arXiv:2302.08088v1 [cs.CL] 16 Feb 2023

TAPLOSS：一种时序 用于语音增强的声学参数损失

Yunyang Zeng[1†]、*Joseph Konan*[1†]、*Shuo Han*[1†]、*David Bick*[1†]、*Muqiao Yang*[1†]、*Anurag Kumar*[2],*Shinji Watanabe*[1],*Bhiksha Raj*[1]

[1] 卡内基梅隆大学， [2]Meta Reality Labs Research

摘要

近年来，语音增强模型取得了长足进步，但其语音输出的感知质量仍存在局限。我们提出一种基于时序声学参数的感知质量客观指标。这些参数是语音的基本特征，在说话人识别和副语言分析等多种应用中发挥着重要作用。我们提供了一种可微分的估计器，用于估计四类低层声学描述符，包括频率相关参数、能量或振幅相关参数、频谱平衡参数以及时序特征。不同于以往研究只关注聚合声学参数或少数几类声学参数，我们提出的时序声学参数（TAP）损失能够在增强工作流程中辅助优化并改善多种细粒度语音特征。我们展示了在语音增强中将TAPLoss 用作辅助目标，可提升语音的感知质量和可懂度。我们使用来自2020年深度噪声抑制挑战赛的数据，证明时域模型和时频域模型均可受益于我们的方法。

索引词— 语音，增强，声学，感知质量，可解释增强评估，可解释性

1. 引言

语音增强旨在提升退化语音信号的质量和可懂度。在各种语音应用中都有这方面的需求。尽管噪声抑制或消除是语音增强的重要组成部分，但保留语音信号的感知质量同样重要。近年来，基于深度神经网络的方法已成为大多数最先进语音增强系统的核心，尤其是单通道语音增强 [1]–[4]。传统上，这些系统使用时域或时频域中的逐点差异进行训练。

然而，许多研究表明，这些损失存在局限性，包括与语音质量 [5] 和 [6]的相关性较低。其他研究表明，这些损失过度强调高能量音素 [7] ，并且无法改善基频 [8], ，从而导致语音出现伪影或感知质量较差 [9]。这些损失的不足促使大量研究致力于提升增强语音信号的感知质量，这也是我们工作的目标之一。感知损失通常涉及估计标准评估指标，例如语音质量感知评估（PESQ） [10]。然而，PESQ 不可微，这使得优化过程十分困难 [11]，并且通常只能带来有限的改进 [12], [13]。其他方法则使用深度特征损失 [14], [15]；然而，由于感知质量仅受到隐式监督，这些方法带来的改进十分有限。在本文中，我们试图通过使用基础语音特征来解决这些问题，并将其称为声学参数。

研究表明，使用声学参数有助于说话人分类、情感识别以及其他涉及语音特征的监督任务 [16]–[18]。过去，由于声学参数需要不可微的计算，因此未将其纳入深度神经网络的工作流程。然而，这与既有文献中声学参数和语音音质之间显著相关这一事实并不相符 [19]–[21]。近来，一些研究在将声学参数用于深度神经网络优化方面取得了进展。文献 [22]提出使用基频、能量轮廓和基频轮廓来优化感知质量。然而，这三个参数只是我们所考虑特征的一小部分，而且评估并未在标准英语数据集上进行。文献 [23], 提出了广泛的声学参数，并引入了这些参数的可微估计器，以构建一种辅助损失，促使模型保留声学参数。事实证明，这有助于提升语音的感知质量。不同于以往按话语计算声学参数汇总统计量的工作，我们当前的估计器能够在每个时间步进行优化。话语中的各项参数值会随时间变化，因此，在比较干净语音和增强语音时，统计量会损失大量信息。

我们考察了 25 个声学参数——**频率相关参数**：基频、抖动、F1、F2、F3 频率和带宽；**能量或振幅相关参数**：微扰、响度、谐波噪声比（HNR）；**频谱平衡参数**：阿尔法比、汉马伯格指数、频谱斜率、F1、F2、F3 相对能量、谐波差；以及其他**时域参数**：响度峰值率、有声/无声区段长度的均值和标准差，以及每秒连续有声区段数。我们使用 OpenSmile [24] 执行真实值的非可微计算，创建数据集以训练可微估计器。

最后，我们介绍了用于估计这 25 个时域声学参数的估计器。利用该估计器，我们定义了一种声学参数损失，称为 TAPLoss， $\mathcal{L}_{\text{TAP}}$，用于最小化干净语音与增强语音的估计声学参数之间的距离。与以往工作不同，我们不假设用户能够获取真实值的干净声学参数。我们通过实证展示了该方法的有效性：感知质量和可懂度的相对及绝对增强指标均有所提升。

