

基于扩散模型的目标语音提取集成处理方法

王 刚¹, 牟昭曦², 贾春波³, 周立新¹, 顾保国¹

(1. 太原卫星发射中心, 太原 030027; 2. 西安交通大学计算机学院, 西安 710072; 3. 国防科技大学计算机学院, 长沙 410073)

摘 要: 为提升单通道目标语音提取的性能, 本文提出了一种基于深度学习的两阶段处理方法。该方法旨在解决现有判别模型过度优化客观指标而导致输出语音产生伪影、失真, 进而损害主观听感的问题。本文创新性地引入扩散模型作为后处理优化模块, 该模块基于随机微分方程 (Stochastic differential equation, SDE) 框架, 对判别模型的初步输入进行再生式优化。这样不仅有效降低了错误提取, 更显著增强了目标语音的自然度与谐波结构的清晰度。实验结果表明, 在 WSJ0-2mix-extr 数据集上, 所提方法在模拟主观听觉感知的非侵入式语音质量评估 (Non-intrusive speech quality assessment, NISQA) 指标上实现了 11.18% 的提升 (从 3.22 增至 3.58), 显著改善了语音的主观质量与自然度。主观听测结果比较平均意见得分 (Comparative mean opinion score, CMOS) 进一步证实了该方法在提升语音清晰度与可懂度方面的有效性。

关键词: 语音分离; 扩散模型; 语音提取; 说话人信息; 生成模型

中图分类号: TN912.3

文献标志码: A

引用格式: 王刚, 牟昭曦, 贾春波, 等. 基于扩散模型的目标语音提取集成处理方法[J]. 数据采集与处理, 2026, 41(5): 1378-1394. WANG Gang, MU Zhaoxi, JIA Chunbo, et al. Integrated processing method for target speech extraction based on diffusion model[J]. Journal of Data Acquisition and Processing, 2026, 41(5): 1378-1394.

引 言

盲语音分离 (Blind speech separation, BSS)^[1]旨在从混合信号中恢复出各个独立的说话人语音^[2], 是语音识别^[3]、说话人验证^[4]等下游任务的关键预处理环节。尽管盲语音分离已经得到了广泛的研究^[5], 但由于在实际场景中往往无法预知混合语音中包含的说话人数量^[6], 这使得任务变得更加复杂且具有挑战性。为应对此挑战, 目标语音提取 (Target speech extraction, TSE)^[7-8]应运而生。该方法利用目标说话人的参考信息^[8-9], 直接从混合语音中分离出特定个体的语音, 从而规避了对说话人数量的先验需求和输出通道的排列模糊问题^[10]。

随着深度学习技术的不断进步, 语音分离领域的研究主要分为两大类: 判别方法和生成方法。判别方法通过增强混合语音与干净语音之间的信号级度量, 直接实现从混合语音到干净语音的映射^[11]。现有方法可划分为时频域与时域两类: 时频域方法 (如 SpeakerBeam^[12]、Voicefilter^[13] 和 SBF-MTSAL-Concat^[14] 等) 通过对语音频谱特征建模实现分离; 时域方法 (如 SpEx^[15]、SpEx+^[16] 和 CIENet-mDPRNN^[17]) 则直接构建端到端的波形映射。两类方法均以提升混合语音至干净语音的转换性能为核心目标。这些方法在信号质量的客观评价指标上, 如信噪比 (Signal-to-noise ratio, SNR)^[18]、感知语音质量评估 (Perceptual evaluation of speech quality, PESQ)^[19] 和增强型短时可懂度 (Extended

short-time objective intelligibility, ESTOI)^[20]等方面表现出色。然而,判别式模型在优化客观评价指标的同时,往往会引入非自然失真,从而损害用户的主观听觉体验。

