arXiv:2512.18572v1 [eess.AS] 21 Dec 2025

MeanFlow-TSE：基于均值流的单步生成式目标说话人提取

Riki Shimizu[1], *Xilin Jiang*[1],*Nima Mesgarani*[1]

[1]美国哥伦比亚大学电气工程系rs4613@columbia.edu

摘要

目标说话人提取（TSE）旨在利用参考语句等辅助信息，从多说话人混合语音中分离出目标说话人的声音。尽管近期扩散模型和流匹配模型的进展提升了 TSE 的性能，但这些方法通常需要多步采样，因此在低延迟场景中的实用性受到限制。本文提出 MeanFlow-TSE，这是一种采用均值流目标训练的单步生成式目标说话人提取框架，无需迭代细化即可实现快速、高质量的生成。基于 AD-FlowTSE 范式，我们的方法定义了一个由混合比例（MR）控制的背景与目标声源之间的流。在 Libri2Mix语料库上的实验表明，与现有基于扩散和流匹配的 TSE 模型相比，我们的方法在分离质量和感知指标方面表现更优，且仅需单个推理步骤。这些结果表明，均值流引导的单步生成是实时目标说话人提取的一种有效且高效的替代方案。代码见 https://github.com/rikishimizu/MeanFlow-TSE。

索引词：目标说话人提取、语音增强、MeanFlow、单步生成

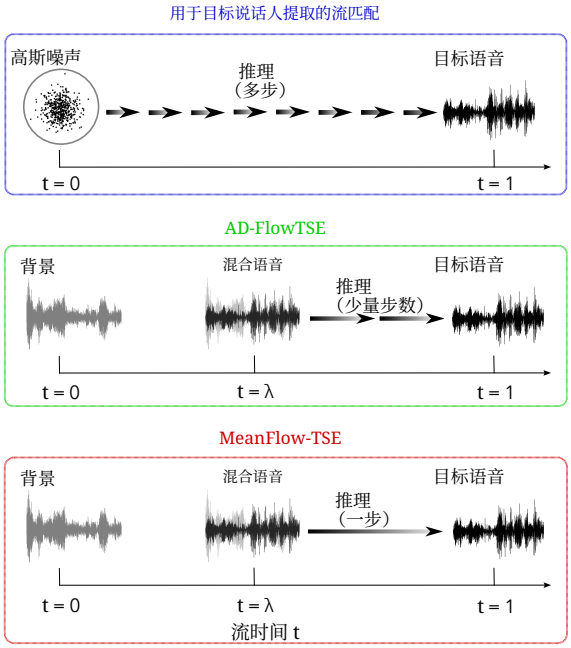

图 1： 生成式目标说话人提取推理范式的比较。上： 标准流匹配通过迭代求解常微分方程，将高斯噪声转换为目标语音（需要多步）。中： *AD-FlowTSE [9]* 定义背景与目标语音之间的流轨迹，并在混合语音的混合比例（$t = \lambda$）处初始化推理，以减少采样步数。下： 本文提出的 *MeanFlow-TSE* 利用均值流目标，实现从混合语音（$t = \lambda$）到目标语音（$t = 1$）的直接高质量单步生成。

1. 引言

目标说话人提取（TSE）旨在从重叠语音和背景噪声的混合语音中分离出目标说话人的声音，并依据说话人特定线索进行引导，例如注册语句 [1, 2], 唇部运动 [3, 4], 或到达方向 [5, 6]。通过有效滤除干扰声音，TSE 对于确保下游应用的鲁棒性至关重要，例如自动语音识别、助听器和电信系统，尤其是在声学环境复杂的现实条件下。

传统的 TSE 方法通常采用判别式模型，直接从输入混合语音中估计目标语音波形。通常，模型经过训练后会预测一个乘性掩码，并将其应用于学习到的音频表示或频谱图，以滤除干扰声音，同时保留目标说话人的声音。通过对 Conv- 等源分离骨干网络进行条件化

TasNet [7] 或 SepFormer [8]利用从辅助线索中提取的说话人嵌入，能够有效地分离出目标说话人。这类判别式训练专注于最小化重构误差，例如尺度不变信噪比（SI-SNR），以恢复所关注的目标说话人。尽管这些方法在信号层面表现出色，但可能引入伪影，并且在未见过的声学条件下的泛化能力可能有限。

