arXiv:2302.08095v1 [cs.SD] 16 Feb 2023

PAAPLOSS：一种用于语音增强的语音对齐声学参数损失

Muqiao Yang[1], *Joseph Konan*[1],*David Bick*[1],*Yunyang Zeng*[1],*Shuo Han*[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

尽管近年来取得了快速发展，当前的语音增强模型生成的语音在感知质量上往往与真实的干净语音有所不同。我们利用声学语音学领域知识，提出一种能够形式化描述感知质量差异的学习目标。我们识别出一些不可微的时序声学参数，例如谱倾斜、谱通量、微扰等，并开发了一种神经网络估计器，能够准确预测这些参数在整个话语中的时间序列值。我们还针对每种特征建立了音素特定权重，因为已知声学参数在不同音素中表现各异。对于任何能够生成语音的模型，我们都可以将这一准则作为辅助损失，以优化语音输出，使这些特征的取值与干净语音相匹配。实验表明，以标准评估指标衡量，该方法能够改进时域和时频域语音增强流程。我们还分析了声学参数方面依赖音素的改进，展示了该方法带来的额外可解释性。这项分析可以指出当前哪些特征是改进的瓶颈。

索引词— 语音增强，声学参数，语音对齐

1. 引言

语音增强（SE）旨在从主要受噪声影响而劣化的信号中提取干净语音。消除语音中噪声的能力极为实用，因为嘈杂环境常常会影响网络电话和电话通话、助听器以及下游语音处理任务等应用。我们关注更为普遍的单通道语音增强，它不需要多麦克风语音采集。

过去十年间，单通道语音增强从传统信号处理技术转向深度神经网络（DNN），取得了显著进步 [1]–[5]。深度噪声抑制（DNS）挑战进一步推动了单通道语音增强的研究：它提供了一个规模庞大的语料库，其中包含在多种噪声类型和强度下合成的音频 [6], [7]。此外，它还提供了一个用于衡量性能的通用测试集。

单通道语音增强模型通常通过比较增强语音与干净语音的波形或频谱图之间的逐点差异来训练。尽管这一范式行之有效，语音增强模型仍常常生成听起来不自然的语音 [8]。这类经典损失的局限包括无法捕捉音高[9], , 以及对低能量音素的改善相对有限 [10]。此外， [11] 和 [12]指出，信号级别的 ℓ_1 或 ℓ_2 差异与语音质量的相关性并不高。

其他方法尝试解决这些问题，包括优化感知评估指标。然而，这些指标不可微，因此近似方法带来的改进有限 [13], [14], , 需要繁琐的优化过程 [14], [15] ，或者难以借助领域知识提供可解释性 [16]。本文旨在解决这些问题，并尝试通过融入领域知识来实现这一目标，具体做法是利用我们称为声学参数的基础语音特征。

在深度神经网络兴起之前，音高、抖动、微扰、谱倾斜等特征曾被用作浅层模型的输入，例如用于说话人识别和情感识别 [17]。随着深度神经网络在直接处理波形或声谱图方面取得更多成功，这些特征逐渐不再流行。由于其计算过程不可微，也难以直接用于深度神经网络的优化。尽管如此，这些参数仍能提供有关语音信号频率成分、能量/振幅及其他频谱特性的关键信息。以往的感知研究表明，这些特征与嗓音质量之间存在重要关联 [18]–[20]。 [21] 提出了一种可微的估计方法，用于计算这些参数的话语级统计量，并通过辅助损失改进了最先进的语音增强模型；该辅助损失旨在缩小干净语音与增强语音的参数值差异。与此类似，我们使用了扩展版日内瓦最小声学参数集 [17] 中列出的 25 个声学参数。然而，不同于以往通过汇总统计量在整体（话语）层面考察这些声学参数的工作，我们将其时间维度纳入考量 [22]。此外，我们还纳入了声学参数与音素之间的关联；这些关联此前已在声学语音学这一子领域中得到研究。例如，塞音通常先具有较高的振幅，随后振幅会降得很低，因为其发音过程是先在声道中完全闭合，再突然释放压力 [23]。声音中的鼻音性会引入反共振峰，因为鼻腔产生的共振会干扰声道的共振 [24]。不同元音也各有不同的共振峰结构，这取决于声道内不同狭窄位置所产生的共振 [25]。