生成式方法在语音信号处理中扮演着重要角色,它们通过深度生成模型来模拟干净语音的分布^[21],利用这些先验知识从噪声或混合语音中恢复出清晰的语音信号。在过去几年中,深度生成模型如变分自动编码器(Variational autoencoder, VAE)^[22]、生成对抗网络(Generative adversarial networks, GANs)^[23]和扩散模型^[24-25]等,在无监督学习领域取得了显著进展,它们能够生成自然且高质量的语音样本。扩散模型以其相对稳定性,已经在视觉领域^[23,26]、语音信号处理^[24,27]等领域展现了突出的有效性。由于扩散模型可以捕获数据分布的固有不确定性,并产生属于该分布的有效样本,因此扩散模型会获得更好的感知指标^[28-29]。与此同时,扩散模型对于训练和推断之间的不匹配条件更加稳健,但会导致更严重的逐点失真,参考度量指标较判别模型会有一定的下降^[25]。

扩散模型在目标说话人语音提取领域的应用尚处于初步阶段。通过对扩散模型在其他语音处理任务中的研究进行分析,在目标说话人语音提取任务中可以采用两种主要方法:(1)利用扩散模型同时进行语音分离和生成^[30-31],这种方法要求模型执行多个复杂的任务,可能会导致性能显著下降;(2)受到去混响任务中随机再生模型(Stochastic regeneration model, StoRM)^[32]的启发,将扩散模型作为解码器来优化判别模型的输出,以实现更好的性能。然而,这种方法中的生成过程仅与源信号的先前估计相关,没有充分探索更灵活的设计策略。在现有融合策略中,判别模型与扩散模型常需进行端到端联合训练^[33]。在这种模式下,判别模型的任何变动都要求扩散模块重新训练,不仅拖慢了整体收敛速度,增加了计算资源消耗,且最终性能也可能受限于联合优化的复杂性而未达最优^[34]。

为了缓解这一问题,文献[34]提出了一种预测与细化相结合的方法,旨在降低扩散模型的计算负担,同时确保生成音频的质量。受到这一方法的启发,本文提出了一种结合判别模型和扩散生成模型的目标语音提取方法“判别-扩散融合网络”(Discriminant-diffusion fusion network, DDFN)。在本方法中,判别模型作为确定性模块负责提供对目标语音的初步估计,而扩散生成模型作为随机模块则在此基础上引入随机性,以增强模型的生成能力。DDFN方法整合了一个基于随机微分方程(Stochastic differential equation, SDE)^[35]的目标说话人提取扩散模型,该模型作为判别模型输出的再生环节,旨在提升语音质量。通过这种方式,DDFN不仅降低了扩散模型在提取过程中引入的错误,还成功生成了具有更加清晰谐波结构的语音。这一改进有效地克服了判别模型可能带来的伪影,进而在无参考感知质量评估和主观听觉感受方面取得了显著进步。

1 方法描述

本文提出了一种基于SDE的目标语音提取的再生改进方法。本方法采用级联架构,融合了判别模型与扩散生成模型,构成一个两阶段的目标语音提取框架。其核心目标是修正判别模型在初步提取过程中引入的信号扭曲与失真。通过这种方法,在保持判别模型的参考度量性能的同时,能够改善无参考感知度量的性能和主观听觉感受。

模型的总体框架如图1所示,主要由3个模块组成:判别网络、参考特征提取网络和扩散再生网络。在目标语音提取过程中,将混合语音 y 和目标说话人参考语音 c 作为输入,利用判别网络从混合语音中提取目标语音信号 y_{predict} ,并实现对参考语音的特征提取。与此同时,特征提取网络将目标说话人索引语音 c 作为输入,生成目标说话人的音色特征嵌入 X_{vec} 。然后,参考特征提取网络实现索引嵌入特征 aux-emb 和音色特征 X_{vec} 的融合。最后,将判别网络生成的预测值 y_{predict} 、参考特征提取网络生成的说话人融合索引特征送到扩散再生网络中。在 N 个扩散步骤中,对生成的样本进行不断更新,最终生成高质量的分离语音信号。