近期关于使用扩散模型 [10] 和流匹配 [11] 标志着从判别式掩码估计转向生成式建模的范式变革，并将 TSE 视为条件生成任务。基于扩散的目标说话人提取模型 [12, 13] 展示了基于得分的生成式建模的潜力，并实现了较高的感知质量。然而，由于采用迭代式反向采样过程，这些方法计算成本高、推理速度慢，要获得高质量输出通常至少需要 10 次函数评估次数（NFEs）。减少函数评估次数对于满足助听器和实时通信系统等应用的延迟约束至关重要。流匹配通过学习更直的概率路径，提供了一种颇具前景的替代方案。FlowTSE [14] 将条件流匹配应用于梅尔频谱域的 TSE，而 AD-FlowTSE [9] 则引入了背景与目标声源之间一种能够感知混合比例的传输方式，从而实现自适应少步生成。

值得注意的是，即使只进行单步采样，AD-FlowTSE 也能取得有竞争力的性能，展现了单步生成应用于 TSE 的潜力。然而，它采用标准的整流流匹配目标，该目标最初是为多步采样设计的，因此其单步推理能力仍有提升空间。

均值流目标为直接单步生成提供了一种有理论依据的解决方案。不同于标准流匹配在轨迹上离散的点处建模瞬时速度，MeanFlow [15] 建模的是整个轨迹上的平均速度。这种表述无需进行小步长常微分方程积分，因此能够实现有效的单步生成，并在图像生成 [15], 语音增强 [16, 17], 和文本转语音合成 [18]方面取得最先进的性能。随后，α-Flow 方法 [19], 进一步改进了这一方法，将均值流目标分解为轨迹流匹配项和轨迹一致性项。通过发现这两项之间存在较强的负相关，作者提出了一种课程学习策略，逐步将目标从轨迹流匹配插值至均值流恒等式。值得注意的是，α-Flow 在训练过程中无需计算雅可比向量积，从而显著降低了计算开销，同时提升了训练稳定性。尽管均值流目标在其他领域取得了这些成果，但尚未应用于目标说话人提取。

在本研究中，我们提出了 MeanFlow-TSE，这是均值流目标首次应用于目标说话人提取。基于 AD-FlowTSE 范式，我们的方法定义了背景与目标声源之间的流，该流由生成混合语音时使用的混合比例控制，从而实现混合比例感知初始化与训练。通过使用均值流目标进行训练，我们实现了快速、高质量的单步生成。在 Libri2Mix 上进行的实验表明，与现有基于扩散和流匹配的目标说话人提取模型相比，MeanFlow-TSE 在分离质量（尺度不变信号失真比）和感知指标（感知语音质量评估、扩展短时客观可懂度）方面表现更佳，同时保持了适用于实时应用的计算效率，证明了均值流引导的单步生成是实际目标说话人提取系统的一种有效替代方案。

2. 背景

2.1. 目标说话人提取

目标说话人提取（TSE）的目标是从单通道混合信号$\mathbf{s} \in \mathbb{R}^L$ 中恢复特定目标说话人的离散语音波形$\mathbf{y} \in \mathbb{R}^L$。混合语音被建模为目标语音与干扰成分$\mathbf{b} \in \mathbb{R}^L$（包括其他说话人和背景噪声）之和：

$$\mathbf{y} = \mathbf{s} + \mathbf{b} \tag{1}$$

目标说话人提取系统利用辅助参考线索$\mathbf{e}$（通常为预先录制的注册语句）来识别目标说话人。该问题被表述为估计$\hat{\mathbf{s}} = f_\theta(\mathbf{y},\mathbf{e})$。传统方法通常估计掩码，而生成式方法则对频谱域中的轨迹进行建模。令 STFT(·) 表示短时傅里叶变换。我们将混合语音、目标语音和背景的频谱表示分别定义为$\mathbf{Y} = \text{STFT}(\mathbf{y})$、$\mathbf{S} = \text{STFT}(\mathbf{s})$ 和$\mathbf{B} = \text{STFT}(\mathbf{b})$。

2.2. 用于目标说话人提取的流匹配

流匹配 [11]为训练连续归一化流（CNFs）提供了一种无需模拟的框架。它将生成过程建模为一个确定性的常微分方程（ODE），用于控制潜在空间中的状态$\mathbf{z}_t$（此处为复数频谱域）：

$$d\mathbf{z}_t = v_\theta(\mathbf{z}_t, t, \mathbf{e})dt, \quad t \in [0, 1] \tag{2}$$

其中，v_θ 是用于预测速度场的神经网络。其目标是将源分布 $p_0(\mathbf{z})$ 转换为数据分布 $p_1(\mathbf{z})$。最优传输（OT）定义了一条直线概率路径：

$$\mathbf{z}_t = (1 - t)\mathbf{z}_0 + t\mathbf{z}_1 \tag{3}$$

标准流匹配目标函数最小化网络输出与路径导数之间的回归误差：

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1,\mathbf{e}}\left[\|v_\theta(\mathbf{z}_t, t, \mathbf{e}) - (\mathbf{z}_1 - \mathbf{z}_0)\|^2\right] \tag{4}$$

