提升互联网语音协议平台上的感知质量、可懂度和声学表现

Joseph Konan*†、Ojas Bhargave†、Shikhar Agnihotri†、Hojeong Lee†、Ankit Shah†、Shuo Han†、Yunyang Zeng†、Amanda Shu†、Haohui Liu†、Xuankai Chang†、Hamza Khalid†、Minseon Gwak†、Kawon Lee†、Minjeong Kim†、Bhiksha Raj†

*KonanAI，匹兹堡，美国，konan@konanai.com。

†卡内基梅隆大学，匹兹堡，{jkonan, obhargav, sagnihot, hojeongl, aps1, shuohan, yunyangz, amshu, haohuil, xuankaic, hkhalid, mgwak, kawonl, minjeong, bhikshar }@andrew.cmu.edu

arXiv:2303.09048v1 [cs.SD] 16 Mar 2023

摘要—**本文提出一种方法，用于微调在深度噪声抑制（DNS）2020 挑战赛数据上训练的模型，以提升其在互联网语音协议（VoIP）应用中的性能。我们的方法通过使 DNS 2020 模型适应 VoIP 通信特有的声学特征，来实现这一目标；这些特征包括由压缩、传输和平台特定处理造成的失真与伪影。为此，我们提出了一种面向 VoIP-DNS 的多任务学习框架，通过联合优化噪声抑制和 VoIP 特定声学特征来增强语音。我们在多种 VoIP 场景中评估了该方法，结果表明，它在 VoIP 应用语音增强方面优于业界性能水平和当前最先进的方法。我们的结果展示了通过 VoIP-DNS 改进在 DNS-2020 上训练的模型，并使其适用于不同 VoIP 平台的潜力；相关发现对于语音识别、语音助手和电信等领域具有重要的应用价值。**

索引词—**语音、声学、降噪、增强、VoIP**。

I. 引言

互联网语音协议（VoIP）已成为一种无处不在的技术，用于通过互联网进行实时语音通信。然而，说话者通话时所处的环境常有背景噪声，会降低感知质量并损害通信的可懂度。为解决这一问题，降噪和语音增强已成为改善噪声环境下语音信号质量的有前景的技术。由 INTERSPEECH 2020 主办的深度噪声抑制（DNS）挑战赛[1]，制定了系统的训练和评估方法，以提升先进降噪和语音增强模型在各种真实背景噪声场景中的性能。

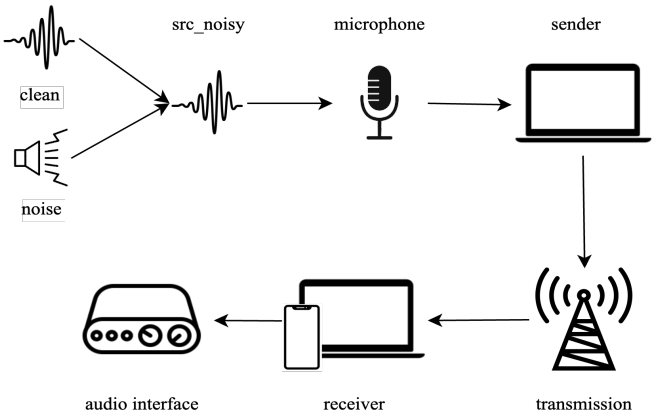
尽管在 DNS-2020 挑战赛数据上训练的模型取得了成功，但由于缺乏针对 VoIP 平台特性的优化，它们在 VoIP 应用中的表现仍然有限。VoIP 音频通常通过低带宽网络传输，使用有损编解码器进行压缩，并且采样率不一致，因此可能产生额外的失真和伪影。为支持这一研究领域的分析与优化，研究人员开发了 VoIP 深度噪声抑制（VoIP-DNS）数据集，作为 DNS-2020 的扩展，为 VoIP 应用 Zoom 和 Google Meet 提供训练和测试数据。研究人员采用受控的实验设计，使用一致的系统、硬件、网络和数据采集流程创建了 VoIP-DNS 数据集，以支持可复现研究。