2. 方法

2.1. 背景

在时域中，令$\mathbf{y}$ 表示一个离散时长为 M的信号，使得$\mathbf{y} \in \mathbb{R}^M$。我们定义干净语音信号$\mathbf{s}$、噪声信号$\mathbf{n}$,

†贡献相同（随机排序）

	Demucs $\mathcal{L}_{\text{TAP}}$ λ_1 消融（$\lambda_2 = 0$）					Demucs $\mathcal{L}_{\text{TAP}}$ λ_2 消融（$\lambda_1 = 1$）					FullSubNet $\mathcal{L}_{\text{TAP}}$ γ 消融				
权重	0.01	0.03	0.1	0.3	1	0.01	0.03	0.1	0.3	1	0.01	0.03	0.1	0.3	1
PESQ	2.788	2.841	2.824	2.834	**2.859**	2.899	2.903	2.926	**2.958**	2.958	2.979	**2.981**	2.979	2.969	2.965
STOI	0.9697	0.9698	0.9689	0.9689	**0.9694**	0.9707	0.9712	0.9714	**0.9722**	0.9720	0.9654	**0.9654**	0.9654	0.9648	0.9654

表 1：$\mathcal{L}_{\text{TAP}}$ 消融研究　Demucs 超参数的 y、λ_1 和 λ_2，以及 FullSu　bNet 超参数 γ。

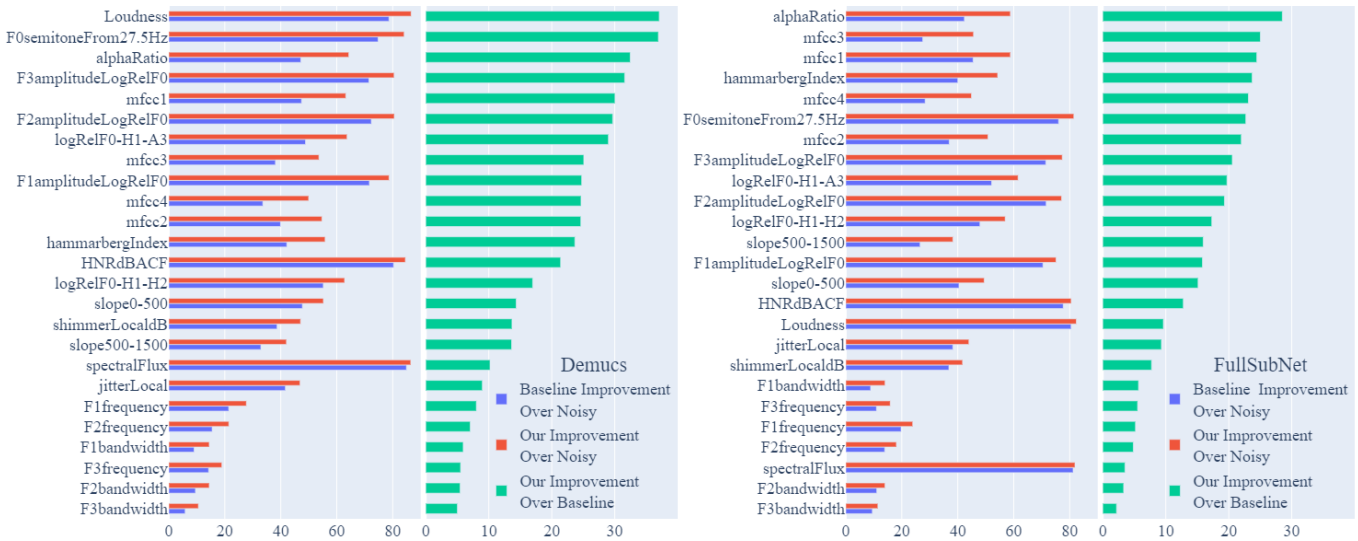

图 1：DNS-2020 合成测试（无混响）中的声学改善百分比 **PAI**。比较的是基线相对于含噪语音的改善（蓝色）$\mathbf{PAI(s_1, x)}$、我们相对于含噪语音的改善（红色）$\mathbf{PAI(s_2, x)}$，以及我们相对于基线的改善（绿色）$\mathbf{PAI(s_2, s_1)}$。按$\mathbf{PAI(s_2, s_1)}$排序。