2.3. 自适应确定性流匹配

AD-FlowTSE [9] 通过定义背景表征$\mathbf{B}$ 与目标表征$\mathbf{S}$之间的流来调整这一框架。尽管物理混合是加性的（$\mathbf{y} = \mathbf{s} + \mathbf{b}$），但考虑到提取任务的尺度不变性，该方法假设混合语音是由标量混合比例 $\lambda \in [0, 1]$ 控制的凸组合，从而使其与最优传输轨迹保持一致：

$$\mathbf{y} = \lambda\mathbf{s} + (1 - \lambda)\mathbf{b} \tag{5}$$

在频谱流域中，路径定义为使得$\mathbf{z}_0 = \mathbf{B}$和$\mathbf{z}_1 = \mathbf{S}$：

$$\mathbf{z}_t = t\mathbf{S} + (1 - t)\mathbf{B} \tag{6}$$

其中 $t \in [0, 1]$。这使得能够感知混合比例地进行初始化。由于混合语谱图$\mathbf{Y}$ 位于该轨迹上的 $t = \lambda$处，因此可使用估计的混合比例 λ，将推理初始化为$\mathbf{z}_{\hat{\lambda}} \approx \hat{\mathbf{Y}}$，从而将积分范围缩小至 $[\hat{\lambda}, 1]$。

2.4. MeanFlow 与 α-Flow 目标

通过标准流匹配或扩散有效提取目标说话人，需要进行多次函数评估（函数评估次数）和常微分方程步进。为缓解这一计算瓶颈，MeanFlow [15]通过建模在时间步 t 与 r 之间跳转所需的平均速度（其中 $r > t$），实现单步生成。状态$\mathbf{z}_r$ 可通过以下方式恢复：

$$\mathbf{z}_r = \mathbf{z}_t + (r - t)v_{\text{avg}}(\mathbf{z}_t, t, r) \tag{7}$$

为提高训练效率，α-Flow 框架 [19]定义了一种混合目标速度$v^{\alpha}_{t,\,r}$ 该速度在真实轨迹与自一致性目标之间进行插值：设真实方向为$\mathbf{u} = \mathbf{S} - \mathbf{B}$。目标速度$v^{\alpha}_{t,\,r}$ 定义如下：

$$v^{\alpha}_{t,r} = \alpha\mathbf{u} + (1 - \alpha)v_\theta(\mathbf{z}_\tau, \tau, r) \tag{8}$$

其中 $\alpha \in (0, 1]$ 是一个标量超参数，$\tau = \alpha r + (1 - \alpha)t$ 是一个中间时刻。该目标函数通过最小化与此插值目标之间的距离来进行优化：

$$\mathcal{L}_\alpha(\theta) = \mathbb{E}_{t,r,\mathbf{B},\mathbf{S}}\left[\left\|v_\theta(\mathbf{z}_t, t, r) - v^{\alpha}_{t,r}\right\|^2\right] \tag{9}$$

3. MeanFlow-TSE

3.1. 提出的方法与推理

我们通过使用注册语音 $\mathbf{e}$ 对速度估计器进行条件化，将 MeanFlow 应用于目标说话人提取。网络 $v_\theta(\mathbf{z}_t, t, r, \mathbf{e})$ 预测从时刻 t 到目标时刻 r 的传输。

在推理时，我们旨在通过单步恢复干净的目标$\mathbf{S}$（其中 $r = 1$）。使用估计的混合比例 ˆratio λ，我们将流初始化于混合语音$\mathbf{z}_{start} = \mathbf{Y}$ 处（对应于 $t_{start} =$ ˆλ）。目标频谱图 ˆ$\mathbf{S}$ 的估计如下：

$$\hat{\mathbf{S}} = \mathbf{Y} + (1 - \hat{\lambda})v_\theta(\mathbf{Y}, \hat{\lambda}, 1, \mathbf{e}) \tag{10}$$

这实际上跳过了噪声占主导的轨迹，直接从混合语音跳转到目标。

3.2. 自适应权重损失

遵循 α-流框架 [19],，我们采用一个由 和 参数化的自适应权重 w α c Δ $=$。令 $(\mathbf{z},,,\mathbf{e})$ v_θ $_t$ t r $-$ v_t^α $_r$,为误差向量。权重定义如下：

$$w = \frac{\alpha}{||\Delta||_2^2 + c} \tag{11}$$

其中 $c = 10^{-3}$。训练目标使用停止梯度算子 sg(·) 应用此权重：

$$\mathcal{L}_{\text{Adaptive}}(\theta) = \mathbb{E}_{t,r,\ldots}\left[\text{sg}(w) \cdot ||\Delta||_2^2\right] \tag{12}$$

