arXiv:2601.04459v1 [eess.AS] 8 Jan 2026

基于流匹配的潜在层级增强：实现鲁棒自动语音识别

Da-Hee Yang 和 Joon-Hyuk Chang *IEEE* 高级会员，

摘要——噪声鲁棒自动语音识别（自动语音识别）通常通过在识别前于波形层面进行语音增强（语音增强）来实现。然而，由于残余失真以及与自动语音识别编码器潜在空间的不匹配，语音层面的增强并不总能带来稳定的识别效果提升。本文提出一种互补策略，称为潜在层级增强，即在自动语音识别推理过程中细化失真的表征。具体而言，我们提出一个即插即用的流匹配细化模块（FM-Refiner），用于处理预训练的基于CTC的自动语音识别编码器输出的潜在表征。FM-Refiner 经过训练，可将不完善的潜在表征——无论其直接来自噪声输入，还是来自增强后但仍不完善的语音——映射至对应的干净表征；它仅在推理阶段应用，无需微调自动语音识别参数。实验表明，无论是直接应用于噪声输入，还是与传统语音增强前端结合，FM-Refiner 都能持续降低词错误率。这些结果表明，基于流匹配的潜在层级细化能够作为现有语音增强方法的轻量而有效的补充，以实现鲁棒自动语音识别。

索引词——**噪声鲁棒自动语音识别、潜在表征级增强、流匹配、细化模块**

I. 引言

ROBUST 噪声条件下的自动语音识别（ASR）长期以来一直是一项挑战 [1]–[10]。一种广泛采用的解决方案是使用语音增强（SE）前端，在识别前处理含噪波形 [3], [5]。这种策略颇具吸引力，因为语音增强模型可以独立于自动语音识别系统开发。然而，波形层面语音质量的提升并不一定能带来稳定的识别效果提升。残余失真、增强伪影，以及增强语音与自动语音识别编码器潜在空间之间的不匹配，往往会限制词错误率（WER）的降低，使噪声鲁棒自动语音识别面临根本性瓶颈 [11]–[13]。

在这封信中，我们提出了一种互补方案，直接在自动语音识别模型的潜在表征级处理失真。我们引入了流匹配细化模块（FM-Refiner），这是一种生成式细化模型，可对预训练的基于 CTC 的自动语音识别编码器 [14]–[16] 的输出表征进行处理。FM-Refiner 学习将不完善的潜在表示（可能来自原始噪声输入，也可能来自仍带有残留伪影的增强语音）转换为具有干净潜在表示特征的分布。与传统的去噪网络不同，它利用流匹配学习失真分布与自动语音识别编码器的干净潜在表示分布之间的显式传输映射。重要的是，这种细化在推理期间以即插即用方式应用，无需重新训练或微调自动语音识别参数。

D. -H. Yang 和 J. -H. Chang（通讯作者）任职于韩国首尔 04763 汉阳大学电子学院（电子邮箱：jchang@hanyang.ac.kr）。

该设计具有三项优势。首先，它直接改进对识别最为关键的表征，从而减少仅依靠波形级语音增强无法解决的不匹配问题。其次，它能够与现有语音增强前端互补，实现先进行波形增强、再进行潜在表征细化的两阶段策略。第三，通过利用流匹配[17] 学习确定性的传输映射，FM-Refiner无需随机采样 [18]–[21]即可实现稳定的细化。

我们在多种噪声条件和语音增强前端 [22]–[24]下评估了所提出的方法。实验结果表明，无论是直接应用于含噪输入，还是与语音增强模型结合使用，FM-Refiner都能持续降低词错误率。这些发现凸显了潜在层级细化作为语音级增强的轻量且有效补充，对于噪声鲁棒自动语音识别具有重要作用，也进一步彰显了将生成式细化直接整合到自动语音识别流程中的新颖性。

本文的主要贡献有三点：(i) 我们针对鲁棒自动语音识别构建了潜在层级增强方法；(ii) 我们设计了即插即用的 FM-Refiner，在无需重新训练自动语音识别模型的情况下，将不完善的潜在表示映射至干净潜在表示；(iii) 我们提供了实验证据，表明潜在表征精炼能够在多种语音增强模型和噪声条件下持续提升识别性能。