本文介绍一种方法，对两种先进的 DNS-2020 模型 Demucs [2] 和 FullSubNet [3],进行微调，并在 VoIP-DNS数据集上提升其在 VoIP 应用中的表现。我们的方法通过时域声学参数损失（TAPLoss） [4] 优化细粒度语音特征，使模型适应 VoIP 音频流中的声学变化。我们在两种 VoIP 平台——Zoom 和 Google Meet——上评估了该方法，涵盖云端和蜂窝电话应用场景。结果表明，通过微调 Demucs 和 FullSubNet，可以超越行业性能，并在 VoIP-DNS 上达到先进水平，从而提升 VoIP 平台上的语音通信质量；这对语音识别、语音助手和电信领域具有重要应用价值。

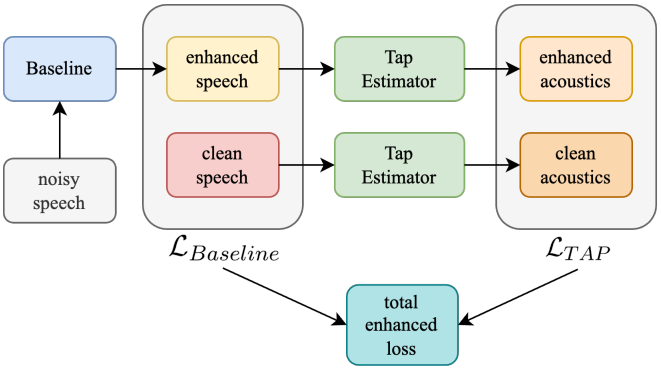
II. 实验设置

VoIP-DNS数据集 [5] 包含 1,200 对 30 秒的训练对和 150 对 10 秒的测试对，内容分别为源信号干净语音和源信号含噪语音。VoIP-DNS数据集的源信号对采用与 DNS-2020 相同的训练合成流程和测试集构建。训练合成流程包括从 LibriSpeech 中选取 30 秒的干净语音样本，以及从 FreeSound [6] 或 AudioSet [7],中选取 30 秒的噪声信号，然后将二者混合，生成 30 秒的含噪语音样本。源信号含噪语音通过虚拟麦克风传送到 VoIP 平台后，会被传输至接收设备。最后，通过音频接口获取转发后的音频。如果接收设备是蜂窝电话，则音频接口由 VoIP-DNS 技术人员管理；如果接收设备位于云端，则音频接口由 VoIP 平台管理。

本文重点研究如何优化 VoIP-DNS，以利用来自 Zoom 和 Google Meet 的互联网语音协议数据，提升云端和蜂窝电话上的语音质量。对于每个平台，我们都考虑开启或关闭降噪的使用场景。由于不同平台使用的术语有所差异，因此需要进行辨析。对于 Zoom，“关闭降噪”指“选择了低降噪级别的 Zoom”，而“开启降噪”指“Zoom



(a) 端到端数据处理流程图：从合成到录制



(b) Demucs 和 FullSubNet 的训练流程。

图 1：数据合成与训练流程

选择了自动降噪”。对于 Google Meet，“关闭降噪”指“选择了关闭降噪的 Google Meet”，而“开启降噪”指“选择了开启降噪的 Google Meet”。因此，工业降噪是根据平台开启降噪时的表现来定义的。

VoIP-DNS 数据收集流程捕捉了互联网语音协议音频流的独特特征，这些音频流受到低带宽、压缩、非均匀采样率以及额外的平台特定处理的影响。VoIP-DNS数据集由本文作者设计，并公开提供，同时详尽记录了实验设计，以确保结果可复现。

III. 模型与训练

本研究评估了两种最先进的深度学习模型 Demucs 和 FullSubNet 在互联网语音协议音频流上的降噪性能。选择这两种模型，是因为它们在 2020 年深度噪声抑制（DNS）挑战赛中表现出色，且其源代码、训练流程和最终检查点均公开可用。此外，Demucs 在时域中运行，而 FullSubNet 在时频域中运行，因此我们能够在一致且可复现的实验设置下比较这两种模态。

为了训练模型，我们使用 VoIP-DNS 中带噪语音中继和干净语音源。每次训练迭代都从中采样一段带噪语音信号及其对应的干净语音源。带噪语音信号经过基础模型处理后，生成增强语音信号。为了优化增强语音信号的细粒度声学特征，我们使用时序声学参数（TAP）估计器，从增强语音信号中提取增强声学特征。同样，我们使用 TAP 估计器从干净语音源中提取干净声学特征。