3.3. AlphaCurriculum调度

我们采用一种课程学习策略，逐步将训练目标从轨迹流匹配（$\alpha = 1$）过渡到均值流恒等式（$\alpha \to 0$）。超参数 α 同时控制公式 (8) 中目标速度的插值，以及公式 (11) 中的自适应权重。我们在训练迭代 k 期间采用基于 Sigmoid 的调度：

$$\alpha(k) = \text{clip}\left(1 - \sigma\left(\gamma \cdot \left(\frac{k - k_s}{k_e - k_s} - 0.5\right)\right), \alpha_{\text{min}}, 1.0\right) \tag{13}$$

其中 k_s 和 k_e 表示过渡的起点和终点，γ 控制陡峭程度，$\sigma(\cdot)$ 是 Sigmoid 函数。由此形成三个训练阶段：(1) 轨迹流匹配预训练（$k < k_s, \alpha = 1$），(2) 课程过渡（$k_s \leq k < k_e, \alpha$ 从 1 递减至 α_{min}），以及 (3) 均值流微调（$k \geq k_e, \alpha \approx \alpha_{\text{min}}$）。我们在第 $\gamma = 25.0$ 轮训练时设置 $k_s = 0$, k_e 2000 ，并设置 $\alpha_{\text{min}} = 0.005$。

3.4. 混合比例预测器

由于推理期间未知 λ，我们使用辅助网络 g_ϕ 对其进行估计。该预测器使用 ECAPA-TDNN [20] 作为特征提取器 $w(\cdot)$，并采用多层感知机 $h(\cdot)$，输出：

$$\hat{\lambda} = g_\phi(\mathbf{y}, \mathbf{e}) = \sigma\left(h\left([w(\mathbf{y}); w(\mathbf{e})]\right)\right) \tag{14}$$

ˆ 该模型通过均方误差损失单独训练：

$$\mathcal{L}_{\text{MR}}\ \phi\ = \mathbb{E}[\ \lambda - ()\ (\ )^2]。$$

λ

4. 实验

4.1. LibriMix 数据集

与 AD-FlowTSE 类似，我们采用了 [23, 24] 中使用 Libri2Mix 数据集所描述的设置。我们将 train-360 和 train-100 子集合并用于训练，并分别保留开发集和测试集用于验证和评估。所有音频均以 16 千赫采样。我们的输入流程使用 6 秒片段，由 3 秒注册语音和 3 秒混合语音组成。为进行频谱特征提取，我们采用了帧移为 128、窗长为 510 的短时傅里叶变换。

4.2. 训练配置

4.2.1. 模型与优化

骨干架构是一种 U-Net 风格的扩散变换器（UDiT） [25]，由 16 个变换器层 [26] 组成，每层包含 16 个注意力头，隐藏维度为 768。输入和输出维度均设置为 512，位置长度为 500，且不应用位置编码。我们使用 AdamW [27] 对模型进行优化，权重衰减为 0.01。学习率采用余弦退火调度（$T_{\text{max}} = 50$），预热 5 个训练轮次，从初始值 1×10^{-4} 衰减至最小值 1×10^{-5}。

4.2.2. 训练设置

模型训练了 2,000 个训练轮次，批量大小为 32，并在 8 个 NVIDIA L40 GPU 上采用分布式数据并行。我们采用 16 位混合精度训练，以降低内存消耗并提高计算效率。为防止梯度爆炸，采用了梯度裁剪，阈值设为 0.5。在推理过程中，除图 2 中的函数评估次数分析外，验证和测试均使用欧拉求解器执行单步计算。最终模型根据在验证集上取得最高尺度不变信号失真比的检查点进行选取。

4.2.3. MeanFlow 配置

时间步 r 和 t 从由 $\mu = -0.4$ 和 $\sigma = 1.0$参数化的对数正态分布中采样。我们采用 0.5 的流比例：具体而言，在 50% 的训练迭代中，条件 $t = r$ 会被强制执行，从而将目标有效地简化为标准整流流匹配。为平衡各损失分量，自适应权重系数 c 被设为 1×10^{-3}。

