，允许的处理延迟最高可达 40 ms [36] ，因此通常采用较大的帧移（例如 8、10 和 16 ms）。然而，在助听器设置中，算法延迟通常要求低于 5 ms [37]。这意味着，若重叠相加采用 50% 的帧重叠，帧移不能超过 2.5 ms；如此小的帧移会使待处理的帧序列变长，并且要实现实时增强，硬件延迟必须低于帧移。在这种情况下，如何设计一种低复杂度的深度神经网络架构，利用 LSTM 实现低算法延迟的单通道和多通道增强，是一个值得研究的重要问题。

在我们的实验中，我们观察到，即使帧移低至 1 毫秒，对全频带信息进行建模的多层单向长短期记忆网络仍能表现出令人印象深刻的性能，这表明长短期记忆网络非常适合用于助听器设计。我们进一步将全频带长短期记忆网络块与子带长短期记忆网络块相结合，从而融合互补的全频带和子频带信息，以实现更好的增强效果，由此提出了一种名为 FSB-LSTM 的新型深度神经网络架构，用于具有极低算法延迟和低复杂度的短时傅里叶变换域语音增强。在噪声混响

[1]处理延迟包括由算法设计（例如使用重叠相加）导致的算法延迟，以及每帧计算所需的硬件延迟 [3]。

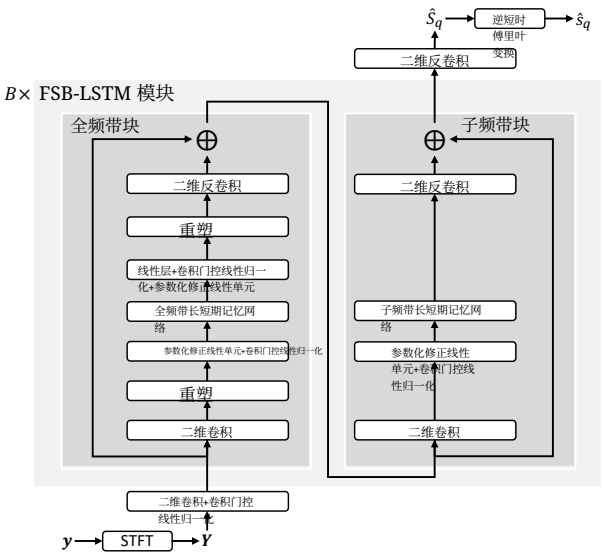

图 1：所提出系统的概览。

环境下对单通道与多通道语音增强的评估结果表明，与时域和复数时频域中的其他先进低延迟流式模型相比，FSB-LSTM 更为有效。消融实验也证实了所提出的集成全频带和子频带处理方法的有效性。

2. 所提算法

给定由 P 通道麦克风阵列在噪声混响环境下录制的单说话人 N 点混合信号，其时域物理模型可写为 $\mathbf{y}[n] = \mathbf{s}[n] + \mathbf{v}[n]$，其中 $\mathbf{y}[n]$、$\mathbf{s}[n]$ 和 $\mathbf{v}[n] \in \mathbb{R}^P$ 分别代表样本 n 处的混合信号、目标说话人的直达路径信号和非目标信号。本研究旨在低延迟、低复杂度的设置下，根据混合信号估计由参考麦克风 q（即 s_q）采集的目标直达路径信号。

在 STFT 域中，我们将混合信号表示为 $\mathbf{Y}(t,f) = \mathbf{S}(t,f) + \mathbf{V}(t,f) \in \mathbb{C}^P$，其中 $\mathbf{Y}$、$\mathbf{S}$ 和 $\mathbf{V}$ 分别是 $\mathbf{y}$、$\mathbf{s}$ 和 $\mathbf{v}$ 的短时傅里叶变换频谱，t 表示 T 帧，f 表示 F 个频率。本系统在 STFT 域中运行。遵循 [3], ，我们对 STFT 使用较大的输入窗长（iWS），并对逆短时傅里叶变换（iSTFT）中的重叠相加使用小得多的输出窗长（oWS）；iWS 和 oWS 均设为帧移（HS）的整数倍。这样，系统可以采用频率分辨率通常较高的 STFT，同时将算法延迟保持在较小的 oWS，而非通常较大的 iWS。有关这种 STFT-iSTFT机制的详细信息，请参见 [3] 。