基础损失函数由基础模型定义，用于最小化增强语音信号与干净语音信号之间的散度，计算方式可以在时域（针对 Demucs）或时频域（针对 FullSubNet）进行。为了进一步提升模型在 VoIP 音频流上的性能，我们提出了一种新颖的声学损失函数，用于最小化增强声学特征与干净声学特征之间的 L1 散度。实验结果表明，采用声学损失函数可显著提升模型在 VoIP 音频流上的降噪性能。

除了评估 Demucs 和 FullSubNet 的性能外，我们还将其性能与 Zoom 和 Google Meet 等热门互联网语音协议平台所使用的工业级降噪模型进行比较。虽然这些工业级降噪模型无法公开检查，但可以通过相应平台使用 VoIP-DNS数据集进行查询。这使我们能够直接比较工业级降噪模型与应用于 VoIP 音频流的最先进深度学习模型的性能。

IV. 消融结果

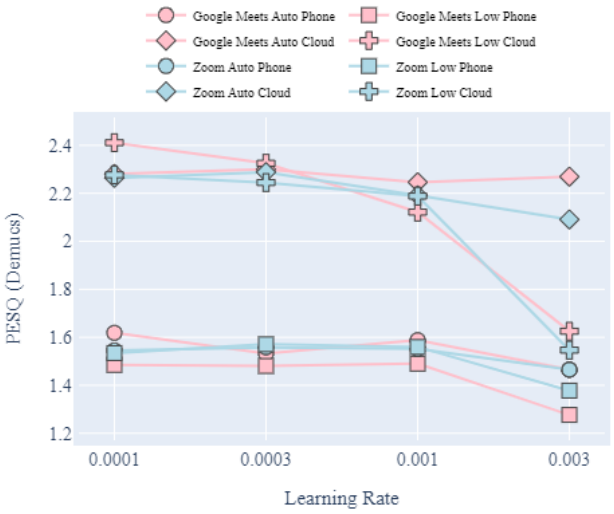


图 2：不同学习率下 Demucs 的消融实验

表 I：感知质量与可懂度的客观相对评估

		PESQ						STOI					
		源信号		Demucs		FullSubNet		源信号		Demucs		FullSubNet	
平台	接收器	Low	Auto	Low	Auto	Low	Auto	Low	Auto	Low	Auto	Low	Auto
Google Meet	电话	1.549	1.976	1.381	1.726	1.398	1.694	0.748	0.884	0.748	0.884	0.700	0.841
Google Meet	云端	1.640	2.255	2.272	2.264	2.435	2.281	0.890	0.923	0.933	0.924	0.920	0.923
Zoom	电话	1.548	1.701	1.548	1.513	1.382	1.343	0.797	0.810	0.766	0.810	0.733	0.724
Zoom	云端	1.450	1.919	2.089	1.996	2.141	1.992	0.859	0.909	0.906	0.913	0.900	0.911

我们开展了一系列消融实验，以评估所提出的时域声学参数（TAP）损失函数对模型降噪性能的影响。我们采用不同的学习率，分别在使用和不使用 TAP 损失函数的情况下进行训练，并在一个大规模数据集上评估模型性能。该数据集包含 1,200 个 30 秒的含噪语音样本，这些样本使用 DNS-2020 合成流程生成。

我们的实验结果表明，与基线模型相比，加入 TAP 损失函数能够提升降噪性能。具体而言，在采用最优学习率并加入 TAP 损失函数时，Demucs 和 FullSubNet 在所选指标上均取得了显著提升。不过，我们也观察到，TAP 损失函数的最优学习率会因互联网语音协议平台和传输过程的具体特征而异。

在某些情况下，我们发现，关闭工业降噪时，较高的学习率能够提升蜂窝设备上的降噪性能。这表明，在训练过程中采用更强的 TAP 损失函数，可能有助于蜂窝设备取得更好的降噪效果。相比之下，对于云中继传输以及启用了工业降噪的传输，较低的学习率能够提升降噪性能。

此外，我们观察到，所提出的降噪模型的效果会因互联网语音协议平台和传输过程的具体特征而异。根据平台、是否启用了工业降噪，以及接收音频的设备不同，可能会听到不同的声学失真和伪影。

