arXiv:2302.07928v1 [eess.AS] 15 Feb 2023

带精细化的多通道目标说话人提取：WAVLAB 提交第二届 Clarity 增强挑战赛的方案

Samuele Cornell[1,2], *Zhong-Qiu Wang*[2],*Yoshiki Masuyama*[3,2],*Shinji Watanabe*[2]*Manuel Pariente*[4],*Nobutaka Ono*[3]

[1]马尔凯理工大学`意大利 [2]卡内基梅隆大学，美国[3]东京都立大学，日本 [4] Pulse Audition，法国{cornellsamuele, wang.zhongqiu41}@gmail.com

摘要

本文介绍了我们参加第二届 Clarity 增强挑战赛（CEC2）的方案。该挑战赛旨在面向助听器（HA）设备，在包含音乐和竞争说话人等多个干扰源的噪声混响环境中进行目标语音增强。我们的方法基于近期工作中提出的强大迭代神经网络/波束形成增强（iNeube）框架，并在此基础上将其扩展用于目标说话人提取。因此，我们将所提出的方法命名为 iNeuBe-X，其中“X”代表提取。为应对 CEC2 场景中的挑战，我们引入了四项主要创新：（1）将最初为单声道说话人分离设计的先进 TF-GridNet 模型 [1], 扩展用于多通道因果语音增强；用这一新架构替换 iNeube 中使用的 TCNDenseNet [2] 后，性能获得了显著提升；（2）采用近期提出的双窗口大小方法与未来帧预测 [3]，确保 iNueBe-X 满足 CEC2 要求的 5 ms 算法延迟限制；（3）为 TF-GridNet 引入新型说话人条件分支，以实现目标说话人提取；（4）提出一个微调步骤，根据经过听者听力图补偿的目标说话人信号计算附加损失。在不使用外部数据的情况下，在官方开发集上，我们的最佳模型取得了 0.942 的助听器语音感知指数（HASPI）得分，以及 18.8 dB 的尺度不变信号失真比提升（SI-SDRi）。鉴于 CEC2 数据极具挑战性（例如，在开发集上，混合信号的 SI-SDR 为 -12.3dB），这些结果颇具前景。我们提交系统的演示可在 WAVLab CEC2 演示中查看。

1. 所提方法

1.1. iNeuBe-X 系统概述

我们提出的 iNeuBe-X 系统如图 1 所示。该系统的设计灵感来自 iNeuBe 框架 [2], , 其中包含两个多麦克风
输入单麦克风输出（MISO）深度神经网络 [4], , 其中第一个网络生成初始目标语音估计，用于波束形成；第二个网络执行后滤波，以获得更好的增强效果。一个关键区别是，我们在此使用从第三个深度神经网络$_{\text{spk}}$ 模块中提取的说话人注册嵌入来调节每个 MISO 深度神经网络。该模块与其余部分联合训练，并以目标说话人的注册语句为输入。给定混合信号的复数多通道短时傅里叶变换（STFT）频谱图 $\boldsymbol{Y}$，第一个深度神经网络（图 1 中记作 DNN$_1$ ）会估计在参考 HA 阵列麦克风处采集的无混响目标语音的复数 STFT 频谱图 $\hat{S}_1^{(n)}$（第 $n=1$ 次迭代），该参考麦克风在 CEC2 中设为“CH1 左”。随后可通过第二步对该估计进行细化，该步骤包括一个深度神经网络支持的波束形成器，以及第二个 MISO 深度神经网络（记作 DNN$_2$），其输入还包括以下输出：

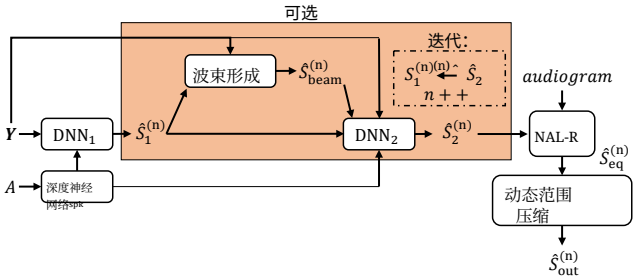

图 1：所提出的 iNeuBe-X 框架概览。我们在 DNN$_1$ 和 DNN$_2$ 之间使用因果多通道维纳滤波波束形成器。

深度神经网络$_1$ 和波束形成器。按照 [2], 可以迭代执行这一细化过程。我们的系统经过训练，可在 32 kHz 采样率下运行，并使用窗长为 16 ms、步长为 4 ms 的短时傅里叶变换以及平方根汉宁窗。这样系统的理论延迟为 16 ms [3]。为克服这一问题，我们训练深度神经网络$_1$和深度神经网络$_2$，使其预测当前帧及目标无混响信号未来 3 帧的内容，从而每隔 4 ms 就能合成输出一帧。实际操作中，这是通过在输入混合信号短时傅里叶变换的左侧补零来实现的。波束形成以逐帧在线方式执行，以满足 CEC2 对算法延迟不超过 5 ms 的要求。我们采用了 [2, 3] 中的因果多通道维纳滤波器。由于听者头部在 CEC2 中可能转动，我们没有简单地累积统计量，而是采用遗忘因子为 0.5 的递归平均策略 [5] 。