II. 建模方法

A. 框架概述

图 1 展示了所提出方法的整体框架。在传统的噪声鲁棒自动语音识别流程中，首先通过 SE前端在波形层面抑制噪声，然后将得到的增强语音输入预训练自动语音识别模型。在本研究中，我们假设自动语音识别模型保持固定，不进行任务特定再训练，因为使用新的噪声条件或领域特定数据进行再训练通常并不现实。尽管语音增强可以改善语音感知质量，但残留伪影以及与自动语音识别编码器特征空间的不匹配往往仍然存在，从而限制识别性能。

为解决这一局限，我们引入了一个额外的精炼阶段，在自动语音识别模型的潜在表征级运行。具体而言，语音增强前端生成增强波形后，信号会被送入预训练ASR编码器。随后，FM-Refiner会对其输出的潜在表征 $\boldsymbol{H}^{L}$进行精炼。FM-Refiner单独进行预训练

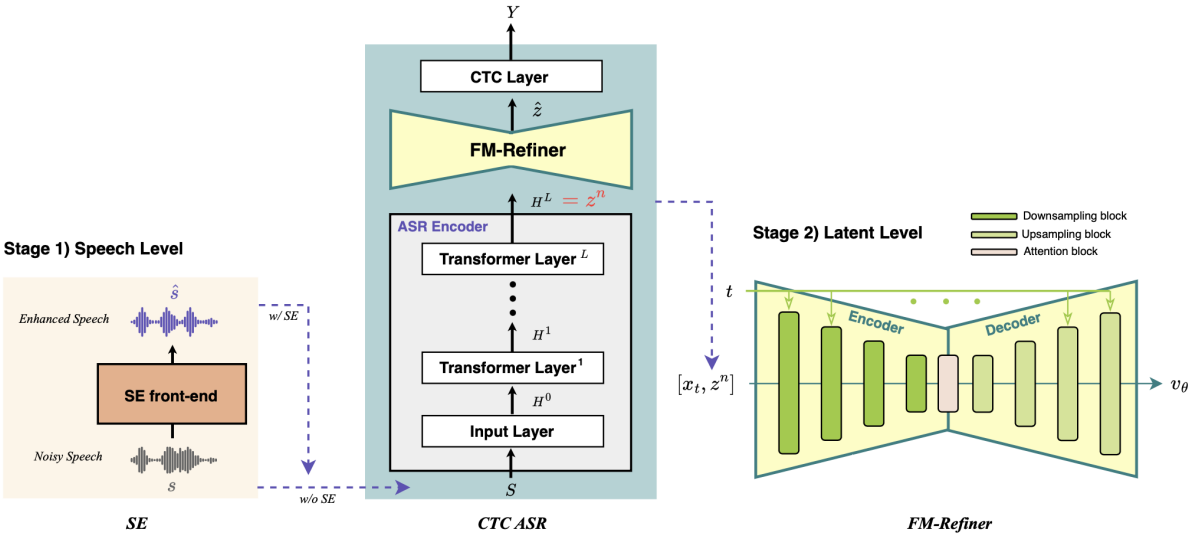

图1. **所提出的两阶段框架概览**。在阶段1中，含噪语音 s 可选择由语音增强前端进行增强，生成增强语音，供预训练的基于 CTC 的自动语音识别编码器使用。在阶段2中，FM-Refiner 对编码器输出潜在表示 z^n 进行细化，并为 CTC层生成改进后的表征$\hat{z}$。所有模块（语音增强、自动语音识别、FM-Refiner）均分别进行预训练；该图展示了解码过程，其中 FM-Refiner 以即插即用方式应用。FM-Refiner采用U-Net架构实现，并通过流匹配进行训练。

用于将噪声或增强后的潜在表示映射为干净潜在表示，并且可以即插即用方式插入，无需对自动语音识别模型进行任何微调。精炼后的潜在表示$\hat{z}$最终被送入CTC层，以生成文本输出。该架构融合了语音级与潜在表征级处理：语音增强可减轻波形域中的噪声，而FM-Refiner则进一步校正自动语音识别模型潜在空间中的残余失配。通过结合这两个阶段，该框架生成的CTC输入与分类器学习到的决策空间更加对齐。