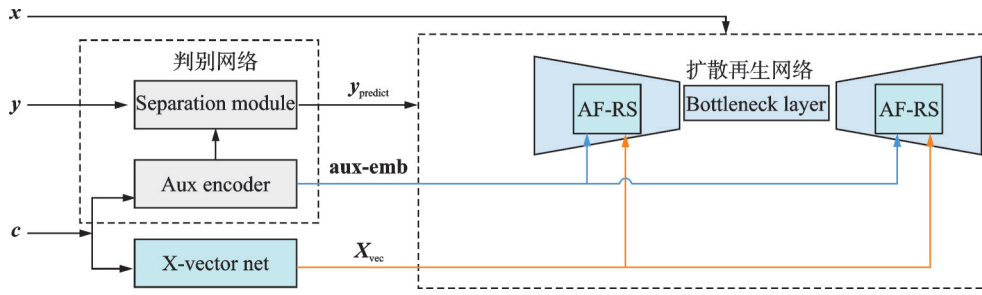


图1 DDFN模型架构图

Fig.1 Architecture diagram of DDFN model

图1中 Separation module 为判别网络中的分离模块, Aux encoder 为参考语音特征提取模块, X-vector net^[36] 为参考语音 X_{vec} 特征提取网络, AF-RS 为得分网络中的特征融合残差网络。图中输入包括清晰语音 x 、混合语音 y 、参考语音 c , y_{predict} 为初始估计目标语音, aux-emb 为参考语音特征提取模块提取的参考语音特征, X_{vec} 为参考语音的音色特征表示。

本方法是集成了判别模型 $\text{Pr}_{\delta}(\cdot)$ 和扩散生成模型 $D_{\varphi}(\cdot)$ 。判别模型利用混合语音 y 和参考语音 c 产生一个初始估计 $\text{Pr}_{\delta}(y, c)$, 这个初始估计存在扭曲失真、细节缺失等问题, 初始估计 y_{predict} 表示为

$$y_{\text{predict}} = \text{Pr}_{\delta}(y, c) = x_0 + d - m \quad (1)$$

式中: x_0 为清晰语音, d 为语音分离时的残余部分或噪声, m 为判别模型的参考信息判断失效导致的失真部分。

扩散生成模型 $D_{\varphi}(\cdot)$ 是在判别模型 $\text{Pr}_{\delta}(y, c)$ 的基础上引导生成清晰语音 x_0 , 生成的估计值表示为

$$\hat{x} = D_{\varphi}(\text{Pr}_{\delta}(y, c), \text{aux}) = D_{\varphi}(y_{\text{predict}}, \text{aux}) \quad (2)$$

式中 aux 为判别模型 $\text{Pr}_{\delta}(\cdot)$ 提取的参考语音线索特征和音色特征 X_{vec} 融合得到的参考语音特征。

1.1 判别网络

本方法采用 SpEx+ 作为全时域的目标说话人提取判别模型。SpEx+ 模型由 4 个主要组件构成: 语音编码器、语音解码器、说话人参考语音编码器和说话人分离器。在 SpEx+ 模型中, 掩码频谱被视为调制响应, 这与选择性听觉注意机制相呼应。在训练和推理过程中, 说话人编码器网络作为 SpEx+ 架构的一部分, 通过多任务学习算法同时对目标参考语音特征提取和语音分离任务进行联合优化。这种方法使得模型能够更有效地学习和提取目标说话人的特征, 从而提高了语音分离的性能。

1.2 参考语音特征提取模块

目标说话人的参考语音在目标语音提取任务中扮演着关键角色, 因为它直接影响到特征提取的效果和网络性能。参考语音嵌入是为了区分不同说话人语音特征的差异, 使得参考语音特征能够描述说话人的物理特性, 分离系统从而能够关注特定的语音信号。因此, 参考语音嵌入的目标是增加同一说话人相似性, 并最大化不同说话人之间的差别。为了进一步提高参考音频特征提取效果, 丰富结构信息, 使用 X-vector^[36] 特征作为说话人的音色声纹特征提取, 并使用多尺度语音编码器提取参考语音在分辨率时间尺度上特征。最后对两种进行特征融合处理。

1.2.1 X-vectors 特征提取器

X-vectors 方法是一种用于说话人识别的特征提取技术, 它的核心结构包括多层时延神经网络 (Time-delay neural network, TDNN) 层、统计池化层以及嵌入表示层。这种模型的一个显著特点是其