本文提出了一种音素对齐的声学参数（PAAP）损失，用于改善语音增强系统输出的语音质量。我们通过最小化增强语音与干净语音中音素对齐的声学参数之间的差异来实现这一目标。具体而言，我们采用两步方法。首先，提出一种可微分的时序声学参数估计器，以获取整个话语中每个参数的时间序列。其次，根据各声学参数预测音素逻辑值的能力，为每个声学参数计算可微分的音素特定权重。这样一来，我们便可根据同一时间步所预测的音素，对某一时间步的声学参数赋予不同的侧重点。这两个组件使我们能够端到端地优化原始模型，使其通过音素对齐的声学参数匹配干净语音的结束

到结束。我们的方法相比具有竞争力的语音增强模型取得了更好的效果。更重要的是，我们通过分析声学参数上依赖音素的改进，展示了该方法的可解释性。

2. 相关工作

许多研究尝试引入旨在提升感知质量的损失函数。一些技术包括通过生成对抗网络（GAN）[15], 强化学习 [14], 优化不可微的感知度量，以及对度量进行凸近似 [13]。然而，如 [21], 所示，现有方法未能捕捉上述声学参数，而对保留这些参数进行显式监督则能改善模型输出。

其他方法尝试在提升感知质量时利用语音信息，例如 [16]。然而，这些方法的损失函数并未显式利用音素的领域知识，语音信息也仅通过 wav2vec 嵌入被隐式捕捉。最近， [26] 对语音增强中的语音感知技术进行了研究，但依赖于不可解释的 HuBERT 特征 [27]。这两种技术均在 Valentini 数据集上进行评估，该数据集比我们实验所用的数据集小得多，变化也少得多。此外，如实验部分所示，我们的方法能够通过声学参数和音素实现可解释性。最后， [21] 也使用声学参数来优化感知质量，但没有考虑时间或语音信息。由于这些声学参数在一段话语中以及不同音素之间变化很大，因此对音素和时间依赖关系进行建模有助于提升性能。

3. 方法

我们提出使用语音对齐声学参数损失对语音增强模型进行微调。需要注意的是，该目标函数可应用于任何架构，甚至适用于任何涉及语音输出的任务。本节以语音增强为具体示例进行说明。不过，该方法只需输入波形，且可进行端到端微分，因此 PAAP损失可应用于任何生成波形的模型。

整体学习范式总结于算法 1。我们将在第 3.1 小节介绍时序声学参数估计，在第 3.2 小节介绍语音对齐和加权，并在第 3.3 小节介绍使用所提出的 PAAP损失进行整体微调的过程。

3.1. 时序声学参数估计

首先，我们将预训练语音增强模型作为种子模型 Φ，并输入含噪音频 $\mathbf{X}^N$，以获得增强波形 $\mathbf{X}^E$（第 3 行）。在种子模型的基础上，我们使用预训练估计器网络 Ψ，根据原始波形预测声学参数。这些声学参数包括一组由 25 个低层描述符组成的特征，涵盖韵律、激励、声道和频谱描述符；这些描述符被发现是标准化特征集中最能表征声学特性的部分。

与先前通过汇总统计量在话语级对这些声学参数进行建模的工作不同，我们在建模时纳入了这些声学参数的时间特征。我们将增强后的波形和干净波形输入模型，分别预测时序声学参数矩阵，$\mathbf{D}^E$ 和$\mathbf{D}^C$（第 4–5 行）。估计器网络首先对原始波形执行短时傅里叶变换（STFT），然后将频谱图输入一个