B. 基于*CTC*的自动语音识别架构

本文采用的自动语音识别模型基于连接主义时序分类（CTC）框架 [14]。基于 CTC 的自动语音识别系统旨在根据长度为 T的输入语音序列 $\boldsymbol{S}$预测目标转写序列 $\boldsymbol{Y}$，无需对输入与输出进行显式的帧级对齐。自动语音识别模型的编码器由12个 Transformer 层组成，每层均包含多头自注意力、前馈网络、残差连接和层归一化 [25]。给定从含噪（增强）波形中提取的声学特征序列 $\hat{\boldsymbol{S}}$（$\boldsymbol{S}$），编码器将其映射为更高层次的潜在表征：

$$\begin{aligned} \boldsymbol{H}^0 &= \text{InputLayer}(\boldsymbol{S}\ (\hat{\boldsymbol{S}})), \\ \boldsymbol{H}^l &= \text{TransformerLayer}^l(\boldsymbol{H}^{l-1}), \quad (1 \le l \le L) \end{aligned} \tag{1}$$

其中，$\boldsymbol{H}^l$ 表示第 l个 Transformer 层的输出表征。最终的潜在表征 $\boldsymbol{H}^{12}$（其中$L = 12$）被输入线性投影层，随后经过 Softmax，以计算子词单元上的逐帧后验分布。CTC损失被定义为给定输入序列时目标序列 $\boldsymbol{Y}$ 的负对数似然 $\boldsymbol{S}$，对所有可能的对齐路径 $B(\boldsymbol{Y}, \boldsymbol{S})$ 进行边缘化，其中每条对齐路径$\hat{\boldsymbol{y}}$ 的长度均为 N：

$$P_{CTC}(\boldsymbol{Y}|\boldsymbol{S}) = \sum_{\hat{\boldsymbol{y}} \in B(\boldsymbol{Y},\boldsymbol{S})} \prod_{n=1}^{N} P(\hat{\boldsymbol{y}}_n|\boldsymbol{S}), \tag{2}$$

$$\mathcal{L}_{CTC} = -\ln P_{CTC}(\boldsymbol{Y}|\boldsymbol{S}). \tag{3}$$

这种表述使自动语音识别系统能够从编码器输出中学习，而无需显式的帧到标签的对齐，从而使潜在空间与所提出的 FM-Refiner 直接兼容。

C. FM-Refiner

1) 流匹配预备知识： 流匹配（FM） [17]是一种生成建模框架，通过估计控制中间样本演化的底层向量场，学习先验分布与目标数据分布之间的连续变换。形式化地，设$x_0 \sim q_0(x)$表示来自初始分布的样本，$x_1 \sim q_1(x)$表示来自数据分布的目标样本。FM 定义了一个与时间相关的映射 $f(x_0,\ t) = x_t$，其中$t \in [0, 1]$ 表示 x_0与 x_1之间的连续插值。该映射的动态遵循一个常微分方程（ODE）：

$$\frac{df}{dt} = \frac{dx_t}{dt} = u_t, \tag{4}$$

其中 u_t 是向量场，描述了在时间 t从 x_0 到 x_1 的瞬时输运方向。

在实际应用中，精确的 u_t 未知，通常由神经网络 v_θ ($x_t,\ t$)进行近似。流匹配目标训练v_θ，使其最小化真实向量场与预测向量场之间的差异：

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x_0,x_1,t}\left[\|v_\theta(x_t,t) - u_t\|_2^2\right]. \tag{5}$$

为了使训练目标易于处理，我们引入了连接 x_0 和 x_1 的仿射概率路径。常见的选择是最优传输（OT）路径，它通过噪声控制因子 $\sigma_{\min}$对两个端点进行线性插值：

$$f(x_0, t) = tx_1 + (1 - (1 - \sigma_{\min})t)\, x_0. \tag{6}$$

相应的向量场定义如下

$$u_t = x_1 - (1 - \sigma_{\min})x_0, \tag{7}$$

它与 t 无关，因此能够实现高效且稳定的学习。

训练完成后，FM模型通过使用常微分方程求解器，随时间积分学习到的向量场 v_θ 来生成样本。给定初始样本 $x_0 \sim q_0(x)$，求解器会根据以下公式迭代更新 x_t

$$x_{t+\Delta t} = x_t + \Delta t \cdot v_\theta(x_t, t), \tag{8}$$

其中， $\Delta t = 1/N$， N 为采样步数。

2）用于自动语音识别潜在表示的 *FM-Refiner*：基于这一基础，我们提出 FM-Refiner，这是一种利用流匹配增强自动语音识别潜在表示的精炼模块。其核心思想是通过 FM 网络估计的向量场，将不完美的自动语音识别潜在表示与其干净对应项对齐，从而提升噪声条件下的识别鲁棒性。