4.3. 基线模型

我们将 MeanFlow-TSE 与一组多样化的生成式TSE基线进行比较，包括 DiffSep+SV、 [13],DDTSE、 [13], DiffTSE、 [12], FlowTSE、 [14], SR-SSL、 [21], 以及 SoloSpeech [22]。对于这些模型，我们采用 Navon等人 [14] 报告的综合基准结果，涵盖干净和噪声配置。此外，我们将 AD-FlowTSE [9] 纳入主要比较对象。由于 AD-FlowTSE 与我们的方法采用相同的骨干架构和超参数，因此可作为直接对照，用于单独评估 MeanFlow 范式的具体效果。

4.4. 评估指标

为全面评估模型性能，我们采用涵盖语音质量、可懂度和说话人相似度的 6 项指标。我们使用尺度不变信号失真比 (SI-SDR) [28]，语音质量感知评估 (PESQ) [29]，和扩展短时客观可懂度 (ESTOI) [30]来衡量信号保真度和感知质量。我们还报告 DNSMOS [31] 和 OVRL分数，以估计平均意见分。最后，为评估目标身份的保持程度，我们基于目标嵌入与估计嵌入之间的余弦相似度计算说话人相似度 (SIM) [32] 。

表1： 生成式目标说话人提取模型在 *Libri2Mix* 噪声集和干净集上的性能比较。

方法	Type	Libri2Mix 噪声集						Libri2Mix 干净集					
		感知语音质量评估	扩展短时客观可懂度	尺度不变信号失真比	OVRL	DNSMOS	SIM	感知语音质量评估	扩展短时客观可懂度	尺度不变信号失真比	OVRL	DNSMOS	SIM
混合语音	–	1.08	0.40	-1.93	1.63	2.71	0.46	1.15	0.54	0.00	2.65	3.41	0.54
DiffSep+SV [13]	G	1.32	0.60	–	2.78	3.63	0.62	1.85	0.79	–	3.14	**3.83**	0.83
DDTSE [13]	G	1.60	0.71	–	3.28	3.74	0.71	1.79	0.78	–	**3.30**	3.79	0.73
DiffTSE [12]	G	–	–	–	–	–	–	3.08	0.80	11.28	–	–	–
FlowTSE [14]	G	1.86	0.75	–	**3.30**	**3.82**	0.83	2.58	0.84	–	3.27	3.79	0.90
SR-SSL [21]	G	–	–	–	–	–	–	2.99	–	16.00	–	–	–
SoloSpeech [22]	G	1.89	0.78	11.12	–	3.76	–	–	–	–	–	–	–
AD-FlowTSE [9]	G	2.15	0.81	12.69	3.11	3.48	**0.87**	2.89	0.90	17.49	3.15	3.59	**0.95**
MeanFlow-TSE	G	**2.21**	**0.82**	**12.85**	3.17	3.55	0.73	**3.26**	**0.93**	**18.80**	3.21	3.69	0.92

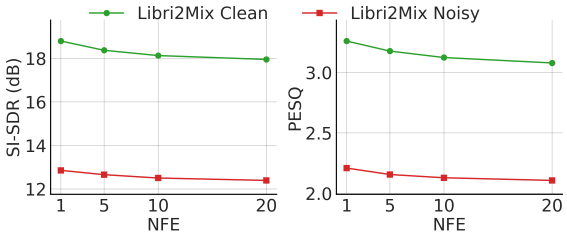

图 2： 不同函数评估次数下性能指标（尺度不变信号失真比和感知语音质量评估）的比较。

5. 结果

5.1. 总体性能

在表1中，我们报告了 MeanFlow-TSE 与其他生成式基线目标说话人提取模型在 Libri2Mix 噪声子集和干净子集上的性能对比。

总体而言，在语音提取和基于参考的感知质量方面，我们的方法始终优于其他生成式基线。在 Libri2Mix 干净集测试集上，MeanFlow-TSE 的感知语音质量评估得分为 3.26，尺度不变信号失真比为 18.80 dB，分别比性能强劲的 AD-FlowTSE 基线高 0.37 和 1.31 dB。这一显著提升凸显了均值流目标在学习更准确的单步传输映射方面的有效性，相较之下，AD-FlowTSE 使用的是标准整流流匹配。在 Libri2Mix 噪声集上也观察到类似趋势：我们的方法在侵入式指标上取得了最先进的结果，感知语音质量评估得分为 2.21，尺度不变信号失真比为 12.85 dB。

对于 OVRL 和 DNSMOS 等非侵入式感知指标，它们无需真实参考即可评估语音质量。MeanFlow-TSE 的表现具有竞争力，但略逊于 FlowTSE 和 DDTSE。这与先前在 [9] 中报告的侵入式指标和非侵入式“自然度”评分之间的权衡一致。至关重要的是，MeanFlow-TSE 在 OVRL 和 DNSMOS 两项指标上均优于 AD-FlowTSE，在保持显著更高语音提取质量的同时，有效缩小了与多步生成模型之间的差距。