算法 1： 在一次迭代中应用 PAAP损失的整体流程
语音增强范式。

输入： 含噪波形 $\mathbf{X}^N$、干净波形 $\mathbf{X}^C$、种子模型 Φ、预训练的声学低层描述符估计器 Ψ、估计得到的声学-音素权重$\mathbf{w}$。
输出： 计算得到的 PAAP损失 ℓ_{PAAP}
$\mathbf{X}^E \leftarrow \Phi(\mathbf{X}^N)$；　// 当前模型生成的增强波形
$\mathbf{D}^C \leftarrow \Psi(\mathbf{X}^C)$；　// 估计的干净声学参数
$\mathbf{D}^E \leftarrow \Psi(\mathbf{X}^E)$；　// 估计的增强后的声学参数
$\ell_{\text{PAAP}} \leftarrow 0$
$N \leftarrow \text{len}(\mathbf{X}^C)$；　// 帧总数
对于 $i \leftarrow 1$ **到** N **执行**
　$j \leftarrow$ 音素在$\mathbf{X}_i^C$处的索引
　$\ell_{\text{PAAP}} \leftarrow \ell_{\text{PAAP}} + (\mathbf{D}_i^E - \mathbf{D}_i^C)^2 \cdot \mathbf{w}_j$
$\ell_{\text{PAAP}} \leftarrow \frac{1}{N} \cdot \ell_{\text{PAAP}}$
返回 ℓ_{PAAP}

序列神经网络，以获得预测的时序声学参数。我们指出，在 PAAP损失中使用估计得到的干净声学参数而非真实标注，能让其他研究人员更容易使用，因为他们无需像 [21] 中那样借助另一个工具包合成标签。这也让我们的损失更容易应用于任意语音增强网络。

3.2. 语音对齐

PAAP损失的下一个组成部分是一组声学-音素权重$\mathbf{w}$，因为我们希望根据声学参数对预测音素逻辑值的重要性，对其进行不同的加权。这些声学-音素权重通过干净语音估计得到，具体方法是对声学参数及其对应的分段音素逻辑值进行线性回归:

$$\mathbf{w} = ((\mathbf{D}^C)^\top \mathbf{D}^C))^{-1}((\mathbf{D}^C)^\top \mathbf{P}^C) \tag{1}$$

其中，$\mathbf{P}^C$ 表示干净波形的音素逻辑值。每一列$\mathbf{w}_i$ 是从25个声学参数到音素 i的权重向量，并包含一个偏置项。每个权重$\mathbf{w}_{ij}$ 对应于声学参数 i 发生单位变化时，音素 j的对数概率变化程度。

这些权重反映了每个特征包含的有关各音素的信息量，因此我们可以利用它来强调优化干净参数值与增强后的参数值之间的差异，尤其是对当前音素更为显著的差异。

我们使用一个无监督语音对齐器获得$\mathbf{P}^C$，该对齐器的词汇表包含40个音素以及一个用于静音的索引。我们保留静音索引，因为我们预计在话语的非语音区域中，声学参数与音素之间的关系会有所不同，并且希望将其纳入建模。无监督语音对齐器使我们能够灵活地将该方法应用于没有真实转录文本的数据集。

3.3. 使用 PAAP 损失进行微调

微调期间，我们首先使用干净语音预测出的音素逻辑值的最大值索引，预测各个时间帧对应的音素索引 j。然后，我们使用$\mathbf{w}_j$，即音素 j 的声学 - 音素权重。我们计算当前时间步干净语音与增强后的声学参数之间的平方差，然后与 $\mathbf{w}_j$ 执行点积（第 8–10 行）。请注意，这些

指标	含噪	FullSubNet		Demucs	
		基线	PAAP Loss	基线	PAAP Loss
PESQ (↑)	1.58	2.89	**3.00**	2.65	**2.99**
STOI (↑)	91.52	96.41	**96.70**	96.54	**97.12**
DNSMOS (↑)	2.48	3.21	**3.27**	3.31	**3.34**
NORESQA (↑)	2.92	4.08	**4.13**	3.93	**3.99**
WER (↓)	19.0	12.6	**12.1**	15.0	**13.2**