图 1 展示了我们提出的系统。该系统经过训练，可执行基于多麦克风复数谱映射的语音增强 [23, 29, 38], ，其中将混合信号 $\mathbf{Y}$ 的实部和虚部（RI）分量堆叠为输入特征，以预测目标语音 S_q 的 RI 分量。给定形状为$2P \times T \times F$的输入张量，其中 $2P$ 是因为我们堆叠了所有 P 个麦克风的 RI 分量，我们首先使用卷积核尺寸为 1×3（沿时间和频率方向）的二维卷积层，为每个时频单元生成 D维嵌入表示，从而得到一个 $D \times T \times F$ 张量。随后，我们使用 B 个 FSB-LSTM 模块，每个模块都包含一个全频带块和一个子频带块，以利用频谱、空间和时序信息，逐步优化时频嵌入。接下来，使用卷积核尺寸为 1×3 的二维反卷积（Deconv2D）层预测目标 RI 分量。最后，应用逆短时傅里叶变换（iSTFT）进行信号重合成。损失函数基于重合成信号及其幅度定义，遵循 [3]中的 Wav+Mag 损失。其余内容

表 1：模型超参数摘要。

符号	说明
B	FSB-LSTM 模块数量
D	每个时频单元的嵌入维度
E	全频带块中二维卷积的输出通道数
I	全频带块中二维卷积和二维反卷积沿频率方向的卷积核尺寸
J	全频带块中二维卷积和二维反卷积沿频率方向的步幅大小
H	全频带长短期记忆网络中的隐藏单元数量
E'	子频带块中二维卷积的输出通道数
I'	子频带块中二维卷积和二维反卷积沿频率方向的卷积核尺寸
J'	子频带块中二维卷积和二维反卷积沿频率方向的步幅大小
H'	子带长短期记忆网络中的隐藏单元数量

本节将介绍 FSB-LSTM 中的全频带块和子频带块。为避免混淆，我们在表 1 中总结了用于描述 FSB-LSTM 的超参数。

2.1. 全频带块

给定形状为 $D \times T \times F$ 的输入张量，我们将每帧内的 D 维时频嵌入压缩为帧级嵌入，使用长短期记忆网络对帧嵌入进行精炼，并根据精炼后的帧嵌入重新计算 D 维时频嵌入。这样，长短期记忆网络就能同时对所有频率进行建模，从而捕获全频带信息。

具体而言，我们首先使用一个二维卷积层，其输入通道数为 D、输出通道数为 E、卷积核尺寸为 $1 \times I$、步幅为 $1 \times J$，在频率维度进行零填充后，沿维度一和维度三压缩 $D \times T \times F$ 张量至 $E \times T \times (\frac{Q-I}{J} + 1)$（$Q = \lceil \frac{F-I}{J} \rceil \times J + I$）。然后，将其重塑为一个 D 维张量，即展平第一维和第三维，以获得形状为 $T \times A$、具有 $A = E \times (\frac{Q-I}{J} + 1)$ 的张量；接着应用参数化修正线性单元，并执行因果全局层归一化（cGLN） [4], 。该归一化以因果方式基于这两个维度计算均值和方差，并使用两个 A 维向量分别沿第一维进行缩放和平移。接下来，我们使用具有 H 个隐藏单元的长短期记忆网络对 A 维帧嵌入进行建模，得到形状为 $T \times H$ 的张量。之后，应用一个线性层，将 H 维嵌入映射为 A 维，随后进行 cGLN 和参数化修正线性单元。最后，我们将 $T \times A$ 张量重塑回 $E \times T \times (\frac{Q-I}{J} + 1)$，并使用一个二维反卷积层（输入通道数为 E、输出通道数为 D、卷积核尺寸为 $1 \times I$、步幅为 $1 \times J$）计算出一个 $D \times T \times F$ 张量。移除填充的零后，通过残差连接将其与输入该全频带块的原始输入张量相加。