在某些场景中，我们发现，在未启用工业降噪的情况下增强音频，能够获得更好的降噪效果。我们认为，这是因为平台处理造成的信息损失较少，从而有利于恢复原始音频。不过，在其他情况下，我们发现启用工业降噪能带来更好的降噪效果。这可能是因为某些失真和伪影在传输过程中会被放大，而且在未经工业降噪的音频中更加明显。在这些情况下，启用工业降噪或许会从音频中去除一些信息，但在传输前消除大部分失真和伪影所带来的益处，超过了由此产生的弊端。

综上所述，我们的结果表明，降噪模型在 VoIP 平台上的有效性高度取决于平台的具体特性、传输过程，以及训练过程中采用的最佳学习率。尽管如此，我们提出的降噪模型在关闭工业降噪时，始终优于工业级降噪模型这表明我们提出的模型有望提升 VoIP 平台上的降噪性能。

V. 感知评估

本节对微调结果进行感知分析，考察模型在不同 VoIP 应用中的可懂度和感知质量。为评估可懂度，我们采用短时客观可懂度（STOI） [8] 指标，该指标量化人类听众理解语音的程度。为评估感知质量，我们采用语音质量感知评估（PESQ） [9] 指标，该指标用于近似衡量人类对语音质量的主观判断。我们设有一个小节进行总体比较（Demucs 与 FullSubNet），并对低规格与自动规格进行差异分析。

A. 总体比较

我们的微调方法达到了或超过了行业可懂度水平。云端应用的差异不大，而蜂窝电话应用的 STOI 提升了 0.7 %。此外，与行业表现相比，我们观察到云端应用在 Google Meet 上的感知质量提升了 1.2%，在 Zoom 上提升了 4.0%。然而，移动应用的感知表现不如行业标准，这可能是因为将 DNS-2020 迁移到受蜂窝传输频带限制的环境中颇具挑战。

B. Demucs 与 *FullSubNet* 的比较

比较 Demucs 和 FullSubNet 后，我们发现 Demucs 的可懂度更高，其 PESQ 提升幅度从蜂窝应用的 0.2-1.1% 到云端应用的 5.1-10.5% 不等。感知质量的提升取决于接收媒介：在移动应用中，Demucs 的表现比 FullSubNet 高出 1.8-12.0%；而在云端应用中，FullSubNet 的表现则比 Demucs 高出 2.5-7.2%。虽然详细的误差分析超出了本文的讨论范围，但我们观察到，对时频域进行建模时的频带限制造成了明显的分布偏移，使得迁移学习在蜂窝应用上的收敛速度较慢。收集更多数据，尤其是蜂窝应用的数据，或许有助于克服这一挑战。

C. 关闭降噪（低）与开启降噪（自动）设置

为了更好地了解传输前未进行降噪的中继设备（low）与传输前进行了降噪的中继设备（auto）的相对性能，我们进行了差异分析。



(a) Demucs 标准化声学改善对比

(b) FullSubNet 标准化声学改善对比

图 3：平均绝对误差的声学改善：从 Google Meet 向蜂窝电话传输的用例

对于传输前未进行降噪的情况，我们观察到，各应用的感知质量更佳，PESQ 也有小幅提升，但 Google Meet 到蜂窝电话的情况除外。与 Zoom 相比，Google Meet 的感知质量更差，这是因为音频传输受到噪声干扰；PESQ 差异约为 0.14，这或许可以解释为何难以学习如何增强语音信号。

在 Demucs 内部比较（自动与低档）时，传输前经过降噪的中继设备，其可懂度与行业标准相当或更高。从 Google Meet 传输到蜂窝电话（提高 18.2% ）、从 Zoom 传输到蜂窝电话（提高 5.7%），以及在一定程度上从 Zoom 传输到云端（提高 0.7%）时，均观察到 STOI 有所改善。然而，从 Google Meet 传输到云端时，STOI 下降了 1.0%，降噪低档的表现因此优于自动档。我们推测，具体在从 Google Meet 到云端的用例中，低档设置加剧失真或伪影问题的程度，并不及传输前降噪造成的信息损失严重。这些细微差异与复杂性凸显了平台特定评估的重要性，因为需要改进的方面及其相对改进难度可能各不相同。