5.2. 函数评估次数分析

为验证我们单步生成策略的有效性，我们分析了函数评估次数（NFE）对语音提取性能的影响。图 2 显示，MeanFlow-TSE 在 NFE = 1 时达到了最佳性能，无论是 SI-SDR 还是感知语音质量评估（PESQ）均如此。这一行为

表2： *Libri2Mix* 上实时因子（*RTF*）和 *GPU* 峰值内存（显存）的比较。基线方法的函数评估次数取自各自原始研究 *[22, 9]* 中报告的最优配置。

方法	NFE	参数量（M）	RTF	内存（MB）
SoloSpeech [22]	50	589	0.75	3738
AD-FlowTSE [9]	1	358	0.017	1531
	5		0.047	
MeanFlow-TSE	1	359	0.018	1536

从经验上验证了 MeanFlow-TSE 背后的理论动机：训练目标成功校正了流轨迹，使模型能够在单步中将含噪混合语音直接映射为干净的目标语音。因此，额外的欧拉积分步骤不仅没有必要，反而会略微降低性能，这可能是由于离散化误差的累积所致。

5.3. 计算开销分析

表2对比了 MeanFlow-TSE 与其最具竞争力的基线模型 SoloSpeech 和 AD-FlowTSE 的计算效率。所有测量均在单张 NVIDIA L40 GPU 上进行，批量大小为 1，音频片段时长为 3 秒。

作为一种基于扩散的模型，SoloSpeech 需要 50 次函数评估次数（NFE），因此其实时因子（RTF）和内存占用都显著高于基于流的对应模型。相比之下，MeanFlow-TSE 和 AD-FlowTSE 采用了几乎相同的骨干架构；尽管 MeanFlow-TSE 额外包含一个小型时间步嵌入器，但由此增加的参数量和峰值内存使用量都十分有限。更重要的是，这些结果表明，表1中报告的性能提升是在相对于 AD-FlowTSE 几乎没有额外计算开销的情况下实现的。凭借单步推理和极小的资源占用，MeanFlow-TSE 特别适合在边缘设备上进行实时部署。

6. 结论

我们提出了 MeanFlow-TSE，这是一种用于单步生成式目标说话人提取的新颖框架。我们的方法在 Libri2Mix 基准测试中取得了最先进的 SI-SDR、感知语音质量评估（PESQ）和扩展短时客观可懂度（ESTOI）表现。未来的工作将把这一框架扩展到混响环境和多通道场景。