具体而言，给定噪声输入话语 $\mathbf{s}$，我们首先使用预训练ASR编码器获取其潜在表征。令$\mathbf{z}^n$ 和$\mathbf{z}^c$ 分别表示噪声和干净的ASR潜在表征。在训练过程中，$\mathbf{z}^n$ 作为起始分布 x_0，而 $\mathbf{z}^c$ 则提供目标 x_1。随后，沿着$\mathbf{z}^n$ 与$\mathbf{z}^c$之间的OT路径构造中间状态 x_t ，并训练FM-Refiner预测相应的向量场。

FM-Refiner采用专为序列潜在特征设计的多分辨率 U-Net 骨干网络[26]。在每个时间步，将扩散后的潜在表征 x_t 与噪声潜在表征$\mathbf{z}^n$ 拼接后输入编码器，如图1所示。编码器通过带有组归一化和SiLU激活函数 [27], 的残差卷积块逐步降低序列分辨率，而解码器则利用上采样层和跳跃连接重建分辨率，以保留细粒度细节。网络还引入了全局注意力瓶颈，以捕获序列中的上下文依赖关系。此外，时间步信息 t通过位置嵌入进行编码，并注入每个残差块，使网络能够沿概率路径调节其特征变换。

训练目标采用条件流匹配损失：

$$\mathcal{L}_{\text{FM-Refiner}} = \mathbb{E}_{\mathbf{z}^n, \mathbf{z}^c, t} \left\| v_\theta(x_t, \mathbf{z}^n, t) - (\mathbf{z}^c - (1 - \sigma_{\min})\mathbf{z}^n) \right\|_2^2. \tag{9}$$

通过对 $\mathbf{z}^n$ 进行条件约束，确保细化过程始终以输入信号为基准，避免过度平滑或丢失说话人特定信息。

在推理过程中，只有带噪的潜在表征 $\mathbf{z}^n$ 可用。FM- Refiner 从 $\mathbf{z}^n$ 初始化，并对学习到的向量场进行迭代积分，以生成精炼后的潜在表征 $\hat{\mathbf{z}}$，随后将其传递至 CTC层。这一精炼过程有效弥合了带噪表征与干净表示之间的差距，从而在各种噪声条件下提升词识别准确率。当应用 SE前端时，带噪的潜在表征 $\mathbf{z}^n$ 会被增强潜在表示取代，后者随后作为精炼过程的初始化。

III. 实验设置

A. 数据集

自动语音识别模型使用 Wall Street Journal (WSJ) 语料库 [28] 进行训练，采用标准数据增强技术，且未在含噪语音上进行微调。为训练语音增强前端，通过将 WSJ 训练集和验证集中的语句与随机选取的 CHiME-4 噪声混合，生成成对的干净语音和含噪语音，信噪比均匀采样自 -5 至 10 分贝。CHiME-4 数据集包含来自街道、咖啡馆、公交车和行人环境的录音，提供了多样且真实的噪声条件 [29]。评估时，将 WSJ 测试语句与信噪比为 $\{-5, -2, 0, 2, 5, 10\}$ 分贝的 CHiME-4 噪声混合。在两种条件下测量词错误率： (i) 直接解码含噪语音；(ii) 含噪语音经语音增强前端处理后，再解码增强语音。

B. 模型配置

- 基于 CTC 的自动语音识别

  我们使用 ESPnet 工具包 [30], [31]构建了一个基于 CTC 的自动语音识别模型。输入特征为 80 维对数梅尔滤波器组系数，窗口大小为 25 ms，帧移为 10 ms。编码器由 12 个 Transformer 层组成，每层配备 4 个注意力头，隐藏维度为 256。逐位置前馈网络的内部维度为 2048。编码器之后接入 CTC层，用于预测输出标签集上的后验分布。模型训练了 100 个 epoch，批量大小为 32。训练使用了 Adam 优化器 [32]，并结合 SpecAugment 进行数据增强，以及使用检查点平均来稳定最终的模型参数。