总之，我们的感知分析展现了将微调方法应用于不同 VoIP 平台和场景的潜在益处与挑战。结果强调了进行平台特定评估的重要性，以优化影响语音增强性能的因素。

VI. 声学评估

在本节中，我们将对降噪模型进行声学分析，重点关注其在改善含噪语音细粒度语音特征方面的效果，并与 Zoom 和 Google Meet 的工业级降噪模型及基线模型进行比较。我们使用扩展版 Geniva 极简声学参数集 (eGeMAPS) [10] forthis 进行评估。

对于每个测试样本，我们提取了原始干净语音、含噪语音以及由所提出模型、工业模型和基线模型生成的降噪语音的 eGeMAPS特征。随后，我们计算了原始干净语音的 eGeMAPS特征与语音退化版本和增强版本之间的平均绝对误差（MAE）。受篇幅所限，我们重点关注 Google Meet 到电话的声学改善，比较我们对 Demucs 和 FullSubNet 进行调优的方法与工业模型，并评估它们相较于各自开源基线的相对提升。

在几乎所有场景中，无论仅在传输前、仅在传输后，还是在传输前后都进行降噪，细粒度语音特征的声学表现均有所改善。我们的方法在达到工业水平方面表现不一，但与开源替代方案相比，使用 DNS-2020 对 Demucs 和 FullSubNet 进行调优并应用于 VoIP-DNS 数据集，FullSubNet 的提升最高可达 15%，Demucs 最高可达 40%；声学保真度平均分别提升约 5% 和 21%。

与行业表现相比，我们在振幅对数相对基频方面实现了更高的声学保真度。Demucs 的表现比行业水平约高 40%，比基线高 51%。在第一共振峰、第二共振峰和第三共振峰方面，FullSubNet 比基线略有提升，幅度为 4%，比行业水平高 40%。然而，在共振峰特定带宽的改善方面，Demucs 或 FullSubNet 与行业水平之间的差异微乎其微。

其他方面显著优于行业表现的声学指标包括响度（Demucs 提升 48%，FullSubNet 提升 47%）和谐噪比自相关函数（HNRdbACF）（Demucs 提升 39%，FullSubNet 提升 37%

），这表明我们的方法进行了适当的归一化，并能够区分谐波信号和非谐波信号。虽然还需进一步分析这些声学变化与感知质量之间的关系，但减少谐波失真并提高响度准确度或许有助于提升主观评价。

在梅尔频率倒谱系数、哈马伯格指数和阿尔法指数等某些指标上，Demucs 和 FullSubNet 难以改善声学表现。不过，与基线相比，我们的方法在这些方面的退化程度较轻。我们相信这些方面仍有进一步改进的空间，目前正在开展消融实验以优化模型。我们将在近期报告反映这些发现的更新结果。

第七节：结论

本研究考察了两种最先进的深度学习模型 Demucs 和 FullSubNet 在互联网语音协议音频流中的降噪性能，尤其关注它们在 Zoom 和 Google Meet 等热门平台上的应用。通过在 VoIP-DNS数据集上对这些模型进行微调，并引入一种新颖的声学损失函数，我们相较于其开源基线显著提升了降噪性能。

我们的感知分析表明，经过微调的模型在各种场景下，其可懂度能够达到或超过业界降噪模型；在 PESQ 和 STOI 指标方面，Demucs 通常优于 FullSubNet。对低规格与自动规格的差异分析显示，模型的表现会因中继条件和所用平台而异，这凸显了开展平台特定评估与优化的重要性。我们还观察到电话与云端应用之间存在性能差异，其中云端应用通常能带来更显著的感知质量提升。

使用 eGeMAPS特征进行的声学分析进一步证实了我们的方法在改善细粒度语音特征方面的有效性。与行业基线和开源基线相比，我们的模型在振幅对数相对基频、响度和 HNRdBACF 等某些声学特征上取得了显著提升。

不过，梅尔频率倒谱系数、哈马伯格指数和阿尔法指数等方面仍有进一步改进的空间。我们认为，进一步微调和开展消融研究将有助于解决这些问题，并在未来进一步提升降噪性能。

本研究为不断扩展的 VoIP 应用语音增强研究贡献了新的成果，并凸显了基于深度学习的降噪模型在提升实时通信质量方面的潜力。随着对有效远程通信的需求持续增长，本研究提出的进展为进一步开发和优化语音增强技术奠定了基础，使其能够适应各种 VoIP 平台、传输场景和终端用户设备（包括电话和云端应用）的具体需求。