静态池化层,它允许模型处理任意长度的语音输入,并将它们转换为统一长度的特征向量。

X-vector特征能够高效捕捉并区分不同个体的音色特性,通过最大化类间(不同说话人)差异,显著提升说话人识别的准确性。在实际应用中,参考语音样本 c 通过X-vectors特征提取器处理后,可以得到对应的X-vector特征。这些特征随后通过式(3)转化为参考语音的X-vector嵌入表示 X_{vec} ,为后续的说话人识别任务提供了关键的信息。

$$X_{\text{vec}} = F(\pi|c) \quad (3)$$

式中: $F(\cdot)$ 为X-vectors特征提取器, π 是预训练模型参数。

1.2.2 多尺度特征提取器

参考SpEx+模型,本方法采用短、中、长这3个不同尺度的一维卷积块提取参考语音的上下文多尺度特征嵌入表示。每个卷积块包含多个卷积核,共同覆盖不同长度的窗口。这些并行的一维卷积块分别用于提取短、中和长尺度的特征,对应的卷积核长度分别为 L_1 (短)、 L_2 (中)和 L_3 (长)。如图2所示,该方法采用残差网络模块编码参考语音,并通过平均池化操作提取关键特征。

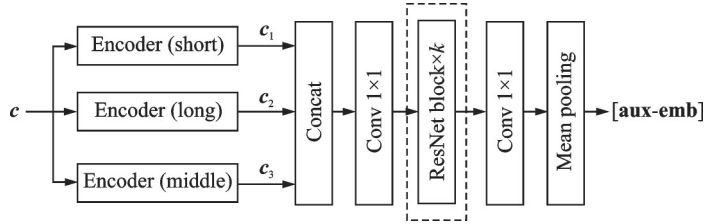


图2 多尺度参考语音特征提取网络结构图^[17]

Fig.2 Architecture diagram of multi-scale reference speech feature extraction network^[17]

图2中, c_1 、 c_2 、 c_3 分别表示不同尺度编码器提取的参考语音的潜空间特征。参考语音 c 在经过语音编码器处理后,形成了多尺度的参考语音嵌入特征表示 C_{emb} ,嵌入表示为

$$C_{\text{emb}} = [c_1, c_2, c_3] = E(\theta|c, L_1, L_2, L_3) \quad (4)$$

式中: $E(\cdot)$ 为语音编码器, θ 为共享参数。

语音信号在不同的时间尺度上呈现出多样的结构特性,涵盖了音素层面的细节、韵律的变化以及内容的丰富性等多个层面。多尺度编码器的优势在于能同时捕获语音信号的宏观结构(如韵律)与微观细节(如音素)^[37],这对构建全面的语音特征表示至关重要。语音信号在较短的时间尺度内表现出细微的变化,而在较长的时间尺度内则表现出稳定特征。多尺度编码器可以通过整合宏观特征和微观细节,进而获得语音信号的全面特征表示。为了提高判别网络的分离能力和多尺度特征网络的特征提取能力,本文方法在多任务学习框架下进行加权联合优化。

1.3 扩散再生网络

扩散模型已成功应用于时域中的离散扩散过程^[38],以及时频域内的随机微分方程的扩散过程^[39]。在使用扩散模型进行语音增强时,反向过程的目标就是估计扰动数据中的白噪声,也就是从语音信号中去除干扰噪声或混响。但是,在目标语音提取任务中,除去目标语音外,混合语音中还有白噪声、其他说话人语音和非结构噪声等,从而导致混合语音或噪声与平稳白噪声的假设不匹配。因此,实际的解决方法是通过沿着过程在干净和混合语音的语音之间进行线性插值^[40],或者通过在SDE的漂移项中定义这样的变换。

本节将探讨基于评分的语音生成模型在目标语音提取方面的应用。为了解决条件扩散模型在缺少参考语音的身份信息,以及无法直接将信息表征集成到网络中的问题,对NCSN++网络架构进行了改进。这一优化使得新的扩散生成模型在目标语音提取任务上的性能得到了显著提升。在本任务中,需要通过条件扩散模型在判别模型结果 $y_{\text{predict}} = D_{\theta}(y, c)$ 的基础上,获得清晰的目标语音信息。本