以及带噪语音信号$\mathbf{x}$，其满足以下加性关系：

$$\mathbf{x} = \mathbf{s} + \mathbf{n} \tag{1}$$

类似地，在时频域中，令$\mathbf{Y} \in \mathbb{C}^{T\times F}$表示一个包含 T个离散时间帧和 F个离散频率分箱的复数频谱图。 $\mathfrak{Re}\{\mathbf{Y}\} \in \mathbb{R}^{T\times F}$表示实部，$\mathfrak{Im}\{\mathbf{Y}\} \in \mathbb{R}^{T\times F}$表示虚部。令 $Y(t, f)$为$\mathbf{Y}$在离散时间帧$t \in [0, T)$和离散频率分箱 $f \in [0, F)$处的复数时频分箱。根据傅里叶变换的线性性质，干净语音$\mathbf{S}$、噪声$\mathbf{N}$和噪声语音$\mathbf{X}$ 的复数频谱图满足以下加法关系：

$$\mathbf{X} = \mathbf{S} + \mathbf{N} \tag{2}$$

语音增强模型 G 输出增强信号$\hat{}\mathbf{s}$，使得：

$$\left\{\hat{\mathbf{s}} = G(\mathbf{x}) \middle| \hat{\mathbf{s}}, \mathbf{x} \in \mathbb{R}^M\right\} \tag{3}$$

在优化过程中， G 最小化$\mathbf{s}$ 与$\hat{}\mathbf{s}$之间的散度。我们将由$\hat{}\mathbf{s}$得到的增强复数频谱图记为$\hat{}\mathbf{S}$。

2.2. 时间声学参数估计器

令$\mathbf{A_y} \in \mathbf{R}^{T\times 25}$ 表示信号$\mathbf{y}$ 的 25 个时域声学参数，其中包含 T 个离散时间帧。我们将 $A_{\mathbf{y}}(t, p)$ 表示为离散时间帧 t处的声学参数 p。我们对声学参数进行标准化，使其在时间维度上的均值为 0、方差为 1。标准化有助于通过确保各特征的单位一致来进行优化和分析。为预测$\mathbf{A_y}$，我们定义估计器：

$$\hat{\mathbf{A}}_{\mathbf{y}} = \mathcal{TAP}(\mathbf{y}) \tag{4}$$

$\mathcal{TAP}$ 接收信号输入$\mathbf{y}$，生成包含$F = 257$ 个频率分箱的复数频谱图$\mathbf{Y}$，然后将该复数频谱图输入循环神经网络，以输出时域声学参数的估计值$\hat{}\mathbf{A_y}$。

为计算损失，我们将总平均绝对误差定义为：

$$\text{MAE}(\mathbf{A_y}, \hat{\mathbf{A}}_{\mathbf{y}}) = \frac{1}{TP}\sum_{t=0}^{T-1}\sum_{p=0}^{P-1}|A_{\mathbf{y}}(t, p) - A_{\hat{\mathbf{y}}}(t, p)| \in \mathbb{R} \tag{5}$$

在训练过程中， $\mathcal{TAP}$ 参数通过 Adam 优化来学习最小化 $\text{MAE}(\mathbf{A_s}, \mathbf{A_{\hat{s}}})$ 的散度。

2.3. 时间声学参数损失

我们提出了时间声学参数损失 $\mathcal{L}_{\text{TAP}}$，用于最小化干净声学参数与增强声学参数之间的差异。本节将阐述 $\mathcal{L}_{\text{TAP}}$ 的数学公式。

设幅度频谱图表示复数频谱图 $\hat{}\mathbf{S}$ 的幅度。根据帕塞瓦尔定理，帧能量权重 $\boldsymbol{\omega}$ 可由沿频率轴计算的幅度频谱图均值得出：

$$\boldsymbol{\omega} = \frac{1}{F}\sum_{f=0}^{F-1}||\hat{S}(t, f)||^2 \in \mathbb{R}^T \tag{6}$$

由于高能量更容易被感知，我们使用 sigmoid 函数 σ，以有界尺度模拟人类听觉，从而得到平滑能量权重 $\sigma(\omega)$。

最后，我们将时间声学参数损失定义为 $\mathcal{L}_{\text{TAP}}$，即干净声学参数估计值与增强声学参数估计值之间经平滑帧能量加权的平均绝对误差：

$$\mathcal{L}_{\text{TAP}}(\mathbf{s}, \hat{\mathbf{s}}) = \text{平均绝对误差}\left(\mathcal{TAP}(\mathbf{s}) \odot \sigma(\omega)\ \mathcal{TAP}(\hat{\mathbf{s}}) \odot \sigma(\omega)\right) \quad (7)$$