参考文献

[1] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun 等，“2020年INTERSPEECH深度噪声抑制挑战赛：数据集、主观测试框架与挑战赛结果”，收录于*INTERSPEECH*，2020年。[2] A. Defossez, G. Synnaeve 和 Y. Adi，“波形域实时语音增强”，收录于*Interspeech*，2020年。[3] X. Hao, X. Su, R. Horaud 和 X. Li，“Fullsubnet：用于实时单通道语音增强的全频带与子频带融合模型”，收录于*ICASSP 2021-2021年IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2021年，第6633–6637页。[4] Zeng、Konan、Han、Bick、Yang、Kumar、Watanabe 和 Raj，“Taploss：用于语音增强的时序声学参数损失”，未发表手稿，2022年。[5] “VoIP-DNS数据集”，https://github.com/deepology/ VoIP-DNS-Challenge，访问日期：2023年3月14日。[6] E. Fonseca、J. Pons Puig、X. Favory、F. Font Corbera、D. Bogdanov、A. Ferraro、S. Oramas、A. Porter 和 X. Serra，“Freesound数据集：一个用于创建开放音频数据集的平台”，收录于*Hu X*、*Cunningham SJ*、*Turnbull D*、*Duan Z*，编者。《第*18*届国际音乐信息检索学会会议论文集》；*2017*年*10*月*23*日至*27*日；中国苏州。*[*加拿大*]*：国际音乐信息检索学会；*2017*年。第*486–493*页。国际音乐信息检索学会（ISMIR），2017年。[7] J. F. Gemmeke、D. P. Ellis、D. Freedman、A. Jansen、W. Lawrence、R. C. Moore、M. Plakal 和 M. Ritter，“Audio Set：一种用于音频事件的本体和人工标注数据集”，收录于*2017*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2017年，第776–780页。[8] C. H. Taal、R. C. Hendriks、R. Heusdens 和 J. Jensen，“一种用于时频加权含噪语音的短时客观可懂度测量方法”，收录于*2010*年*IEEE*声学、语音与信号处理国际会议，2010年，第4214–4217页。[9] A. Rix、J. Beerends、M. Hollier 和 A. Hekstra，“语音质量感知评估（PESQ）——一种用于评估电话网络和编解码器语音质量的新方法”，收录于*2001*年*IEEE*声学、语音与信号处理国际会议论文集（目录号：*01CH37221*），第2卷，2001年，第749–752页，第2卷。[10] F. Eyben、K. R. Scherer、B. W. Schuller、J. Sundberg、E. Andr´e、C. Busso、L. Y. Devillers、J. Epps、P. Laukka、S. S. Narayanan 等，“用于语音研究与情感计算的日内瓦极简声学参数集（GeMAPS）”，*IEEE*情感计算汇刊，第7卷，第2期，第190–202页，2015年。

附录

A. 时域声学参数损失（*TAPLoss*）

为计算 TAPLoss，我们定义时序声学参数估计器 A_y。$A_y(t,\ p)$ 表示离散时间帧 t 上的声学参数 p。为预测 A_y，我们定义该估计器：

$$\hat{A_y} = \mathcal{TAP}(y)$$

$\mathcal{TAP}$ 函数接收信号输入 y。它计算包含 $F = 257$ 个频率分箱的复数频谱图 Y。随后，它将该复数频谱图输入循环神经网络 $\hat{}$，输出时序声学参数 A_y。TAP损失定义为实际值与预测值之间的平均绝对误差。

$$MAE(A_y, \hat{A_y}) = \frac{1}{TP}\sum_{t=0}^{T-1}\sum_{p=0}^{P-1}|A_y(t,p) - A_{\hat{y}}(t,p)| \in R$$

训练期间，$\mathcal{TAP}$ 参数通过 Adam优化，学习最小化 $MAE(A_s,\ \hat{}A_s)$ 的散度。由于该损失可端到端微分，且仅以波形作为输入，因此能够利用干净参考信号，对任何语音模型和任务进行声学优化。