表 1：在合成测试集上使用 PAAP损失与含噪音频及基线模型的评估结果。

权重用于将语音信息融入声学参数差异中，而不是直接预测音素逻辑值。通过这种方式，PAAP损失会计算每个时间步声学参数之间的加权差异。

在我们的实现中，我们使用 hop length 为 160、window length 为 512 的 STFT 来确定帧总数 N。音素逻辑值和声学参数均包含 N 个数值向量。我们对话语中的所有帧重复上述过程，并用帧总数对 PAAP损失取平均。PAAP损失作为辅助损失，与语音增强模型的原始损失共同用于微调网络。我们遵循 [21] 的最优设置，仅保留语音增强模型的权重可训练，冻结其他所有权重。在我们的工作中，这同时适用于声学-音素权重 $\mathbf{w}$ 以及时序声学估计器网络的权重 Ψ。

4. 实验

4.1. 数据

我们使用来自 Interspeech 2020 [6] 深度噪声抑制（DNS）挑战赛的数据和脚本，合成了 50,000 对时长为 30 秒（s）的含噪语音和干净语音，用于训练。我们进一步合成了另外 10,000 对语音，作为验证集。合成过程采用默认设置，其中信噪比（SNR）在 0 到 40 分贝（dB）之间均匀采样。随后，从 DNS 噪声集中选择时长足以覆盖所选 Librivox 干净话语的噪声音频，并将其添加到干净语音 [28]中。

我们的基线模型会在训练前对输入数据进行少量预处理，并在微调期间遵循各模型各自的配置。Demucs 将 30 秒音频切分为 10 秒片段，步长为 2 秒；FullSubNet 则在每次迭代中从 30 秒音频中随机采样一个 3.072 秒片段。

在模型的最终评估中，我们使用无混响的 DNS 2020 合成测试集。该测试集包含 150 条来自格拉茨大学干净语音数据集 [29], 的语句，并混入从 100 多种噪声类别中随机抽取的噪声。测试集的信噪比（SNR）在 0 至 25 dB 之间均匀采样。

4.2. 实验结果

为展示我们提出的方法在提升各种架构方面的稳健性，我们选取了最先进的 Demucs [30] 和 FullSubNet [31]，分别代表时域和时频域模型。这些模型也均已开源，因此我们使用其预训练检查点，以便复现本研究的结果。对于无监督语音对齐器，我们使用一种基于 wav2vec2 的方法 [32]。

在我们的实验中，我们先将 PAAP损失乘以 0.1，再将其加到原始损失中，以微调种子语音增强模型。在

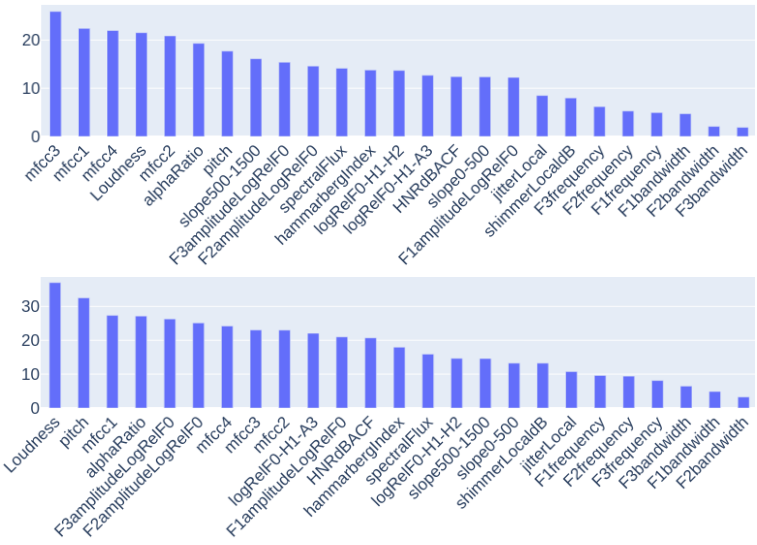

图 1：采用所提出的 PAAP损失后，FullSubNet（上）和 Demucs（下）的声学改善（以 % 计），其中声学改善定义为平均绝对误差的降低，具体定义见第 4.3.1 节。