这里，“⊙”表示带广播的逐元素乘法。请注意，该损失可进行端到端微分，且仅以波形作为输入。因此，借助干净参考信号，该损失可用于对任何语音模型和任务进行声学优化。

3. 实验

3.1. TAPLoss工作流程

本节介绍将 TAPLoss 应用于语音增强模型的工作流程。为了证明我们的方法能够泛化到时域和时频域模型，我们将 TAPLoss $\mathcal{L}_{\text{TAP}}$ 应用于两种具有竞争力的语音增强模型：Demucs [25] 和 FullSubNet [26]。Demucs 是一种基于映射的时域模型，采用编码器-解码器结构，以含噪波形作为输入，并输出估计的干净波形。FullSubNet 是一种基于掩码的时频域融合模型，将全频带模型与子频带模型相结合。FullSubNet 根据输入信号的复数频谱图估计复数理想比率掩码（复数理想比率掩码），并将该复数理想比率掩码与输入信号的复数频谱图相乘，从而获得增强信号的复数频谱图。增强后的复数频谱图通过逆短时傅里叶变换（逆短时傅里叶变换）转换到时域。

我们的目标是使用 $\mathcal{L}_{\text{TAP}}$ 微调这两个基线增强模型，以提升其感知质量和可懂度。在前向传播过程中，增强模型以含噪信号为输入，并输出增强信号。TAP 估计器会预测干净信号和增强信号的时域声学参数。随后，使用上一小节讨论的方法计算 $\mathcal{L}_{\text{TAP}}$。Demucs 和 FullSubNet 也各有自己的损失函数。FullSubNet 使用估计的复数理想比率掩码（cIRM）与真实 cIRM 之间的均方误差（MSE）作为损失（$\mathcal{L}_{\text{cIRM}}$）。Demucs 有两个损失函数：L1波形损失（$\mathcal{L}_{\text{wave}}$）和多分辨率短时傅里叶变换损失（$\mathcal{L}_{\text{STFT}}$）。在 DNS 2020数据集上预训练的基线 Demucs 模型仅使用 L1波形损失。为了公平比较，我们首先使用 L1波形损失和$\mathcal{L}_{\text{TAP}}$微调 Demucs。然而，先前研究表明，Demucs 模型容易产生音调伪影 [27]，而且我们在使用 L1波形损失和 $\mathcal{L}_{\text{TAP}}$进行微调时也观察到了这一现象。此外，我们发现多分辨率短时傅里叶变换损失能够缓解这一问题，因为音调伪影引入的误差在时频域中比在时域中更加显著、明显。因此，在取得最佳微调结果后，我们再次使用 L1波形损失、 $\mathcal{L}_{\text{TAP}}$和多分辨率短时傅里叶变换损失进行微调，以消除音调伪影。以下公式给出了微调 Demucs 和 FullSubNet 所用的最终损失函数，其中 λ_1、 λ_2和 γ 表示权重超参数：

$$\mathcal{L}_{\text{Demucs}} = \mathcal{L}_{\text{wave}} + \lambda_1 \cdot \mathcal{L}_{\text{TAP}} + \lambda_2 \cdot \mathcal{L}_{\text{STFT}} \quad (8)$$

$$\mathcal{L}_{\text{FullSubNet}} = \mathcal{L}_{\text{cIRM}} + \gamma \cdot \mathcal{L}_{\text{TAP}} \quad (9)$$

反向传播期间，TAP 估计器的参数被冻结，仅优化增强模型的参数。

3.2. 数据

本研究使用 2020年 DeepNoise Suppression 挑战赛(DNS)数据 [28]，，其中包括干净语音（来自 Librivox 语料库）、噪声（来自 Freesound 和 AudioSet [29]）以及噪声语音合成方法。我们合成了时长为 30 秒的干净语音与含噪语音配对样本，包括 50,000 个训练样本和 10,000 个开发样本。在实验中，我们使用官方的无混响合成测试集，其中包含 150 个时长为 10 秒的样本。

3.3. 实验细节与消融

我们对 Demucs 的官方预训练检查点进行微调 [1] 和 Full-SubNet [2]。我们使用声学权重λ_1 对 Demucs 微调 40 个周期，然后使用频谱图权重 λ_2再微调 10 个周期。我们使用声学权重 γ对 FullSub Net 微调 100 个周期。TAP 和 TAPLoss 的源代码将在 https://github.com/ YunyangZeng/TAPLoss开放。在我们的实验中，估计器架构为一个具有 256 个隐藏单元的三层长短期记忆（LSTM）双向循环神经网络。经过 200 个周期后，声学参数估计器收敛，训练误差和验证误差均为 0.15。