7. 参考文献

[1] Q. Wang、H. Muckenhirn、K. Wilson、P. Sridhar、Z. Wu、J. R. Hershey、R. A. Saurous、R. J. Weiss、Y. Jia 和 I. L. Moreno，《VoiceFilter：通过说话人条件频谱掩蔽实现目标语音分离》，2019 年，第 2728–2732 页。 [在线]。地址：https://www.isca-archive. org/interspeech 2019/wang19h interspeech.html[2] M. Ge、C. Xu、L. Wang、E. S. Chng、J. Dang 和 H. Li，《SpEx+：完整的时域说话人提取网络》，载于 Interspeech 2020。国际语音通信协会，2020 年 10 月，第 1406–1410 页。 [在线]。地址：https://www.isca-archive. org/interspeech 2020/ge20 interspeech.html[3] A. Ephrat、I. Mosseri、O. Lang、T. Dekel、K. Wilson、A. Hassidim、W. T. Freeman 和 M. Rubinstein，《鸡尾酒会中聆听：用于语音分离的说话人无关的视听模型》，ACM Transactions on Graphics，第 37 卷，第 4 期，第 112:1–112:11 页，2018 年 7 月。 [在线]。地址：https://dl.acm.org/doi/10.1145/3197517.3201357[4] Z. Pan、R. Tao、C. Xu 和 H. Li，《Muse：利用视觉线索进行多模态目标说话人提取》，载于 ICASSP 2021 - 2021 IEEE国际声学、语音与信号处理会议（ICASSP），2021 年 6 月，第 6678–6682 页，ISSN：2379-190X。 [在线]。地址：https://ieeexplore.ieee.org/abstract/document/9414023[5] D.-J. Alcala Padilla、N. L. Westhausen、S. Vivekanan-than 和 B. T. Meyer，《面向助听器的位置感知目标说话人提取》，载于 Interspeech 2025。国际语音通信协会，2025 年 8 月，第 2975–2979 页。 [在线]。地址：https://www.isca-archive.org/interspeech 2025/ alcalapadilla25 interspeech.html[6] S. He、J. liu、H. Li、Y. Yang、F. Chen 和 X. Zhang，《3S-TSE：面向实时和低资源应用的高效三阶段目标说话人提取》，2024 年 1 月，arXiv:2312. 10979 [cs]。 [在线]。地址：http: //arxiv.org/abs/2312.10979[7] Y. Luo 和 N. Mesgarani，《Conv-TasNet：超越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 8 月。 [在线]。地址：https://ieeexplore.ieee.org/ abstract/document/8707065[8] C. Subakan、M. Ravanelli、S. Cornell、M. Bronzi 和 J. Zhong，《语音分离中的注意力即一切》，载于 ICASSP 2021 - 2021 IEEE国际声学、语音与信号处理会议（ICASSP），2021 年 6 月，第 21–25 页，ISSN：2379-190X。 [在线]。地址：https://ieeexplore.ieee.org/abstract/document/9413901[9] T.-A. Hsieh 和 M. Kim，《用于目标说话人提取的自适应确定性流匹配》，2025 年 10 月，arXiv:2510.16995 [电气工程与系统科学]。 [在线]。地址：http://arxiv. org/abs/2510.16995[10] J. Ho、A. Jain 和 P. Abbeel，《去噪扩散概率模型》，载于《神经信息处理系统进展》，第 33 卷。Curran Associates, Inc.，2020 年，第 6840–6851 页。 [在线]。地址：https://proceedings.neurips.cc/paper/2020/hash/ 4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html[11] Y. Lipman、R. T. Q. Chen、H. Ben-Hamu、M. Nickel 和 M. Le，《生成式建模的流匹配》，载于国际学习表征会议，2023 年。 [在线]。地址：https://openreview.net/forum?id= PqvMRDCJT9t[12] N. Kamo、M. Delcroix 和 T. Nakatani，《基于条件扩散模型的目标语音提取》，载于 INTERSPEECH 2023。国际语音通信协会，2023 年 8 月，第 176–180 页。 [在线]。地址：https://www.isca-archive.org/interspeech 2023/ kamo23 interspeech.html[13] L. Zhang、Y. Qian、L. Yu、H. Wang、H. Yang、S. Liu、L. Zhou 和 Y. Qian，《DDTSE：用于目标语音提取的判别式扩散模型》，载于 2024 IEEE口语语言技术研讨会（SLT），2024 年 12 月，第 294–301 页。 [在线]。地址：https://ieeexplore.ieee.org/ abstract/document/10832163

[14] A. Navon, A. Shamsian, Y. Segal-Feldman, N. Glazer, G. Hetz, 和 J. Keshet，《FlowTSE：通过流匹配进行目标说话人提取》，载于*Interspeech 2025*。国际语音通信协会，2025年8月，第2965–2969页。 [在线]。获取地址：https://www.isca-archive.org/interspeech 2025/ navon25 interspeech.html[15] Z. Geng, M. Deng, X. Bai, J. Z. Kolter, 和 K. He，《用于单步生成式建模的平均流》，2025年5月，arXiv:2505.13447 [cs]。 [在线]。获取地址：http: // arxiv.org/abs/2505.13447[16] Y. Zhu, B. Kang, Z. Wang, X. Li, Z. Zhang, W. Li, L. Xiao, W. Xue, 和 L. Xie，《MeanFlowSE：通过 MeanFlow 实现单步生成式语音增强》，2025年9月，arXiv:2509.23299 [cs]。 [在线]。获取地址：http://arxiv.org/abs/2509.23299[17] J. Wang, H. Wang, W. Wang, L. Yang, C. Li, W. Zhang, L. Tan, 和 Y. Qian，《MeanSE：利用平均流实现高效生成式语音增强》，2025年9月，arXiv:2509.21214 [电气工程与系统科学]。 [在线]。获取地址：http://arxiv. org/abs/2509.21214[18] W. Wang, R. Cao, Y. Guo, Z. Chen, K. Chen, 和 Y. Huo，《IntMeanFlow：通过积分速度蒸馏实现少步语音生成》，2025年10月，arXiv:2510. 07979 [cs]。[在线]。获取地址：http://arxiv.org/abs/2510.07979[19] H. Zhang, A. Siarohin, W. Menapace, M. Vasilkovsky, S. Tulyakov, Q. Qu, 和 I. Skorokhodov，《AlphaFlow：理解并改进 MeanFlow 模型》，2025年10月，arXiv:2510.20771 [cs]。 [在线]。获取地址：http: //arxiv.org/abs/2510.20771[20] B. Desplanques, J. Thienpondt, 和 K. Demuynck，《ECAPA-TDNN：基于 TDNN 的说话人验证中的强调式通道注意力、传播与聚合》，载于*Inter- speech 2020*。国际语音通信协会，2020年10月，第3830–3834页。 [在线]。获取地址：https://www.isca-archive.org/interspeech 2020/ desplanques20 interspeech.html[21] P.-J. Ku, A. H. Liu, R. Korostik, S.-F. Huang, S.-W. Fu, 和 A. Juki´c，《生成式语音基础模型预训练：实现高质量语音提取与修复》，载于*ICASSP 2025 - 2025 IEEE*国际声学、语音与信号处理会议 *(ICASSP)*，2025年4月，第1–5页，ISSN：2379-190X。 [在线]。获取地址：https://ieeexplore.ieee.org/document/10888830