2.2. 子频带块

在 [9,39], 中，我们发现，利用子频带模块来利用子频带信息，对于去混响和利用空间信息非常有效。然而， [9, 39] 中提出的 TF-GridNet 会在每个频率上运行一个子频带模块，消耗大量计算量。为减少计算量，我们采用基于卷积的下采样来减少频率数量，并在 LSTM 中使用少得多的输入单元、隐藏单元和输出单元。

具体而言，给定形状为 $D \times T \times F$ 的输入张量，我们首先使用输入通道数为 D、输出通道数为 E'、卷积核尺寸为 $1 \times I'$、步幅大小为 $1 \times J'$ 的二维卷积层，对 $D \times T \times F$ 张量沿第一维和第三维进行下采样，得到 $E' \times T \times (\frac{Q'-I'}{J'} + 1)$，同时先将频率维度进行零填充至 $Q' = \lceil \frac{F-I'}{J'} \rceil \times J' + I'$。接下来，我们应用参数化修正线性单元，并执行卷积门控线性归一化；对于三维张量，该操作以因果方式基于全部三个维度计算用于归一化的均值和方差，并使用两个 E' 维向量分别沿第一维进行缩放和平移。之后，我们将张量视为 $\frac{Q'-I'}{J'} + 1$ 个序列，每个序列长度为 T，并使用具有 H' 个隐藏单元的长短期记忆网络进行细化

这些 E' 维时频嵌入，得到形状为 $H' \times T \times (\frac{Q'-I'}{J'}+1)$ 的张量。需要注意的是，所有子带共享同一个长短期记忆网络，以减少模型参数。最后，我们使用输入通道数为 H'、输出通道数为 D、卷积核尺寸为 $1 \times I'$、步幅大小为 $1 \times J'$ 的二维反卷积层，计算得到一个 $D \times T \times F$ 张量；移除填充的零值后，通过残差连接将其加到子频带块的输入张量上。

2.3. 网络设计讨论

在我们的网络中，我们保留了一条信息高速通道，通过残差连接在网络内部传递多通道输入信号的过完备时频表示（即 D 维的时频嵌入）。在细化时频嵌入时，我们首先执行下采样以提取所需的判别特征，然后使用长短期记忆网络层执行全频带或子频带时序建模，最后进行上采样并将其加回输入张量。这种策略尤其适用于多麦克风增强，因为它既能保留多个输入信号的细粒度信息（例如频谱-时序模式和空间模式），又能在不同模块中提取不同的感兴趣判别信息，从而实现更好的增强效果。

相比之下，早期一种广泛使用堆叠长短期记忆网络的方法，是将 $2P \times F$ 维输入特征直接输入多层长短期记忆网络（其方式类似于深度聚类 [40, 41] 和置换不变训练 [42] 研究中的做法）。然而，由于当 P 较大时，长短期记忆网络的隐藏维度通常远小于输入维度，因此输入特征的压缩会限制模型利用空间信息的能力。类似地，现代时域模型（如 Conv-TasNet）[4, 43] 往往会在编码器之后立即生成瓶颈表示。当输入维度较高时（例如在多麦克风场景中），这种瓶颈可能导致信息损失。一种解决方案是使用基于循环 U-Net 的多通道分离模型 [23,29,38],，其中 U-Net 编码器的较低层（以及解码器中的对应层）可以保留输入特征的过完备表示，以维持细粒度模式，同时逐步将输入特征下采样至适合循环网络的维度。在我们的实验中，我们将展示 FSB-LSTM 的性能优于一种强大的基于循环 U-Net 的模型 [44] 和多通道 Conv-TasNet [43, 45]。

我们模型中的所有卷积在时间维度上的卷积核尺寸均为 1。这样，由于采用因果卷积，我们无需缓冲过去帧，而只需使用长短期记忆网络来建模时序信息。我们模型中的二维卷积层和二维反卷积层计算开销不大，因为它们采用较大的步幅大小、较小的卷积核尺寸，以及较少的输入和输出通道。对于反卷积，我们使用自定义实现[2] 以减少乘累加（MAC）运算次数。