作为一种辅助损失， $\mathcal{L}_{\text{TAP}}$ 需要通过消融实验确定最优的超参数设定。我们发现，在时域模型中，较高的声学权重有利于 Demucs。然而，在聆听音频并查看频谱图时，我们也观察到了音调伪影。经过调查发现，这些伪影由模型架构导致，是某些转置卷积设定中已知的问题。为了解决音调问题，我们在给定声学权重的情况下进行了另一项消融实验，以改善频谱图。我们发现，较高的频谱图权重有所帮助，但其重要性不及优化声学。在时频域中，声学权重取 0.03 时，FullSubNet 的效果最佳。值得注意的是，Demucs 的权重取 1、FullSubNet 的权重取 0.03，分别对应了二者的尺度差异。

3.4. 声学评估

设干净语音目标为 $\mathbf{s}$，噪声语音输入为 $\mathbf{x}$，基线增强语音为$\hat{\mathbf{s}}_1$，，我们的增强语音为 $\hat{\mathbf{s}}_2$。令 $\mathbf{MAE}_0$ 表示沿时间轴（轴 0）的平均绝对误差，并令 $\oslash$ 表示逐元素除法。百分比声学改善 $\mathbf{PAI}$定义如下：

$$\epsilon_{\mathbf{m},\mathbf{m}'} \triangleq \mathbf{MAE}_0(\mathbf{A}_{\mathbf{m}}, \mathbf{A}_{\mathbf{m}'}) \quad (10)$$

$$\mathbf{PAI}(\mathbf{A}_{\hat{\mathbf{s}}_1}, \mathbf{A}_{\mathbf{x}}) = 100\% \cdot (1 - \epsilon_{\hat{\mathbf{s}}_1,\mathbf{s}} \oslash \epsilon_{\mathbf{x},\mathbf{s}}) \quad (11)$$

$$\mathbf{PAI}(\mathbf{A}_{\hat{\mathbf{s}}_2}, \mathbf{A}_{\mathbf{x}}) = 100\% \cdot (1 - \epsilon_{\hat{\mathbf{s}}_2,\mathbf{s}} \oslash \epsilon_{\mathbf{x},\mathbf{s}}) \quad (12)$$

$$\mathbf{PAI}(\mathbf{A}_{\hat{\mathbf{s}}_2}, \mathbf{A}_{\hat{\mathbf{s}}_1}) = 100\% \cdot (1 - \epsilon_{\hat{\mathbf{s}}_2,\mathbf{s}} \oslash \epsilon_{\hat{\mathbf{s}}_1,\mathbf{s}}) \quad (13)$$

声学评估包含三个部分：（1）基线相较于噪声语音输入的改进，$\mathbf{PAI}(\mathbf{A}_{\hat{\mathbf{s}}_1},\mathbf{A}_{\mathbf{x}})$；（2）我们的系统相较于噪声语音输入的改进，$\mathbf{PAI}(\mathbf{A}_{\hat{\mathbf{s}}_2}, \mathbf{A}_{\mathbf{x}})$；以及（3）我们的系统相较于基线的改进，$\mathbf{PAI}(\mathbf{A}_{\hat{\mathbf{s}}_2},\mathbf{A}_{\hat{\mathbf{s}}_1})$。

声学改进用于衡量增强过程将含噪输入转换为清晰音频输出的效果。改进幅度为 0% 表示增强没有改变含噪声学，而 100% 表示达到最大可能改进，即增强声学与清晰声学完全相同。相对声学改进用于衡量增强微调能在多大程度上产生更清晰的音频输出。改进幅度为 0% 表示微调后 TAPLoss 没有改变增强声学。改进幅度为 100% 表示经 TAPLoss 增强后的声学听起来与清晰声学完全相同。

[1] https://github.com/facebookresearch/denoiser [2] https://github.com/haoxiangsnr/FullSubNet

指标	所用损失函数	窄带 PESQ	宽带 PESQ	STOI	ESTOI	CD	LLR	WSS	OVRL	BAK	SIG	NORESQA
干净	–	–	–	–	–	–	–	–	**3.28**	**4.04**	**3.56**	**4.61**
含噪	–	2.454	1.582	0.915	0.810	12.623	0.577	35.546	2.48	2.62	3.39	2.99
Demucs	$\mathcal{L}_{\text{wave}}$	3.272	2.652	0.965	0.921	17.138	0.443	18.239	3.31	**4.15**	3.54	3.95
Demucs	$\mathcal{L}_{\text{wave}} + \lambda_1 \mathcal{L}_{\text{TAP}}$	3.356	2.859	0.969	0.930	17.803	0.334	23.442	3.15	3.78	**3.58**	**4.12**
Demucs	$\mathcal{L}_{\text{wave}} + \lambda_1 \mathcal{L}_{\text{TAP}} + \lambda_2 \mathcal{L}_{\text{STFT}}$	**3.409**	**2.958**	**0.972**	**0.934**	**18.298**	**0.312**	**14.392**	**3.34**	4.14	3.57	4.08
FullSubNet	$\mathcal{L}_{\text{cIRM}}$	3.386	2.889	0.964	0.920	16.962	0.399	20.887	3.21	4.02	3.51	4.09
FullSubNet	$\mathcal{L}_{\text{cIRM}} + \gamma \mathcal{L}_{\text{TAP}}$	**3.417**	**2.981**	**0.965**	**0.922**	**17.677**	**0.310**	**18.946**	**3.25**	**4.05**	**3.53**	**4.14**