文方法将该任务直接融入到基于评分的扩散模型的正向和反向过程。

1.3.1 正向过程

在得分驱动扩散模型中,正向过程 $\{x_\tau\}_{\tau=0}^T$ 在一个连续的时间变量 $\tau \in [0, T]$ 进行索引,数据在正向过程中收到噪声的干扰,表示为

$$dx_\tau = f(x_\tau, y_{\text{predict}})d\tau + g(\tau)d\mathbf{w} \quad (5)$$

式中: x_τ 为当前过程状态, $\tau=0$ 时的状态 x_0 表示干净语音。 τ 在描述中代表了一个连续性的操作步骤,它与时域或时频域中的具体时间位置或索引不同。这种表示强调了过程的连续性和动态性,而不是在离散的时间点上的具体测量。在实验过程中,为了避免出现 $\tau=0$ 时的异常处理情况,设置1个最小时间步 $\tau_{\min}$ 来代替 $\tau=0$ 。 $f(x_\tau, y_{\text{predict}})$ 为漂移函数, $g(\tau)$ 为扩散系数,后者用于调节在每个时间步骤中引入的高斯白噪声强度。 y_{predict} 为判别模型的输出,代表了初步语音分离结果。 $\mathbf{w}$ 代表标准的维纳过程,是一种随机过程,用于模拟连续时间内的噪声。

与传统的得分模型不同,本文方法中漂移项是 y_{predict} 。通过漂移函数,SDE被特别定制用于重建混合语音的任务。在这个模型的正向过程中,一个清晰的语音信号 x_0 会逐渐地通过维纳过程带来的噪声,转换成一个噪声样本 x_T 。漂移项确保了正向过程能够按照从清晰语音到混合噪声语音 y_{predict} 的均值路径进行转换。漂移函数 $f(x_\tau, y_{\text{predict}})$ 和扩散系数 $g(\tau)$ 分别定义为

$$f(x_\tau, y_{\text{predict}}) = \gamma(y_{\text{predict}} - x_\tau) \quad (6)$$

$$g(\tau) = \sigma_{\min} \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^\tau \sqrt{2 \log \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)} \quad (7)$$

式中 $\sigma_{\min}$ 和 $\sigma_{\max}$ 是维纳过程调度参数,控制在每个时间步中注入高斯噪声的量。漂移函数 $f(x_\tau, y_{\text{predict}})$ 是根据当前时间步 τ 和预测的混合语音 y_{predict} 来确定的函数,指导清晰语音信号如何随时间演变为噪声样本。扩散系数 $g(\tau)$ 是根据时间步 τ 来确定的函数,它控制了在每个时间步骤中引入的噪声的强度,从而影响重建过程的随机性。

将式(6,7)代入式(5),正向过程的随机微分方程为

$$dx_\tau = \gamma(y_{\text{predict}} - x_\tau)d\tau + \sigma_{\min} \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)^\tau \sqrt{2 \log \left(\frac{\sigma_{\max}}{\sigma_{\min}} \right)} d\mathbf{w} \quad (8)$$

式中 γ 为刚度常数,控制从 x_0 到 y_{predict} 的过程。

1.3.2 反向过程

反向过程可以看作是语音恢复任务的过程,是以 y_{predict} 为条件生成 x_0 的条件生成任务。反向过程是另一个扩散过程,该过程的随机微分方程为

$$dx_\tau = \left[-f(x_\tau, y_{\text{predict}}) + g(\tau)^2 \nabla_{x_\tau} \log p_\tau(x_\tau | y_{\text{predict}}) \right] d\tau + g(\tau)d\bar{\mathbf{w}} \quad (9)$$