我们强调，所提网络只需缓冲过去帧中长短期记忆网络的隐藏状态和细胞状态。维护该缓冲区的运行时内存开销和复杂度都很低。

[2]反卷积（又称 转置卷积） [30] 通常通过先根据步幅大小在输入张量中插入零值，再执行常规卷积来实现。当步幅大于 1 时，这会增加乘加运算量，因为插入零值后，新的输入张量包含更多需要进行卷积的元素。在我们的研究中，我们将反卷积实现为一个线性层，随后沿频率维度进行重叠相加（请勿将此重叠相加与逆短时傅里叶变换中的重叠相加混淆）。这样可以节省对插入零值所浪费的计算量。另一方面，与线性层相比，重叠相加通常只需进行可忽略不计的乘加运算。我们的实现所需的乘加运算量约为典型实现的 $1/J$，其中 J 为步幅大小。

3. 实验设置

我们在模拟的噪声混响语音增强任务上验证了我们的算法。本节介绍数据集、系统配置、基线系统和评估指标。

3.1. 数据集

我们使用近期研究中用于评估所提算法的模拟数据 [3, 44], 。基于 WSJCAM0 中干净语音的划分，该数据集分别模拟了 39,245（∼77.7 小时）、2,965（∼5.6 小时）和 3,260（∼8.5 小时）的噪声混响混合信号，用于训练、验证和测试。从 FSD50k 的开发集 [46] 中采样音频片段，以模拟训练和验证所用的噪声；测试所用的噪声则取自评估集。每个模拟混合信号最多包含七段噪声音频，其中一段长于十秒，作为背景噪声，其余则作为前景噪声。模拟麦克风阵列包含六个麦克风，均匀排列在直径为 20 厘米的圆周上。每个声源相对于阵列中心的方向从范围 $[0, 2\pi)$ 中采样，距离从 $[0.75, 2.5]$ 米中采样，混响时间则从 $[0.2, 1.0]$ 秒中采样。我们将每个声源视为点声源，将每个声源与模拟房间脉冲响应进行卷积，并将卷积后的声源相加，以生成混合信号。目标直达路径语音与混响噪声之间的信噪比从 $[-8, 3]$ dB 中抽取。采样率为 16 kHz。对于双通道处理，我们使用第一个和第四个麦克风的信号；对于单声道处理，则使用第一个麦克风的信号。第一个麦克风捕获的目标直达路径信号用作模型训练的标签，以及指标计算的参考信号。

3.2. 系统配置

我们的目标是构建一个算法延迟为 4 ms 的增强系统，这略短于近期 Clarity 挑战赛 [37] 针对助听器设计提出的 5 ms 要求。对于 STFT 和 iSTFT，默认情况下，iWS 设为 16 ms，帧移设为 2 ms，oWS 设为 4 ms，因此算法延迟为 4 ms [3]。分析窗采用矩形窗。采样率为 16 kHz 时，使用 256 点离散傅里叶变换，在每一帧提取 129 维复频谱。根据验证集，我们设定 $B = 3$, $D = 32$, $E = 8$, $I = 8$, $J = 4$,$H = 256$ $E' = 64$, $I' = 5$, $J' = 5$ 和 $H' = 64$ （符号的定义见表 1）。在此配置下，共有 26 个子带，子带模块的乘加运算量约为全带模块的两倍。

3.3. 基线系统

我们将 Conv-TasNet [4], 及其多通道扩展 MC-Conv-TasNet [43, 45], LSTM-ResUNet [3], 和仅采用全频带的 LSTM 模型作为主要基线。所有模型均使用与 FSB-LSTM 相同的损失函数进行训练。

Conv-TasNet [4] 是语音分离领域出色的时域模型。它在非常短的信号窗口上使用学习基，实现了算法延迟极低的分离。使用 Conv-TasNet 论文 [4], 表 I 中列出的符号，我们将 Conv-TasNet 和 MC-Conv-TasNet 的超参数设为$N = 512$, $B = 158$, $S_c = 158$, $H = 512$, $P = 3$, $X = 8$和$R = 3$（请勿将这些符号与本文中定义的符号混淆）。 B 和 S_c 设得略大于 [4], 建议的默认值 128，因为 MC-Conv-TasNet 会将额外的空间嵌入表示与频谱嵌入表示拼接，作为 Conv-TasNet 分离器的输入。按照 [43, 45], 的做法，双通道增强的空间嵌入维度设为 60，六通道增强则设为 360。