表 2：语音增强质量的相对指标和绝对指标，与 DNS-2020 测试集（无混响）上的基线进行比较 $\mathcal{L}_{\text{TAP}}$。

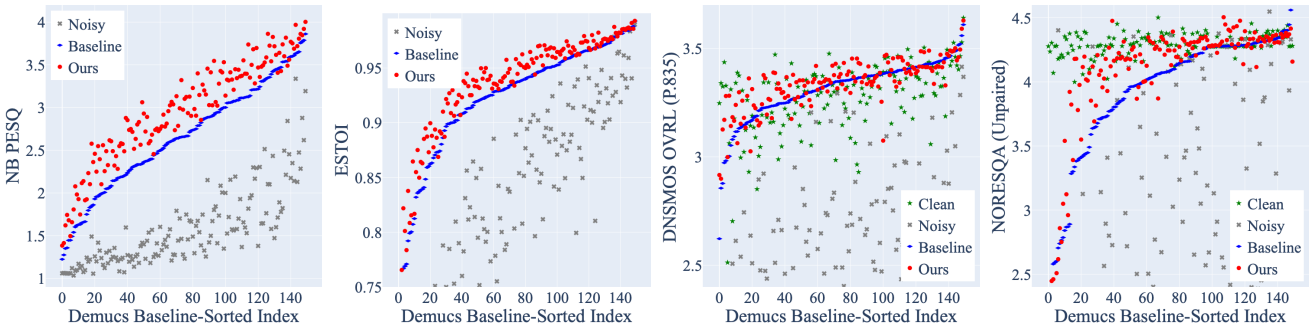

图 2：对比 s　　选定的相对指标和绝对指标，以及 DNS-2020 测试集（无混响）上最终的 $\mathcal{L}_{\text{TAP}}$ Demucs 与基线的比较　　b).

图 1 展示了 Demucs 和 FullSubNet 的声学改进百分比。平均而言，使用 $\mathcal{L}_{\text{wave}}$, $\mathcal{L}_{\text{TAP}}$ 和 $\mathcal{L}_{\text{STFT}}$ 的 Demucs 将含噪声学改善了 53.9%，而基线 Demucs 将其改善了 44.9%。使用 $\mathcal{L}_{\text{cIRM}}$, $\mathcal{L}_{\text{TAP}}$ 的 FullSubNet 将含噪声学改善了 50.3%，而基线 FullSubNet 将其改善了 42.6%。平均而言，TAPLoss 将 Demucs 基线声学改善了 19.4%，将 FullSubNet 基线声学改善了 14.5%。

作为一种分析工具，声学分析能够分解语音增强质量的变化，识别有待改进的潜在架构或优化标准。例如，Demucs 和 FullSubNet 架构在优化共振峰频率和带宽方面存在困难。因此，这些实证结果表明，未来若引入相关的数字信号处理机制，或可提升声学保真度优化能力。本文提供了一个用于声学分析和优化的框架，为理解和改善声学表现与感知质量提供了必要工具。

3.5. 感知评估

增强效果评估包括用于比较信号的相对指标，以及用于评估单个信号的绝对指标。相对指标包括短时客观可懂度（STOI）、扩展短时客观可懂度（ESTOI）、倒谱距离（CD）、对数似然比（LLR）、加权谱斜率（WSS）、宽带（WB）和窄带（NB）语音质量感知评估（PESQ） [30]。绝对指标包括 DNS-MOS P. 中的总体（OVRL）、信号（SIG）和背景（BAK）835 [31]。最后，基于非匹配参考的语音质量评估（NORESQA）包括其绝对（非配对）和相对（配对）平均主观意见分估计 [32]。

许多增强评估指标都受益于使用 TAPLoss 对声学参数进行显式优化。语音质量感知评估中的窄带（PESQ-NB）和宽带（PESQ-WB）指标提升最为显著。虽然 STOI 在时频域模型中提升不大，但在时域模型中有小幅提升。