- FM-Refiner

  我们采用 U-Net 骨干网络实现所提出的 FM-Refiner。该网络由四个下采样层和四个上采样层组成，其中自动语音识别编码器输出作为输入潜在表征。具体而言，自动语音识别编码器生成隐藏维度为 256 的时序特征，并通过流匹配过程逐步细化。模型使用 Adam 优化器进行训练，批量大小为 16，在单个 NVIDIA A100 GPU 上训练 200 个 epoch。推理过程中，FM-Refiner 使用三个采样步数执行。除层数之外的详细架构参数均遵循官方代码[1]中提供的实现。

[1]https://github.com/sp-uhh/sgmse.git

表 I 未经处理的输入与各种语音增强前端的实验结果，有无所提出的 FM-Refiner 均包含在内。FM-Refiner 在所有条件下都能持续降低词错误率（越低越好）。语音增强性能仅供参考，与所提出的方法没有直接关系。

方法		自动语音识别性能							语音增强性能
		词错误率（↓）							DNSMOS(↑)
		-5 分贝	-2 分贝	0 dB	2 dB	5 dB	10 分贝	**Avg.**	[-5, 10] 分贝
未处理		85.6	79.7	73.4	63.9	51.7	39.3	65.60	2.74
	搭配 FM-Refiner	80.9	73.5	65.9	56.4	45.6	35.2	**59.58**	-
语音增强前端									
Conv-TasNet		54.2	44.8	39.7	35.1	31.0	27.7	38.75	3.73
	搭配 FM-Refiner	50.2	40.9	36.9	33.0	28.8	26.8	**36.10**	-
DEMUCS		60.9	49.2	43.7	38.0	31.7	27.2	41.78	3.73
	搭配 FM-Refiner	55.9	46.3	42.3	35.8	30.2	26.7	**39.53**	-
SGMSE+		56.2	45.9	40.8	36.1	31.3	27.0	39.55	3.98
	使用 FM-Refiner	54.2	43.7	39.1	34.4	29.9	26.5	**37.97**	-

C. 语音增强前端模型

为展示所提出 FM-Refiner 的通用性，我们采用了多个语音增强前端模型，其中既包括判别式方法，也包括生成式方法。该设置使我们能够在广泛的增强技术范围内评估潜在表征级细化的有效性，从而确保所提出的方法不局限于特定的语音增强架构或目标。

- Conv-TasNet [22]：一种时域判别式语音增强模型，基于堆叠的时间卷积网络，能够预测输入波形的时变掩码，并应用该掩码生成增强波形。
- DEMUCS [23]：一种基于波形的判别式语音增强模型，采用 U-Net 架构。它通过编码器—解码器结构将带噪波形映射为干净波形。
- SGMSE+ [24]：一种在频谱图域中运行的生成式语音增强模型。它使用噪声条件评分网络对干净语音分布进行建模，从而通过迭代去噪实现样本细化。

IV. 实验结果

表 I 展示了不同信噪比（SNR）条件下各类语音增强模型在使用和不使用所提出的 FM-Refiner 时的词错误率（WER）表现。从这些结果中可以得出几项重要发现。首先，即使在未使用任何语音增强前端的情况下直接应用于含噪语音，FM-Refiner 也能在所有信噪比水平下显著提高自动语音识别准确率，涵盖从 -5 分贝的极端不利条件到 10 分贝的相对干净环境。这表明，预训练 FM-Refiner 可以通过即插即用方式优化自动语音识别潜在表示，而无需重新训练自动语音识别编码器。

其次，与 Conv-TasNet、DEMUCS 和 SGMSE+ 等传统语音增强前端结合时，FM-Refiner 始终能够降低词错误率，证明了其作为互补精炼阶段的有效性。值得注意的是，在所有测试过的条件下均观察到了性能提升，这表明 FM-Refiner 能够适应不同的语音增强架构和噪声严重程度。这种一致性凸显了所提出方法与现有语音增强流程集成时的灵活性。