表 2：FB-LSTM 在不同帧移下的结果（6 通道）。

系统	输入窗口长度（毫秒）	输出窗口长度（毫秒）	帧移（毫秒）	参数数量（百万）	每秒十亿次乘加运算	尺度不变信号失真比 (dB)	感知语音质量评估	扩展短时客观可懂度
未处理	-	-	-	-	-	−6.2	1.44	0.411
FB-LSTM（6 层）16	16	8	3.59	0.58	3.9	2.06	0.721
FB-LSTM（6 层）16	8	4	3.59	1.16	5.8	2.28	0.776
FB-LSTM（6 层）16	4	2	3.59	2.33	6.8	2.37	0.795
FB-LSTM（6 层）16	2	1	3.59	4.65	**8.0**	**2.54**	**0.821**

表3：语音增强结果（6通道）。

系统	参数量（百万）	每秒十亿次乘加运算	尺度不变信号失真比（dB）	感知语音质量评估	扩展短时客观可懂度
未处理	-	-	−6.2	1.44	0.411
FSB-LSTM	1.96	3.37	**7.8**	**2.61**	**0.830**
FB-LSTM（6 层）	3.59	2.33	6.8	2.37	0.795
FB-LSTM（9 层）	5.38	3.43	7.6	2.51	0.816
MC-Conv-TasNet [43,45]	6.37	3.76	5.2	2.24	0.764
LSTM-ResUNet [3]	2.33	3.54	6.2	2.23	0.767

LSTM-ResUNet [3] 是一种具有代表性的复数时频域模型，由夹在 U-Net 中的多层长短期记忆网络组成，并在多个频率尺度上插入残差网络块。它在重叠相加中采用比输入窗口长度更短的输出窗口长度，以实现低算法延迟的增强 [3]。我们强调，它的网络架构与近期用于语音增强的复数时频域模型 [17, 18, 20, 29] 有许多相似之处。因此，除了 Conv-TasNet 之外，我们也将其视为一个重要基线。

为了展示加入子带长短期记忆网络的效果，我们将图 1 中的子带模块替换为全带模块。这样，系统本质上堆叠了多个全频带长短期记忆网络。我们将该系统称为 **FB-LSTM**，其中“FB”表示全频带。我们测试了包含 6 个和 9 个全频带长短期记忆网络块的 FB-LSTM，因为 FSB-LSTM 使用 $B = 3$ 个全带和子带模块（共计 $2 \times 3 = 6$ 个长短期记忆网络），且每个子带模块的乘加运算量约为全带模块的两倍。

3.4. 评估指标

评估指标包括尺度不变信号失真比（SI-SDR）、语音质量感知评估（PESQ）和扩展短时客观可懂度（eSTOI）。对于 PESQ，我们使用*python-pesq*（v0.0.2）工具包报告窄带 MOS-LQO 分数。模型参数量以百万（M）为单位报告。使用*ptflops* 工具包，我们通过统计每秒十亿次乘加运算（GMAC/s）来报告计算量。

4. 评估结果

4.1. LSTM 处理小帧移的有效性

尽管有人批评 LSTM 不擅长对小帧移产生的长序列进行建模 [4], ，但在我们专注于帧在线增强的实验中，即使帧移低至 1 ms，我们仍发现 FB-LSTM 的表现出人意料地出色。结果见表 2。输入窗口长度始终为 16 ms。当算法延迟（等于输出窗口长度）变得更小时，我们同步减小帧移和输出窗口长度，使重叠相加中的帧重叠率始终保持为 50%。每当帧移减半时，由于需要处理的帧数加倍，计算量大约也会加倍。从结果来看，我们观察到，更小的帧移（以及输出窗口长度）能够带来更好的性能，尽管由此产生的帧序列会长得多，并且可利用的未来上下文信息（最多延伸至未来输出窗口长度）会变少。

尽管在 Conv-TasNet [4]和 DPRNN [6], 等时域说话人分离研究中，较小的跳数已被证明有效，但它们通常与更先进的架构搭配使用，而非用于对帧序列进行从左到右建模的简单单向长短期记忆网络。