大多数 DNSMOS 指标（OVRL、SIG、BAK）在时频域和时域模型中均有所提升；然而，增强后的表现优于干净语音，这表明该指标并不可靠。在这一时域应用中，NORESQA 显著提升，不过对于采用时域模型的源模型，显式优化频谱图会损害该指标。根据这项实证分析，对于旨在提升感知质量的任务，我们建议使用 TAPLoss。未来的工作或可根据特定指标的优化目标对声学参数进行加权，从而显著提升感知质量。

图2展示了更多细节，有助于我们分析这150个十秒样本。为了便于成对比较，我们按照基线增强语音评估结果进行排序。相比之下，我们的增强语音在两个相对指标 NB-PESQ 和 ESTOI 上优于基线增强语音。在分析 NORESQA 时也可以观察到类似的模式。对于 NORESQA，我们的增强语音大多优于基线增强语音。

4. 结论

TAPLoss 能够提升时域和时频域语音增强模型的声学保真度。与聚合声学参数相比，优化时序声学参数能够带来更好的增强效果评估，并显著提升声学改进。此外，使用 TAPLoss 实现声学改进有着坚实的数字信号处理理论基础，可为后续针对性开发提供指导，包括通过受声学原理启发的架构调整或损失优化来提升语音增强效果。

5. 致谢

本研究使用了极端科学与工程发现环境（XSEDE） [33], , 该环境由美国国家科学基金会资助，资助编号为 ACI-1548562。具体而言，本研究使用了匹兹堡超级计算中心（PSC）的 Bridges系统 [34], , 该系统由美国国家科学基金会资助，资助编号为 ACI-1445606。

6. 参考文献

[1] F. Weninger、H. Erdogan、S. Watanabe、E. Vincent、J. Le Roux、J. R. Hershey 和 B. Schuller，“采用 LSTM 循环神经网络进行语音增强及其在噪声鲁棒自动语音识别中的应用”，载于潜变量分析与信号分离，E. Vincent、A. Yeredor、Z. Koldovsk´y 和 P. Tichavsk´y 编，沙姆：施普林格国际出版公司，2015年，第 91–99 页。[2] S. Pascual、A. Bonafonte 和 J. Serr`a，“SEGAN：语音增强生成对抗网络”，*arXiv* 预印本 *arXiv:1703.09452*

，2017年。[3] D. Rethage、J. Pons 和 X. Serra，《用于语音降噪的 WaveNet》，载于*ICASSP* 会议论文集，电气与电子工程师学会，2018年，第5069–5073页。[4] D. S. Williamson、Y. Wang 和 D. Wang，《用于单耳语音分离的复数比掩蔽》，*IEEE/ACM*音频、语音与语言处理汇刊，第24卷，第3期，第483–492页，2016年。[5] P. Manocha、A. Finkelstein、R. Zhang、N. J. Bryan、G. J. Mysore 和 Z. Jin，《基于最小可觉差学习的可微感知音频度量》，*Interspeech* 会议论文集，2020年。[6] S.-W. Fu、C. Yu、T.-A. Hsieh、P. Plantinga、M. Ravanelli、X. Lu 和 Y. Tsao，《Metricgan+：一种改进的语音增强 MetricGAN 版本》，*Interspeech* 会议论文集，2021年。[7] P. Plantinga、D. Bagchi 和 E. Fosler-Lussier，《用于单通道增强和鲁棒自动语音识别的识别模型感知损失》，*arXiv*预印本 *arXiv:2112.06068*，2021年。[8] J. Turian 和 M. Henry，《深表遗憾：基于频谱的音频距离不善于衡量基频》，*arXiv*预印本 *arXiv:2012 04572*，2020年。。[9] C. K. Reddy、E. Beyrami、J. Pool、R. Cutler、S. Srinivasan 和 J. Gehrke，《可扩展的噪声语音数据集和在线主观测试框架》，*Interspeech* 会议论文集，第1816–1820页，2019年。[10] 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》，载于*ICASSP* 会议论文集，第2卷，2001年，749 752 2– 卷。[11] S.-W. Fu、C. Yu、T.-A. Hsieh、P. Plantinga、M. Ravanelli、X. Lu 和 Y. Tsao，《Metricgan+：一种改进的语音增强 MetricGAN 版本》，*Interspeech* 会议论文集，2021年。[12] J. M. Martin-Do˜nas、A. M. Gomez、J. A. Gonzalez 和 A. M. Peinado，《基于语音质量感知评估的深度学习损失函数》，*IEEE*信号处理快报，第25卷，第11期，第1680–1684页，2018年。[13] Y. Koizumi、K. Niwa、Y. Hioka、K. Kobayashi 和 Y. Haneda，《基于音质测量、通过强化学习自优化的深度神经网络源增强》，载于*ICASSP* 会议论文集，2017年，第81–85页。[14] S. Kataria、J. Villalba 和 N. Dehak，《融合音频模式识别模型与自监督模型的感知损失语音降噪》，载于*ICASSP* 会议论文集，电气与电子工程师学会，2021年，第7118–7122页。[15] T.-A. Hsieh、C. Yu、S.-W. Fu、X. Lu 和 Y. Tsao，《通过使用 Wasserstein 距离的音素增强感知损失提升语音增强的感知质量》，载于*Interspeech* 会议论文集，2021年，第196–200页。[16] M. Sambur，《说话人识别声学特征的选择》，*IEEE*声学、语音与信号处理汇刊，第23卷，第2期，第176–182页，1975年。