微调过程中，预训练的时序声学参数估计器是一个 3 层双向长短期记忆网络（LSTM）[33]，包含 512 个隐藏单元。表 1 展示了通过添加 PAAP损失对 FullSubNet 和 Demucs 进行微调后的评估结果。

我们首先考察语音质量感知评估（PESQ）和短时客观可懂度（STOI），因为它们是语音增强的经典评估方法。采用我们的 PAAP损失后，这些指标均有显著改善。请注意，这些都是强大的最先进的模型，因此很难取得改进。尤其是，FullSubNet 和 Demucs 的 PESQ 分别提高了近 **4%** 和 **13%** 。

由于我们的目标是提升感知质量，黄金标准的评估方法是来自人类的平均意见分。它通过对 1–5 分量表上的评分取平均值来计算。开展平均意见分（MOS）研究成本较高，因此我们纳入了两种当前最先进的 MOS 估计方法，即 DNS-MOS [34], 和 NORESQA-MOS（基于非匹配参考的语音质量评估） [35]。我们观察到，对于两种模型，我们的 PAAP损失在这些指标上再次表现出改进。

最后，我们还计算词错误率（WER），以评估我们的增强方法是否减少了影响下游语音处理应用的失真。由于我们没有真实转录文本，因此使用 HuggingFace 上的 WavLM [36] base 模型处理干净语音，并将其转录结果作为参考。随后，我们使用同一识别器处理基线语音和增强语音，并进行比较。我们也观察到词错误率有所改善，这表明我们的方法既能提升人类的感知质量，也能增强语音与各类语音技术的交互能力。

4.3. 分析

4.3.1. 声学改进

图 1 展示了使用 PAAP损失对模型进行微调后，25 个声学参数的改善百分比。声学特征的改善通过增强语音与干净语音的声学参数之间平均绝对误差（MAE）的降低来衡量。形式上，若 $\mathbf{D}^E, \mathbf{D}^C \in \mathbb{R}^{N\times 25}$ 分别表示

估计的增强和干净声学特征，对于每个声学参数j，我们计算

$$\mathrm{MAE}(\mathbf{D}_j^E, \mathbf{D}_j^C) = \frac{1}{N}\sum_{i=1}^{N} |\mathbf{D}_{ij}^E - \mathbf{D}_{ij}^C| \tag{2}$$

然后对所有声学参数取平均，得到 $\mathrm{MAE}(\mathbf{D}^E,\mathbf{D}^C)$。形式上，声学改进可表示为 MAE 的降低：

$$\frac{\mathrm{MAE}(\mathbf{D}^E, \mathbf{D}^C) - \mathrm{MAE}(\mathbf{D}^B, \mathbf{D}^C)}{\mathrm{MAE}(\mathbf{D}^B, \mathbf{D}^C)} \cdot 100\% \tag{3}$$

其中，$\mathbf{D}^B$ 表示基线增强模型的声学参数。对于 FullSubNet，我们可以观察到，PAAP损失在梅尔频率倒谱系数特征和响度方面带来的改进最大。另一方面，对于 Demucs，大多数特征的声学改进幅度与 FullSubNet 相近，唯有响度和采用半音频率刻度的基频提升更为显著，增幅达 30%。在所有声学特征中，两种模型在共振峰频率和共振峰带宽方面的改进相对较小，但我们得出结论：不同类别的语音增强模型在所有声学低层描述符上都取得了稳定的改进。

4.3.2. 音素相关的声学改进

在上一节中，我们考察了各声学参数的整体改进情况。现在，我们进一步细分分析，展示按音素划分的各声学参数改进情况。声学改进的计算方法如下：首先，使用语音对齐器对干净语音进行音素对齐。然后，对于每一帧，我们计算干净语音与增强语音在声学参数上的差值，并将该差值累加到对应对齐音素的总值中。最后，根据帧数对每个音素的差值取平均。