表 4：语音增强结果（2 通道）。

系统	参数数量（M）	每秒十亿次乘加运算	SI-SDR（dB）	PESQ	扩展短时客观可懂度
未处理	-	-	−6.2	1.44	0.411
FSB-LSTM	1.96	3.31	**4.9**	**2.20**	**0.753**
FB-LSTM（6 层）	3.59	2.27	4.2	2.07	0.729
FB-LSTM（9 层）	5.38	3.37	4.5	2.14	0.744
MC-Conv-TasNet [43,45]	6.19	3.68	3.6	2.00	0.711
LSTM-ResUNet [3]	2.33	3.48	4.3	2.06	0.726

表 5：语音增强结果（1 通道）。

系统	参数数量 (M)	每秒十亿次乘加运算	尺度不变信号失真比 (dB)	感知语音质量评估	扩展短时客观可懂度
未处理	-	-	−6.2	1.44	0.411
FSB-LSTM	1.96	3.30	**3.1**	**1.92**	**0.688**
FB-LSTM（6 层）	3.59	2.25	2.6	1.84	0.671
FB-LSTM（9 层）	5.38	3.35	2.5	1.83	0.667
Conv-TasNet [4]	6.18	3.67	2.2	1.78	0.657
LSTM-ResUNet [3]	2.32	3.47	2.8	1.90	0.682

我们的研究发现，这种简单的因果 LSTM 即使帧移低至 1 毫秒，也能取得不错的表现。这一发现意义重大，因为它表明，运行时复杂度极低的简单 LSTM，能够在要求处理延迟极低的助听器应用场景中实现令人期待的增强效果。

4.2. FSB-LSTM 的结果

表3、表4和表5分别展示了 FSB-LSTM 在六通道、双通道和单通道语音增强任务上的结果。可以看出，FSB-LSTM 的模型大小更小，且 MAC/s 运算量更少，却比 Conv-TasNet、MC-Conv-TasNet 和 LSTM-ResUNet 实现了更好的增强效果。值得注意的是，Conv-TasNet 和 LSTM-ResUNet 都包含沿时间方向扩张的卷积，因此在运行时需要缓冲许多过去帧。

FSB-LSTM 的结果优于 6 层和 9 层 FB-LSTM。这表明使用子频带块具有优势。

4.3. 运行时复杂度

我们在在线流式设置下，基于单精度浮点运算，计算了各模型的运行时缓冲区大小。FSB-LSTM 只需缓冲过去帧中 LSTM 的隐藏状态和细胞状态。其缓冲区大小为 46.2 千字节（KB）；而 Conv-TasNet 和 LSTM-ResUNet 的缓冲区大小分别为 3133.4 KB 和 1815.1 KB。如此小的缓冲区使缓冲张量能够驻留在更高层级的缓存中；即使在现代处理器上，这些缓存的空间也非常有限（例如，一级缓存为数十 KB，二级缓存为数 MB）。较小的缓冲区和较低的算法复杂度，也使硬件延迟更容易低于帧移，从而在资源受限的助听器场景中实现实时增强。

5. 结论

我们提出了一种新颖的 FSB-LSTM 架构，将全频带与子频带建模相结合，用于实现低复杂度、低算法延迟的语音增强。实验结果表明，与此前提出的先进可流式、低延迟模型相比，FSB-LSTM 只需更少的缓冲内存，并且乘加运算的计算负担更低，同时性能更优。未来的研究将进一步降低算法延迟，并探索深度神经网络量化与蒸馏技术，以进一步降低复杂度，同时保持出色的增强性能。