比较不同的语音增强前端时，可以发现一个有趣的现象。我们使用 DNSMOS [33],这一感知度量评估语音增强质量，该指标反映了语音听起来自然与否。如表 I 所示，生成式模型 SGMSE+取得了最高的 DNSMOS 分数，表明其增强语音听起来更加自然。然而，就下游自动语音识别而言，其词错误率表现不如判别式模型 Conv-TasNet。这种差异凸显了感知层面或信号层面的改进未必能带来更高的自动语音识别准确率，说明仅在波形层面进行增强并不足够。相比之下，FM-Refiner 直接优化自动语音识别潜在空间，弥合含噪与干净表征之间的差距，并带来更符合识别目标的改进。

总体而言，结果表明，FM-Refiner 为优化自动语音识别潜在表示提供了一种系统且有效的解决方案。其在未经处理的输入和语音增强处理后的输入上均取得了稳定的改进，验证了潜在表征级增强对于补充传统语音增强方法的必要性。

V. 结论

我们提出了 FM-Refiner，这是一种潜在表征级细化框架，旨在提升基于 CTC 的自动语音识别在噪声条件下的鲁棒性。FM-Refiner 将流匹配模块应用于自动语音识别编码器输出，逐步将含噪潜在表示转换为更干净的对应表示，作为一种轻量且有效的细化步骤。采用多种语音增强前端进行的实验表明，该方法在不同噪声水平下均能持续提高词识别准确率，凸显了其灵活性与适用性。这些结果表明，将概率精炼融入自动语音识别的潜在表示，为弥合含噪表示与干净表示之间的差距提供了一种新颖途径。

参考文献

[1] H. Xu、Z.-H. Tan、P. Dalsgaard 和 B. Lindberg，“通过非局部均值去噪处理实现鲁棒语音识别”，IEEE信号处理快报，第 15 卷，第 701–704 页，2008 年。[2] J. Li、L. Deng、Y. Gong 和 R. Haeb-Umbach，“噪声鲁棒自动语音识别综述”，IEEE/ACM音频、语音与语言处理汇刊，第 22 卷，777，2014 年 4，745第。页–。[3] F. Weninger、H. Erdogan、S. Watanabe、E. Vincent、J. Le Roux、J. R. Hershey 和 B. Schuller，“采用 LSTM循环神经网络的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国际会议（LCA/ICA）论文集》，第 91–99 页，Springer，2015 年。[4] K. Nathwani、E. Vincent 和 I. Illina，“利用基于 GMM 的不确定性特征进行 DNN不确定性传播，以实现噪声鲁棒自动语音识别”，IEEE信号处理快报，第 25 卷，第 3 期，第 338–342 页，2018 年。[5] Z.-Q. Wang、P. Wang 和 D. Wang，“用于单通道及多通道语音增强和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。[6] C. Fan、J. Yi、J. Tao、Z. Tian、B. Liu 和 Z. Wen，“用于鲁棒端到端语音识别的联合训练框架与门控循环融合”，IEEE/ACM音频、语音与语言处理汇刊，第 29 卷，第 198–209 页，2020 年。[7] Y. Hu、N. Hou、C. Chen 和 E. S. Chng，“用于端到端噪声鲁棒语音识别的交互式特征融合”，载于《IEEE国际声学、语音与信号处理会议（ICASSP）论文集》，第 6292–6296 页，2022 年。[8] D.-H. Yang 和 J.-H. Chang，“将 FiLM条件化与增强特征用于基于 Transformer 的端到端噪声语音识别”，载于《INTERSPEECH论文集》，第 4098–4102 页，2022 年。[9] K. Kinoshita、T. Ochiai、M. Delcroix 和 T. Nakatani，“利用单通道时域增强网络改进噪声鲁棒自动语音识别”，载于《IEEE国际声学、语音与信号处理会议（ICASSP）论文集》，第 7009 7013 页，2020 年–。[10] D.-H. Yang 和 J.-H. Chang，“采用改进语音增强进行联合训练的端到端自动语音识别的基于注意力的潜在特征”，《沙特国王大学学报：计算机与信息科学》，第 35 卷，第 3 期，第 202–210 页，2023 年。[11] P. Wang、K. Tan 和 D. Wang，“通过与失真无关的声学建模弥合单耳语音增强与识别之间的差距”，IEEE/ACM音频、语音与语言处理汇刊，第 28 卷，第 39–48 页，2019 年。[12] M. Fujimoto 和 H. Kawai，“结合含噪特征与增强特征进行单通道噪声语音单遍识别”，载于《INTERSPEECH论文集》，第 486–490 页，2019 年。[13] K. Iwamoto、T. Ochiai、M. Delcroix、R. Ikeshita、H. Sato、S. Araki 和 S. Katagiri，“伪影的影响有多严重？：分析语音增强错误对自动语音识别的影响”，arXiv预印本 arXiv:2201.06685，2022 年。[14] A. Graves 和 N. Jaitly，“迈向采用循环神经网络的端到端语音识别”，载于《国际机器学习会议（ICML）论文集》，第 1764–1772 页，2014 年。[15] D. Amodei、S. Ananthanarayanan、R. Anubhai、J. Bai、E. Battenberg、C. Case、J. Casper、B. Catanzaro、Q. Cheng、G. Chen 等，“Deep Speech 2：英语和普通话的端到端语音识别”，载于《国际机器学习会议（ICML）论文集》，第 173–182 页，2016 年。[16] C.-C. Chiu、T. N. Sainath、Y. Wu、R. Prabhavalkar、P. Nguyen、Z. Chen、A. Kannan、R. J. Weiss、K. Rao、E. Gonina 等，“基于序列到序列模型的先进语音识别”，载于《IEEE国际声学、语音与信号处理会议（ICASSP）论文集》，第 4774–4778 页，2018 年。[17] Y. Lipman、R. T. Q. Chen、H. Ben-Hamu、M. Nickel 和 M. Le，“用于生成建模的流匹配”，载于《国际学习表征会议（ICLR）论文集》，2023 年。[18] Y. Song、J. Sohl-Dickstein、D. P. Kingma、A. Kumar、S. Ermon 和 B. Poole，“通过随机微分方程进行基于分数的生成建模”，载于《国际学习表征会议（ICLR）论文集》，2021 年。[19] D.-H. Yang 和 J.-H. Chang，“Flow-PLC：利用流匹配实现高效丢包隐藏”，IEEE信号处理快报，2025 年。[20] Y. Yuan、X. Liu、H. Liu、M. D. Plumbley 和 W. Wang，“FlowSep：利用整流流匹配进行语言查询声音分离”，载于《IEEE国际声学、语音与信号处理会议（ICASSP）论文集》，第 1–5 页，2025 年。