我们将这一分析与引言中提到的声学语音特性联系起来。回想一下，塞音的振幅特征表现出非常鲜明的模式。还要回想一下，元音和鼻音具有特定的共振峰特征。我们绘制了各音素在响度和F1频率方面的声学参数改进图，分别用于表征振幅特征和共振峰特征。

我们在图 2 中绘制了依赖音素的响度和第一共振峰（F1）频率改进情况。每个音素对应一个点，颜色／形状表示不同的音素类别。我们将元音单独分类，并以发音部位作为辅音的分类标准。这包括舌背音、唇音和舌冠音，分别对应以舌背、双唇和舌前部发音的辅音。我们还将 /HH/ 单独列出，因为它是英语中唯一发音部位在喉部的辅音。因此，图中总共使用五种不同的颜色／形状来表示音素类别。

有了这些认识，我们可以看到，按照上述领域知识，采用语音对齐的声学参数损失后，取得了预期的改进。响度方面最大的改进出现在塞音中，例如 /B/、/P/、/K/、/G/、/D/ 和 /DH/，平均改进幅度约为 90%。PAAP损失的目标是学习音素与声学参数随时间变化的关系，并据此微调增强模型。现在我们发现，使用 PAAP损失微调的模型生成的语音，在与特定参数相关的特定音素上，声学参数的改进幅度更大。

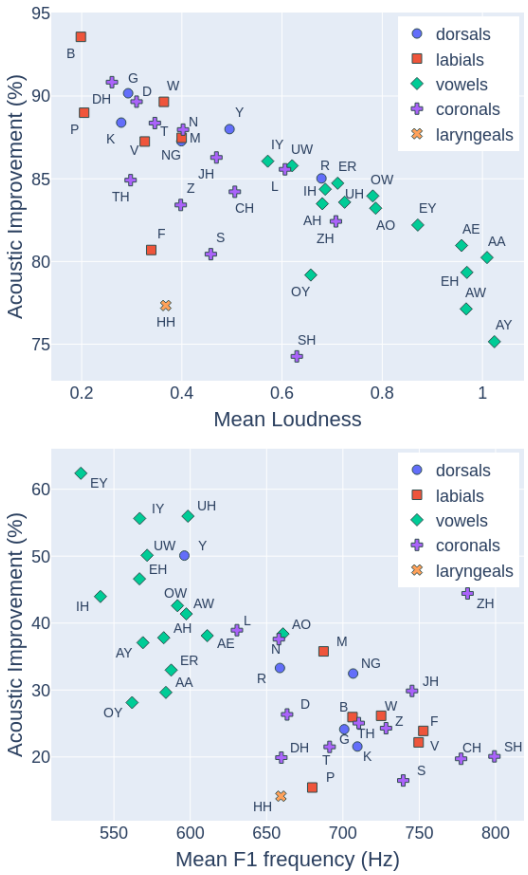

图 2：每个音素的响度/F1频率误差相对于声学参数平均值的降低幅度。

我们还观察到，F1 频率方面的改进呈现出预期的聚类现象。几乎所有改进幅度最大的情况都出现在元音中，因为共振峰结构对元音比对辅音更重要。元音的整体声学改进约为 45%，高于任何一组辅音。鼻音 /N/ 和 /M/ 在引言中也因其共振峰结构而被提及，它们的改进幅度与许多元音相近。其他改进幅度较大的辅音 /L/ 和 /R/ 是流音辅音，众所周知，它们比其他辅音更接近元音。

5. 结论

在本项工作中，我们提出了一种用于语音增强的新型辅助目标，即音素对齐声学参数（PAAP）损失。该损失通过音素类型进行加权，以最小化重要时序声学参数之间的差异。我们在加入 PAAP损失的基础上，对具有竞争力的语音增强模型进行微调；实验表明，包括感知质量评估以及来自具有竞争力的自动语音识别模型的词错误率（WER）在内的所有评估指标均有所提升。我们还对依赖音素的声学改进进行了详细分析，结果表明，声学参数在预期的音素类别中得到了最显著的改善。