式中: $d\bar{\mathbf{w}}$ 为一个布朗运动,时间以相反的方向流动, $\nabla_{x_\tau} \log p_\tau(x_\tau | y_{\text{predict}})$ 为得分函数,表示当前过程状态 x_τ 对数数据分布的梯度,用一个基于神经网络模型 $\text{Score}_\epsilon(\bullet)$ 来逼近。

得分模型可以表示为 $\text{Score}_\epsilon(x_\tau, y_{\text{predict}}, \mathbf{aux}, \tau)$,模型接收当前过程状态 x_τ 、损坏语音 y_{predict} 、当前时间步骤 τ 、参考语音嵌入 $\mathbf{aux}$ 作为输入来进行得分运算。由此得到新的反向微分方程为

$$dx_\tau = [-f(x_\tau, y_{\text{predict}}) + g(\tau)^2 \text{Score}_\epsilon(x_\tau, y_{\text{predict}}, \mathbf{aux}, \tau)] d\tau + g(\tau)d\bar{\mathbf{w}} \quad (10)$$

在反向过程中,假定得分模型 $\text{Score}_\epsilon(\bullet)$ 已经得到充分训练,该模型可以近似所有 τ 时的真实得分值。然后可以利用式(10)的反向SDE,在给定受损语音 y_{predict} 和参考语音嵌入 $\mathbf{aux}$ 的条件下,生成干净语音

x_0 。在生成干净语音的过程就是通过反向过程从 $\tau = T$ 迭代到 $\tau = 0$ 来实现的。

本方法使用预测-校正的采样方法^[32]实现反向推理过程。对于 N 步进行采样生成,时间步 τ 是区间 $[0, T]$ 上的 N 个均匀网格。在每个反向步骤 τ 中,预测器使用欧拉-马鲁雅马(Euler-Maruyama)^[39]单步方法来生成 x_τ ,校正器使用分数模型 $\text{Score}_\epsilon(\cdot)$ 的输出来确保结果样本与估计的边缘分布的一致性。

1.3.3 训练过程

对于给定的初始条件 x_0 和条件变量 y_{predict} ,根据式(8)中描述的高斯过程,可以推导出状态 x_τ 的均值 $\mu(x_0, y_{\text{predict}}, x_\tau)$ 和方差 $\sigma(\tau)^2$,有

$$\mu(x_0, y_{\text{predict}}, x_\tau) = e^{-\gamma\tau} x_0 + (1 - e^{-\gamma\tau}) y_{\text{predict}} \quad (11)$$

$$\sigma(\tau)^2 = \frac{\sigma_{\min}^2 (\alpha^{2\tau} - e^{-2\gamma\tau}) \log \alpha}{\gamma + \log \alpha} \quad (12)$$

式中 $\alpha = \frac{\sigma_{\max}}{\sigma_{\min}}$ 。

根据得到的均值和方差,可以使用扰动核在任意时间步 τ 对 x_τ 进行抽样,有

$$p_\tau(x_\tau | x_0, y_{\text{predict}}) = N_c(x_\tau; \mu(x_0, y_{\text{predict}}, x_\tau), \sigma(\tau)^2 I) \quad (13)$$

式中: N_c 表示圆对称复正态分布, I 表示单位矩阵。

将评分模型 $\text{Score}_\epsilon(\cdot)$ 与扰动核的评分 $\nabla_{x_\tau} \log p_\tau(x_\tau | y_{\text{predict}})$ 进行匹配,本质上是在一定的正则化条件下进行隐式和显式的评分匹配^[41],扰动核评分计算过程见式(14)。这个过程被称为降噪评分匹配,其主要目的是估计干净语音信号的分布。通过这种方式,模型能够学习如何从噪声中恢复出清晰的语音,从而有效地实现语音信号的降噪。

$$\nabla_{x_\tau} \log p_\tau(x_\tau | y_{\text{predict}}) = - \frac{x_\tau - \mu(x_0, y_{\text{predict}}, x_\tau)}{\sigma(\tau)^2} \quad (14)$$