6. 参考文献

[1] D. Wang 和 J. Chen，《基于深度学习的有监督语音分离：综述》，IEEE/ACM 音频、语音与语言处理汇刊，第 26 卷，第 1702–1726 页，2018 年。[2] E. Tzinis、Z. Wang 和 P. Smaragdis，《Sudo RM -RF：用于通用音频源分离的高效网络》，载于 MLSP 会议论文集，2020 年。[3] Z.-Q. Wang、G. Wichern、S. Watanabe 和 J. Le Roux，《具有极低算法延迟的短时傅里叶变换域神经语音增强》，arXiv 预印本 arXiv:2204.09911，2022 年。[4] Y. Luo 和 N. 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 1256–1266 页，2019 年。[5] Y. Liu 和 D. Wang，《用于单声道、说话人无关的说话人分离的因果深度 CASA》，IEEE/ACM 音频、语音与语言处理汇刊，第 1270–1279 页，2020 年。[6] Y. Luo、Z. Chen 和 T. Yoshioka，《双路径循环神经网络：用于时域单通道语音分离的高效长序列建模》，载于 ICASSP 会议论文集，2020 年，第 46–50 页。[7] A. Pandey 和 D. Wang，《采用自注意力的密集卷积神经网络用于时域语音增强》，IEEE/ACM 音频、语音与语言处理汇刊，第 29 卷，第 1270–1279 页，2021 年。[8] L. Yang、W. Liu 和 W. Wang，《TFPSNet：用于语音分离的时频域路径扫描网络》，载于 ICASSP 会议论文集，2022 年，第 6842–6846 页。[9] Z.-Q. Wang、S. Cornell、S. Choi、Y. Lee 等，《TF-GridNet：让时频域模型再次出色地用于单声道说话人分离》，载于 ICASSP 会议论文集，2023 年。[10] A. Vaswani、N. Shazeer、N. Parmar 等，《注意力就是一切》，载于 NIPS 会议论文集，2017 年。[11] C. Subakan、M. Ravanelli、S. Cornell 等，《语音分离中的注意力就是一切》，载于 ICASSP 会议论文集，2021 年，第 21–25 页。[12] ——，《资源高效分离变换器》，arXiv 预印本 arXiv:2206.09507，2022 年。[13] G. Zhang、C. Wang、L. Yu 和 J. Wei，《采用轴向注意力的多尺度时频卷积网络用于多通道语音增强》，载于 ICASSP 会议论文集，2022 年，第 9206–9210 页。[14] A. Pandey 和 D. Wang，《用于提升跨语料库泛化能力的语音增强自注意力循环神经网络》，IEEE/ACM 音频、语音与语言处理汇刊，第 30 卷，第 1374–1385 页，2022 年。[15] K. Tan 和 D. Wang，《用于实时语音增强的卷积循环神经网络》，载于 Interspeech 会议论文集，2018 年，第 3229–3233 页。[16] O. Ernst、S. E. Chazan 等，《使用全卷积网络进行语音去混响》，载于欧洲信号处理会议论文集，2018 年，第 390–394 页。[17] K. Tan 和 D. Wang，《利用门控卷积循环网络学习复数谱映射以进行单声道语音增强》，IEEE/ACM 音频、语音与语言处理汇刊，第 28 卷，第 380–390 页，2020 年。[18] Y. Liu 和 D. Wang，《分而治之：面向说话人无关的单声道说话人分离的深度 CASA 方法》，IEEE/ACM 音频、语音与语言处理汇刊，第 27 卷，第 12 期，第 2092–2102 页，2019 年。[19] U. Isik、R. Giri 等，《PoCoNet：借助频率位置嵌入、半监督对话数据和偏置损失实现更优的语音增强》，载于 Interspeech 会议论文集，2020 年，第 2487–2491 页。[20] Y. Hu、Y. Liu、S. Lv、M. Xing 等，《DCCRN：用于相位感知语音增强的深度复数卷积循环网络》，载于 Interspeech 会议论文集，2020 年，第 2472–2476 页。[21] Z.-Q. Wang 和 D. Wang，《基于深度学习的目标消除用于语音去混响》，IEEE/ACM 音频、语音与语言处理汇刊，第 28 卷，第 941–950 页，2020 年。[22] 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。[23] ——，《用于逐话语和连续语音分离的多麦克风复数谱映射》，IEEE/ACM 音频、语音与语言处理汇刊，第 29 卷，第 2001–2014 页，2021 年。[24] H. Taherian、Z.-Q. Wang、J. Chang 和 D. Wang，《基于单通道和多通道语音增强的稳健说话人识别》，IEEE/ACM 音频、语音与语言处理汇刊，第 28 卷，1293–1302 页，2020 年。