[17] R. Brown，《声学参数对听觉说话人识别相对重要性的实验研究》，*Language and Speech*，第24卷，第4期，第295–310页，1981年。

[18] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, 和 S. Zafeiriou，《使用深度神经网络进行端到端多模态情感识别》，电气与电子工程师学会信号处理专题期刊，第11卷，第8期，第1301–1309页，2017年。

[19] G. d. Krom，“不同类型元音片段中病理性气声和粗糙嗓音质量的一些谱相关特征”，*Journal of Speech, Language, and Hearing Research*，第38卷，第4期，第794–811页，1995年。[20] J. Hillenbrand、R. Cleveland 和 R. Erickson，“气声声质的声学相关特征”，*Journal of speech and hearing research*，第37卷，第769–78页，1994年9月。[21] H. Kasuya、S. Ogawa、Y. Kikuchi 和 S. Ebihara，“病理性嗓音的声学分析及其在喉部病变评估中的应用”，*Speech Communication*，1986年。[22] C.-J. Peng、Y.-J. Chan、Y.-L. Shen、C. Yu、Y. Tsao 和 T.-S. Chi，“基于感知特征的多目标语音增强模型”，载于*Interspeech* 会议论文集，2022年，第211–215页。[23] M. Yang、J. Konan、D. Bick、A. Kumar、S. Watanabe 和 B. Raj，“通过细粒度语音特征改进语音增强”，载于*Interspeech*会议论文集，2022年，第2953–2957页。[24] F. Eyben、M. W¨ollmer 和 B. Schuller，“OpenSmile：慕尼黑多功能快速开源音频特征提取器”，载于第*18*届*ACM*国际多媒体会议论文集，MM ’10系列，意大利Firenze：计算机协会，2010年，第1459–1462页。[25] A. Defossez、G. Synnaeve 和 Y. Adi，“波形域中的实时语音增强”，载于*Interspeech*会议论文集，2020年。[26] 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”，载于*ICASSP*会议论文集，2021年，第6633–6637页。[27] J. Pons、J. Serr`a、S. Pascual、G. Cengarle、D. Arteaga 和 D. Scaini，“用于音乐源分离的上采样层”，*arXiv*预印本 *arXiv:2111.11773*，2021年。[28] C. K. Reddy、V. Gopal、R. Cutler、E. Beyrami、R. Cheng、H. Dubey、S. Matusevych、R. Aichner、A. Aazami、S. Braun、等，“Interspeech 2020深度噪声抑制挑战赛：数据集、主观测试框架和挑战赛结果”，*Interspeech*会议论文集，2020年。[29] J. F. Gemmeke、D. P. Ellis、D. Freedman、A. Jansen、W. Lawrence、R. C. Moore、M. Plakal 和 M. Ritter，“Audio Set：用于音频事件的本体和人工标注数据集”，载于*ICASSP*会议论文集，电气与电子工程师学会，2017年，第776–780页。[30] P. C. Loizou，《语音增强：理论与实践》，第2版。美国：CRC出版社，2013年。[31] C. K. A. Reddy、V. Gopal 和 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*ICASSP*会议论文集，2022年，第886–890页。[32] P. Manocha、B. Xu 和 A. Kumar，“NORESQA：一种使用非匹配参考进行语音质量评估的框架”，载于神经信息处理系统第三十五届会议，2021年。[33] J. Towns、T. Cockerill、M. Dahan、I. Foster、K. Gaither、A. Grimshaw、V. Hazlewood、S. Lathrop、D. Lifka、G. D. Peterson、R. Roskies、J. R. Scott 和 N. Wilkins-Diehr，“XSEDE：加速科学发现”，*Computing in Science & Engineering*，第16卷，第5期，第62–74页，2014年9月。[34] N. A. Nystrom、M. J. Levine、R. Z. Roskies 和 J. R. Scott，“Bridges：面向新兴社区和数据分析的高度灵活高性能计算资源”，载于*2015*年*XSEDE*会议论文集：增强型网络基础设施助力科学进步，2015年，第1–8页。