1.2. 说话人注册嵌入提取

对于 DNN$_{\text{spk}}$，我们使用 TCNDenseNet 架构 [2, 6] 的编码器和 TCN 模块（并移除解码器）来提取说话人嵌入。由于注册语句事先已知，该 DNN 可以采用非因果结构，并且可在针对每个目标说话人进行推理之前缓存其输出。给定单声道适配语句的复数频谱 A，该 DNN 对 TCN 模块的输出沿时间维度进行均值池化，从而提取固定长度的说话人嵌入 $X_{\text{adapt}} \in \mathbb{R}^{128}$。随后，iNeuBe 框架将该说话人嵌入作为额外线索，用于目标提取。与 [2, 6], 不同，我们使用规模较小的 DNN，其中包含 3 个 TCN 重复单元、4 个 TCN 块、128 个 TCN 通道；编码器的下采样二维卷积和 DenseNet 层中则各有 16 个隐藏通道。其参数总量约为 0.6 百万（M）。

1.3. 帧在线说话人条件化 MISO-TF-GridNet

对于 DNN$_1$ 和 DNN$_2$，我们均采用 TF-GridNet [1], 。该模型是当前先进的分离模型，最近在 WSJ0-2mix 数据集 [7] 上取得了令人瞩目的 23.4 dB 尺度不变信号失真比增量。受这一出色结果的启发，我们将该模型扩展到了多通道场景，如

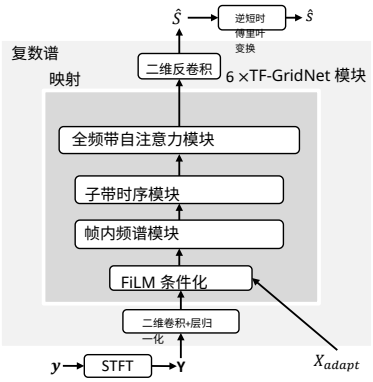

图2：提出的帧在线说话人条件化 MISO-TF-GridNet。

图2所示。由于该模型采用复数谱映射，我们只需将所有输入通道的实部和虚部（RI）分量 [2] 堆叠后输入 TF-GridNet，并将其转换为一个名为 MISO-TF-GridNet 的多输入单输出模型。我们在每个 TF-GridNet 模块的开头，对 DNN_{spk} 提取的说话人嵌入应用特征级线性调制（FiLM） [8]，以对分离过程进行条件控制。请注意， [1] 中提出的 TF-GridNet 是非因果的。为实现帧在线处理，我们进行了以下修改。首先，在第一个 Conv2D 层之后使用层归一化（LN）。其次，通过对注意力矩阵进行掩码，使用因果全频带自注意力模块。第三，我们采用了修改后的子带时序模块：使用单向 LSTM（而不是 [1] 中的双向 LSTM），并在展开操作中仅将前序帧的特征与当前帧堆叠。这样，当步长设置为 1 时，子带时序模块便可针对输入中的每一帧输出一帧。

我们采用 [1], 中建议的相同超参数，唯一的例外是此处将 E 设为 2。由此得到的 TF-GridNet 约有 800 万个参数，而完整的 iNeube-X 模型（包含 DNN_1 和 DNN_2）的参数量大约是其两倍。

1.4. STFT 域 NAL-R 与动态范围压缩

我们发现，仅靠增强不足以获得较高的 HASPI 分数。在未对听者听力损失进行任何补偿的情况下，理想目标语音在开发集上的 HASPI 分数低于 0.7。为了以逐帧在线方式进行拟合，我们在 STFT 域执行 NAL-R 拟合 [9] 和动态范围压缩。我们采用了 [10], 中实现的动态范围压缩算法，并分别将拐点设为 -10 dB、阈值设为 -40 dB 、压缩比设为 1.2 、拐点宽度设为 4 、启动时间设为 0.05 s、释放时间设为 0.2 s。由于 iNeuBe-X 使用 16 ms 的 STFT 窗，NAL-R 均衡滤波器的长度设为 80 个抽头。输入理想目标无混响语音后，我们的 STFT 域 NAL-R 拟合加压缩算法得到 0.987 的 HASPI 分数，略低于 CEC2 基线拟合算法得到的 0.998 分数。这可能是由于我们的算法使用了较短的 NAL-R 滤波器，但我们的理想分数仍然可以接受。