[25] Z.-Q. Wang、G. Wichern 和 J. Le Roux，《用于单耳语音去混响和噪声混响环境下说话人分离的卷积预测》，IEEE/ACM 音频、语音与语言处理汇刊，第 29 卷，3476–3490 页，2021 年。[26] S. Zhao、B. Ma、K. N. Watcharasupat 和 W.-S. Gan，《FRCRN：利用频率递归增强单耳语音增强的特征表示》，载于 ICASSP 会议论文集，2022 年，9281–9285 页。[27] S. E. Eskimez、T. Yoshioka、H. Wang、X. Wang 等，《个性化语音增强：新模型与综合评估》，载于 ICASSP 会议论文集，2022 年。[28] H. Taherian、S. E. Eskimez、T. Yoshioka、H. Wang 等，《一个模型增强所有语音：与阵列几何无关的多通道个性化语音增强》，载于 ICASSP 会议论文集，2022 年，271–275 页。[29] K. Tan 等，《神经频谱空间滤波》，IEEE/ACM 音频、语音与语言处理汇刊，第 30 卷，605–621 页，2022 年。[30] A. Courville、I. Goodfellow 和 Y. Bengio，《深度学习》。MIT出版社，2016 年。 [在线]。网址：http://www.deeplearningbook.org [31] H. Zhao、S. Zarar 等，《用于语音增强的卷积循环神经网络》，载于 ICASSP 会议论文集，2018 年，2401–2405 页。[32] J. M. Valin 、 U. Isik、N. Phansalkar、R. Giri 等，《一种以感知为动机的全频带语音低复杂度实时增强方法》，载于 Interspeech 会议论文集，2020 年，2482–2486 页。[33] Y. Xia、S. Braun、C. K. Reddy、H. Dubey 等，《用于神经网络实时语音增强的加权语音失真损失》，载于 ICASSP 会议论文集，2020 年，871–875 页。[34] S. Braun、H. Gamper 等，《迈向高效的实时深度噪声抑制模型》，载于 ICASSP 会议论文集，2021 年，656–660 页。[35] M. Thakker、S. E. Eskimez、T. Yoshioka 和 H. Wang，《快速实时个性化语音增强：端到端增强网络（E3Net）与知识蒸馏》，载于 Interspeech 会议论文集，2022 年，991–995 页。[36] C. K. Reddy、V. Gopal 等，《INTERSPEECH 2020 深度噪声抑制挑战赛：数据集、主观语音质量与测试框架》，载于 Interspeech 会议论文集，2020 年，2492–2496 页。[37]《Clarity 挑战赛：助听设备的机器学习挑战》。 [在线]。网址：http://claritychallenge.org/ [38] Z.-Q. Wang 和 D. Wang，《用于语音去混响的多麦克风复频谱映射》，载于 ICASSP 会议论文集，2020 年，486–490 页。[39] Z.-Q. Wang、S. Cornell、S. Choi 等，《TF-GridNet：融合全频带与子频带建模的语音分离》，arXiv 预印本 arXiv:2211.12433，2022 年。[40] J. R. Hershey、Z. Chen、J. Le Roux 和 S. Watanabe，《深度聚类：用于分割与分离的判别式嵌入》，载于 ICASSP 会议论文集，2016 年，31–35 页。[41] Z.-Q. Wang、J. Le Roux 和 J. R. Hershey，《多通道深度聚类：用于说话人无关语音分离的判别式频谱与空间嵌入》，载于 IEEE 国际声学、语音与信号处理会议，2018 年，1–5 页。[42] M. Kolbæk、D. Yu、Z.-H. Tan 和 J. Jensen，《采用深度循环神经网络话语级置换不变训练的多说话人语音分离》，IEEE/ACM 音频、语音与语言处理汇刊，第 25 卷，第 10 期，1901–1913 页，2017 年。[43] J. Zhang、C. Zorila、R. Doddipatla 和 J. Barker，《混响环境中的端到端多通道时域语音分离》，载于 ICASSP 会议论文集，2020 年，6389–6393 页。[44] Z.-Q. Wang、G. Wichern 和 J. Le Roux，《利用低失真目标估计改善语音增强》，arXiv 预印本 arXiv:2110.00570，2021 年。[45] Z. Tu、J. Zhang 等，《用于噪声环境助听器处理的两阶段端到端系统》，载于 Clarity 会议论文集，2021 年，3–5 页。[46] E. Fonseca、X. Favory、J. Pons、F. Font 等，《FSD50K：人工标注声音事件的开放数据集》，IEEE/ACM 音频、语音与语言处理汇刊，2021 年。