6. 致谢

本研究使用了由美国国家科学基金会资助（拨款编号 ACI-1548562）的 Extreme Science and Engineering Discovery Environment（XSEDE）[37], 。具体而言，本研究使用了位于匹兹堡超级计算中心（PSC）的 Bridges系统 [38], ，该系统由美国国家科学基金会资助（拨款编号 ACI-1445606）。

7. 参考文献

[1] H. Zhao、S. Zarar、I. Tashev 和 C.-H. Lee，“用于语音增强的卷积循环神经网络”，载于 ICASSP 会议论文集，IEEE，2018 年，第 2401–2405 页。[2] F. Weninger、F. Eyben 和 B. Schuller，“采用记忆增强循环神经网络的单通道语音分离”，载于 ICASSP 会议论文集，2014 年，第 3709–3713 页。[3] Y. Xu、J. Du、L.-R. Dai 和 C.-H. Lee，“一种基于深度神经网络的语音增强回归方法”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 23 卷，第 1 期，第 7–19 页，2015 年。[4] F. Weninger、H. Erdogan、S. Watanabe、E. Vincent、J. L. Roux、J. R. Hershey 和 B. Schuller，“采用 LSTM 循环神经网络的语音增强及其在噪声鲁棒自动语音识别中的应用”，载于 LVA/ICA，2015 年。[5] P. Plantinga、D. Bagchi 和 E. Fosler-Lussier，“有无平行语音数据条件下用于语音增强的语音反馈”，载于 ICASSP 会议论文集，IEEE，2020 年，第 6679–6683 页。[6] 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、主观测试框架与挑战赛结果”，Interspeech 会议论文集，2020 年。[7] H. Dubey、V. Gopal、R. Cutler、A. Aazami、S. Matusevych、S. Braun、S. E. Eskimez、M. Thakker、T. Yoshioka、H. Gamper 等，“ICASSP 2022 深度噪声抑制挑战赛”，载于 ICASSP 会议论文集，IEEE，2022 年，第 9271–9275 页。[8] C. K. Reddy、E. Beyrami、J. Pool、R. Cutler、S. Srinivasan 和 J. Gehrke，“可扩展噪声语音数据集与在线主观测试框架”，Interspeech 会议论文集，第 1816–1820 页，2019 年。[9] J. Turian 和 M. Henry，“很遗憾你的损失：基于频谱的音频距离无法准确衡量音高”，arXiv 预印本 arXiv:2012 04572，2020 年。[10] P. Plantinga、D. Bagchi 和 E. Fosler-Lussier，“结合识别模型的感知损失，用于单通道增强和鲁棒自动语音识别”，arXiv 预印本 arXiv:2112.06068，2021 年。[11] P. Manocha、A. Finkelstein、R. Zhang、N. J. Bryan、G. J. Mysore 和 Z. Jin，“从最小可觉差学习得到的可微感知音频度量”，Interspeech 会议论文集，2020 年。[12] S.-W. Fu、C. Yu、T.-A. Hsieh、P. Plantinga、M. Ravanelli、X. Lu 和 Y. Tsao，“MetricGAN+：用于语音增强的 MetricGAN 改进版本”，Interspeech 会议论文集，2021 年。[13] J. M. Martin-Do˜nas、A. M. Gomez、J. A. Gonzalez 和 A. M. Peinado，“一种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。[14] Y. Koizumi、K. Niwa、Y. Hioka、K. Kobayashi 和 Y. Haneda，“利用音质测量通过强化学习自我优化的基于深度神经网络的声源增强”，载于 ICASSP 会议论文集，2017 年，第 81–85 页。[15] S.-W. Fu、C.-F. Liao、Y. Tsao 和 S.-D. Lin，“MetricGAN：基于生成对抗网络的语音增强黑箱度量分数优化”，载于国际机器学习会议，PMLR，2019 年，第 2031–2041 页。[16] T.-A. Hsieh、C. Yu、S.-W. Fu、X. Lu 和 Y. Tsao，“利用基于 Wasserstein 距离的语音音素增强感知损失提升语音增强的感知质量”，Interspeech 会议论文集，2021 年。[17] 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，“用于语音研究和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。[18] G. d. Krom，“不同类型元音片段中病理性气息声和粗糙嗓音质量的一些频谱相关特征”，《言语、语言与听力研究杂志》，第 38 卷，第 4 期，第 794–811 页