[21] D.-H. Yang、J. Lee 和 J.-H. Chang，“借助语言模型和流匹配的标记化生成式语音增强”，*IEEE*信号处理快报，2025年。[22] 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8期，第1256–1266页，2019年。[23] 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，第3291–3295页，2020年。[24] J. Richter、S. Welker、J.-M. Lemercier、B. Lay 和 T. Gerkmann，“基于扩散生成模型的语音增强与去混响”，*IEEE/ACM*音频、语音与语言处理汇刊，第31卷，第2351–2364页，2023年。[25] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力即所需”，载于神经信息处理系统进展会议论文集（NeurIPS），第30卷，2017年。[26] Y. Song、J. Sohl-Dickstein、D. P. Kingma、A. Kumar、S. Ermon 和 B. Poole，“通过随机微分方程进行基于分数的生成建模”，载于国际学习表征会议论文集（ICLR），2021年。[27] S. Elfwing、E. Uchibe 和 K. Doya，“用于强化学习中神经网络函数逼近的Sigmoid加权线性单元”，神经网络，第107卷，第3–11页，2018年。[28] L. D. Consortium 等，“CSR-II（WSJ1）完整数据集”，语言数据联盟，费城，第*LDC94S13A*卷，1994年。[29] J. P. Barker、R. Marxer、E. Vincent 和 S. Watanabe，“CHiME挑战赛：日常环境中的鲁棒语音识别”，载于鲁棒语音识别新时代：利用深度学习，第327–344页，Springer，2017年。[30] S. Watanabe 等，“Espnet：端到端语音处理工具包”，载于*INTERSPEECH*会议论文集，2018年。[31] “端到端语音处理工具包”，*https://github.com/espnet/espnet*，引用于2018年3月21日。[32] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，载于国际学习表征会议论文集（ICLR），2015年。[33] C. K. Reddy、V. Gopal 和 R. Cutler，“DNSMOS：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*IEEE*国际声学、语音与信号处理会议论文集（ICASSP），第6493–6497页，2021年。