[22] H. Peng, T. Thebaud, Y. Wang, J. Villalba, N. Dehak, M. Hasegawa-Johnson, A. Velazquez, J. Trmal, 和 S. Khudanpur，“SoloSpeech：通过级联生成流程提升目标语音提取的可懂度和质量”，2025年9月，arXiv:2505.19314 [电气工程与系统科学]。 [在线]。可访问：http://arxiv.org/abs/2505.19314

[23] K. ˇZmol´ıkov´a、M. Delcroix、K. Kinoshita、T. Ochiai、T. Nakatani、L. Burget 和 J. ˇCernock´y，《SpeakerBeam：用于语音混合信号中目标说话人语音提取的说话人感知神经网络》，发表于*IEEE* 信号处理精选主题期刊，第13卷第4期，第800–814页，2019年。[24] M. Delcroix、T. Ochiai、K. Zmolikova、K. Kinoshita、N. Tawara、T. Nakatani 和 S. Araki，《利用时域 SpeakerBeam 提高目标语音提取中的说话人区分能力》，发表于*ICASSP 2020 - 2020 IEEE*国际声学、语音与信号处理会议（*ICASSP*），2020年，第691–695页。[25]A. H. Liu、M. Le、A. Vyas、B. Shi、A. Tjandra 和 W.-N. Hsu，《利用流匹配对语音进行生成式预训练》，发表于国际学习表征会议，2024年。 [在线]。可访问：https: //openreview.net/forum?id=z1Qk28Q3L7

[26]A. Vaswani、N. Shazeer、N. Parmar、J. Uszko- reit、L. Jones、A. N. Gomez、L. Kaiser 和 I. Polosukhin，《注意力就是一切》，发表于神经信息处理系统进展，第30卷。Curran Associates, Inc.，2017年。 [在线]。可访问：https://proceedings.neurips.cc/paper files/paper/2017/ hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[27]I. Loshchilov 和 F. Hutter，《解耦权重衰减正则化》，发表于国际学习表征会议，2019年。 [在线]。可访问：https://openreview.net/forum?id=Bkg6RiCqY7[28] J. L. Roux、S. Wisdom、H. Erdogan 和 J. R. Hershey，《SDR——半生不熟还是恰到好处？》，发表于*ICASSP 2019 - 2019*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2019年5月，第626–630页，ISSN：2379-190X。 [在线]。可访问：https://ieeexplore.ieee.org/document/8683855[29] 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》，发表于*2001*年*IEEE*声学、语音与信号处理国际会议。会议论文集（目录号：*01CH37221*），第2卷，2001年5月，第749–752页，第2卷，ISSN：1520-6149。 [在线]。可访问：https: // ieeexplore.ieee.org/document/941023[30] 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度量》，发表于*2010 IEEE*国际声学、语音与信号处理会议，2010年3月，第4214–4217页，ISSN：2379-190X。 [在线]。可访问：https://ieeexplore.ieee.org/document/5495701[31]C. K. A. 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》，发表于*ICASSP 2021 - 2021 IEEE*国际声学、语音与信号处理会议（*ICASSP*），2021年6月，第6493–6497页，ISSN：2379-190X。 [在线]。可访问：https://ieeexplore.ieee.org/document/9414878[32] H. Wang、C. Liang、S. Wang、Z. Chen、B. Zhang、X. Xiang、Y. Deng 和 Y. Qian，《Wespeaker：面向研究与生产的说话人嵌入学习工具包》，发表于*ICASSP 2023 - 2023 IEEE*国际声学、语音与信号处理会议（*ICASSP*），2023年6月，第1–5页，ISSN：2379-190X。 [在线]。可访问：https://ieeexplore.ieee.org/document/10096626/