1.5. 系统训练配置与外部数据使用

我们在两个数据集上训练了模型：原始 CEC2 数据，我们使用提供的场景生成脚本将其扩展到 16k 个场景；以及这 16k 个场景加上另外生成的 6k 个场景。生成后者时使用了相同的配置，但外部干净语音取自 LibriVox，以增加说话人变化。

关于训练，我们首先使用 DNN_{spk} 训练 DNN_1 直至收敛，然后冻结这两个网络，并使用它们的输出以及 MCWFouput 训练 DNN_2。如第 1.1 节所述，这两个多输入单输出网络通过多通道混合信号，使用复数频谱映射来估计

表 1：CEC2 开发集上的结果。

方法	尺度不变信号失真比增量（dB）	短时客观可懂度	感知语音质量评估	HASPI
挑战赛基线（混合信号 CH1 左） [11]	0.0	0.563	1.197	0.245
挑战赛基线（预言机 CH1 左）	∞	1.0	5.0	0.998
BLSTM-加权预测误差+MVDR [12]	10.481	0.743	1.931	0.436
iNeuBe 深度神经网络$_1$ [2]	16.567	0.887	1.946	0.723
iNeuBe-X 深度神经网络$_1$	17.97	0.92	1.951	0.842
+深度神经网络$_2$ +NAL-R 微调	19.082	0.947	2.269	0.942
+深度神经网络$_2$ +NAL-R 微调 + 外部数据	19.514	0.954	2.336	0.943

目标语音在 CH1 左声道的无混响信号。我们使用 Adam优化器，学习率为 0.001，批量大小为 1。对于范数大于 1 的梯度，我们进行裁剪。多输入单输出深度神经网络以输入信号中随机截取的 4 秒片段进行训练，并确保目标语音和仅含干扰语音的起始部分都至少出现 1 秒。我们使用整个开发集进行验证。如果验证损失连续 5 个训练轮次未改善，则将学习率减半。我们采用 [2], 提出的尺度不变损失函数，该函数基于估计的时域信号及其幅度定义。与 [2], 不同，我们使用包含 512、1024、2048、256 和 128 个采样点窗口的多分辨率短时傅里叶变换滤波器组来计算多个幅度损失。我们发现，幅度损失能显著提升 HASPI。由于该损失具有尺度不变性，为了避免推理时发生削波，我们逐帧在线地使用多通道维纳滤波器的输出对多输入单输出深度神经网络的估计结果重新缩放。

为进一步提升 HASPI，在深度神经网络$_2$ 收敛后，我们通过添加一个附加损失项来微调深度神经网络$_2$。该损失项计算经 NAL-R 拟合的目标语音与经 NAL-R 拟合的深度神经网络估计结果之间的相同多分辨率损失。由于训练集中没有听力图，我们从开发集的听者中随机抽取听力图，以进行感知拟合的训练。

2. 开发集上的结果

我们在表 1 中报告了在 CEC2 开发集上取得的结果，指标包括尺度不变信号失真比（SI-SDR）、短时客观可懂度（STOI）、感知语音质量评估（PESQ）和 HASPI。需要注意的是，SI-SDR、STOI 和 PESQ 在听者适配和压缩之前计算，而 HASPI 则在之后计算。除所提出的 iNeuBe-X 外，我们还报告了两种非因果方法的性能：iNeube（仅基于 TCNDenseNet 的深度神经网络$_1$, ） [2]以及 ESPNet-SE++ 工具包中的一种基于深度神经网络的掩蔽波束形成方法 [12], ，该方法采用双层双向长短期记忆网络（BLSTM）和加权预测误差（WPE），随后进行最小方差无失真响应（MVDR）波束形成。作为另一种比较，我们报告了未经处理的混合语音和理想目标语音的得分。后者可以视为我们所提系统的合理上限，因为我们采用了类似的补偿策略。需要注意的是，CEC2基线系统不执行增强，仅进行听者适配和动态范围压缩。对于我们提出的系统，我们在第三个面板中报告了不同配置和训练策略下的性能：仅使用深度神经网络的 iNeube-X$_1$, 、使用听者适配后的附加损失进行训练的 iNeube-X（$_2$深度神经网络+$_2$NAL-R 微调），以及使用外部数据的系统（+ 深度神经网络$_2$ + NAL-R 微调 + 外部数据）。后两个系统是我们提交给 CEC2 的系统。我们发现，再进行一次波束形成并随后进行后滤波可以使 HASPI 略有提升，但代价是增加计算开销。因此，我们仅提交了在第一次深度神经网络$_2$ 细化步骤后得到的信号。