，1995年。

[19] J. Hillenbrand、R. Cleveland 和 R. Erickson，《气声嗓音质量的声学相关特征》，《言语与听力研究杂志》，第 37 卷，第 769–78 页，1994 年 9 月。[20] 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》，《语音通信》，1986 年。[21] M. Yang、J. Konan、D. Bick、A. Kumar、S. Watanabe 和 B. Raj，《通过细粒度语音特征改善语音增强》，载于 Interspeech 会议论文集，2022 年。[22] Y. Zeng、J. Konan、S. Han、D. Bick、M. Yang、A. Kumar、S. Watanabe 和 B. Raj，《TAPLoss：用于语音增强的时域声学参数损失》，载于 ICASSP 会议论文集，2023 年。[23] A. Alwan、J. Jiang 和 W. Chen，《噪声环境中塞音和擦音发音部位的感知》，《语音通信》，第 53 卷，第 2 期，第 195–209 页，2011 年。[24] W. Styler，《英语和法语元音鼻化的声学特征》，《美国声学学会杂志》，第 142 卷，第 4 期，第 2469–2482 页，2017 年。[25] H. G. Yi、M. K. Leonard 和 E. F. Chang，《颞上回中的语音编码》，《神经元》，第 102 卷，第 6 期，第 1096–1110 页，2019 年。[26] O. Tal、M. Mandel、F. Kreuk 和 Y. Adi，《语音增强中语音学感知技术的系统比较》，载于 Interspeech 会议论文集，2022 年，第 1193–1197 页。[27] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，《HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习》，《IEEE/ACM音频、语音与语言处理汇刊》，第 29 卷，第 3451–3460 页，2021 年。[28] J. F. Gemmeke、D. P. Ellis、D. Freedman、A. Jansen、W. Lawrence、R. C. Moore、M. Plakal 和 M. Ritter，《 AudioSet：用于音频事件的本体与人工标注数据集》，载于 ICASSP 会议论文集，IEEE，2017 年，第 776–780 页。[29] G. Pirker、M. Wohlmayr、S. Petrik 和 F. Pernkopf，《一个基频跟踪语料库及其在多基频跟踪场景中的评估》，载于 Interspeech 会议论文集，2011 年。[30] A. Defossez、G. Synnaeve 和 Y. Adi，《波形域中的实时语音增强》，arXiv 预印本 arXiv:2006 12847，2020 年。[31] 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，IEEE，2021 年，第 6633–6637 页。[32] J. Zhu、C. Zhang 和 D. Jurgens，《无需文本的音素到音频对齐：一种半监督方法》，载于 ICASSP 会议论文集，IEEE，2022 年，第 8167–8171 页。[33] S. Hochreiter 和 J. Schmidhuber，《长短期记忆》，《神经计算》，第 9 卷，第 8 期，第 1735–1780 页，1997 年。[34] C. K. Reddy、V. Gopal 和 R. Cutler，《DNSMOS P.835：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，IEEE，2022 年。[35] P. Manocha 和 A. Kumar，《使用不匹配参考进行平均意见分语音质量评估》，载于 Interspeech 会议论文集，2022 年。[36] S. Chen、C. Wang、Z. Chen、Y. Wu、S. Liu、Z. Chen、J. Li、N. Kanda、T. Yoshioka、X. Xiao 等，《WavLM：面向全栈语音处理的大规模自监督预训练》，《IEEE信号处理选题期刊》，2022 年。[37] 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》，《科学与工程计算》，第 16 卷，第 5 期，第 62–74 页，2014 年 9 月。[38] 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 XSEDE会议论文集：《增强型网络基础设施支持的科学进展》，2015 年，第 1–8 页。