我们强调，我们的系统算法延迟为 4 ms，并通过这份公开代码[1]确认所提出的系统满足算法延迟不超过 5 ms 的限制。

[1]https://github.com/zqwang7/CausalityCheck

3. 致谢

S. Cornell 的部分研究经费由马尔凯大区资助，来自其资助项目“Miracle” POR MARCHE FESR 2014-2020。Z.-Q. Wang 使用了由美国国家科学基金会通过项目编号 ACI-1548562 资助的 Extreme Science and Engineering Discovery Environment [13]，，以及匹兹堡超级计算中心的 Bridges系统 [14]，，该系统由美国国家科学基金会通过项目编号 ACI-1445606 资助。Y. Masuyama 和 N. Ono 的部分研究经费由 JSPS KAKENHI项目编号JP21J21371 和 JST CREST项目编号JPMJCR19A3 资助。

4. 参考文献

[1] Z.-Q. Wang, S. Cornell, S. Cho, Y. Lee, B.-Y. Kim, 和 S. Watanabe，“TF-GridNet：让时频域模型在单声道说话人分离中再次出色”，载于投稿至 *ICASSP 2023*，2022。 [2] Y.-J. Lu, S. Cornell, X. Chang, W. Zhang, C. Li, Z. Ni, Z.-Q. Wang, 和 S. Watanabe，“迈向低失真多通道语音增强：ESPnet-SE 对 L3DAS22挑战赛的投稿”，载于*ICASSP* 会议论文集，2022，第 9201–9205 页。 [3] Z.-Q. Wang, G. Wichern, S. Watanabe, 和 J. L. Roux，“具有极低算法延迟的 STFT域神经语音增强”，*arXiv* 预印本 *arXiv:2204.09911*，2022。 [4] Z.-Q. Wang, P. Wang, 和 D. Wang，“用于逐话语和连续语音分离的多麦克风复数频谱映射”，*IEEE/ACM*《音频、语音与语言处理汇刊》，第 29 卷，第 2001–2014 页，2021。 [5] S. Gannot, E. Vincent, S. Markovich-Golan, 和 A. Ozerov，“多麦克风语音增强和源分离的综合视角”，*IEEE/ACM*《音频、语音与语言处理汇刊》，第 25 卷，第 692–730 页，2017。 [6] Z.-Q. Wang, G. Wichern, 和 J. Le Roux，“利用低失真目标估计改进语音增强”，*arXiv* 预印本 *arXiv:2110.00570*，2021。 [7] J. R. Hershey, Z. Chen, J. Le Roux, 和 S. Watanabe，“深度聚类：用于分割和分离的判别式嵌入”，载于*ICASSP* 会议论文集，2016，第 31–35 页。 [8] E. Perez, F. Strub, H. De Vries, V. Dumoulin, 和 A. Courville，“FiLM：具有通用条件层的视觉推理”，载于 *AAAI* 会议论文集，第 32 卷，第 1 期，2018。 [9] D. Byrne 和 H. Dillon，“国家声学实验室（NAL）用于选择助听器增益和频率响应的新程序”，《耳与听力》，第 7 卷，第 4 期，第 257–265 页，1986。 [10] L. McCormack, V. V¨alim¨aki 等，“基于 FFT 的动态范围压缩”，载于声音与音乐计算会议论文集，2017，第 5–8 页。 [11] X. Ren, L. Chen, X. Zheng 等，“用于 B 格式 3D 语音增强和识别的神经波束形成网络”，载于*MLSP* 会议论文集，2021。 [12] Y.-J. Lu, X. Chang, C. Li, W. Zhang, S. Cornell, Z. Ni, Y. Masuyama, B. Yan, R. Scheibler, Z.-Q. Wang 等，“ESPnet-SE++：用于鲁棒语音识别、翻译和理解的语音增强”，*Interspeech* 会议论文集，2022。 [13] J. Towns, T. Cockerill, M. Dahan 等，“XSEDE：加速科学发现”，《科学与工程计算》，第 16 卷，第 5 期，第 62–74 页，2014。 [14] N. A. Nystrom, M. J. Levine, R. Z. Roskies, 和 J. R. Scott，“Bridges：面向新社区和数据分析的独特灵活高性能计算资源”，载于*XSEDE* 会议论文集，2015，第 1–8 页。