图3描述了目标语音提取过程中的扩散再生步骤。首先,目标语音提取的判别网络 $\text{Pr}_\phi(y, c)$ 会输出1个初步的语音 y_{predict} ,该语音可能包含随机噪声、结构噪声或其他形式的干扰。扩散生成模型 $D_\phi(\cdot)$ 通过引入结构噪声和高斯随机噪声,生成初始样本 x_T 并进行得分网络的训练。最后,依据式(10)逆向求解扩散的SDE,从而从估计的后验概率 $p(x | y_{\text{predict}})$ 中得到样本 x_0, x_0 便是最终重建的目标语音。

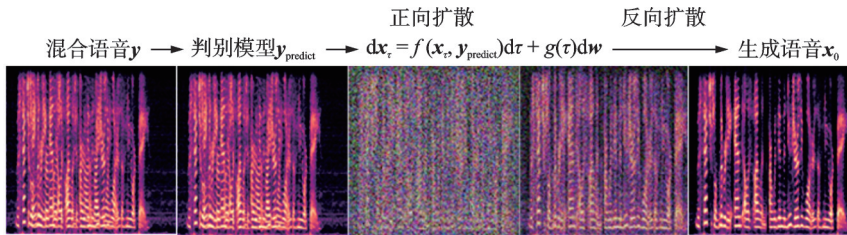


图3 扩散再生过程

Fig.3 Diffusion reverse process

1.3.4 网络架构

本文方法在扩散再生网络中使用的网络架构 NCSN++T 是基于噪声条件化评分网络(NC-SN++)^[35]改进得到的。在扩散再生网络中,使用 NCSN++T 作为条件评分网络。NCSN++T 是在 NCSN++ 的基础上修改以更好地完成目标语音提取任务。在 NCSN++ 网络的基础上,将下采样-上采样结构中的残差块从2个减少到1个。在残差块结构中添加参考语音融合模块,使得网络能够

更好地完成目标说话人语音提取任务。NCSN++T网络因为在时频域内进行运算,所以相关变量都需要通过短时傅里叶变换(Short-time Fourier transform, STFT)转换到时频域。

图4中, h 为上一个残差网络块的输出,同时也是本级残差网络块的输入。对于第一个残差网络来说,正向过程时, h 是当前过程状态 x_τ , $\tau=0$ 时状态 x_0 表示干净语音 x ;逆向过程时, h 是预测网络的输出 y_{predict} 。NCSN++T网络在实现 $\text{Score}_\epsilon(x_\tau, y_{\text{predict}}, \text{aux}, \tau)$ 得分判别过程中,其预测过程和预测结果与参考语音的特征向量直接相关。为了更好地完成目标语音提取任务,需要将说话人参考语音融入到网络中。为了引导生成过程,可采用参考语音 c 作为条件变量。但若直接将参考语音与预测输出 y_{predict} 转换为复数频谱后输入网络,会导致参考语音特征在多次采样转换中可能减弱,难以精确引导分离与生成。因此,本文方法将参考语音特征整合入残差模块,使得网络在采样过程中能根据参考语音的特征引导分离和生成。

1.4 优化目标

本文提出的DDFN模型是一个包括判别阶段和扩散生成阶段两阶段架构。在判别阶段,模型 $\text{Pr}_\delta(\bullet)$ 被训练为根据参考语音 c 从混合语音 y 中分离目标语音信号 x ,并得到参考语音的特征表示 aux-emb 。第一阶段的输出 y_{predict} 和 aux 作为扩散生成阶段的输入, y_{predict} 作为扩散模型初始条件。

1.4.1 判别模型优化目标

判别模型通过多任务学习来进行优化,包括说话人嵌入损失 $\mathcal{L}_{\text{CE}}$ 和目标语音提取损失 $\mathcal{L}_{\text{MS}}$ 。说话人嵌入提取器的优化是通过减少说话人分类任务中的交叉熵损失 $\mathcal{L}_{\text{CE}}$ 。通过最小化损失函数,提取器能够更准确地学习并区分不同说话人的特征表示。

判别模型的优化是将目标说话人参考语音编码器网络和目标语音提取网络进行联合训练,多任务学习不仅有利于说话人提取任务,而且可以使目标说话人编码网络更加有效地提取目标说话人参考语音的特征表示。总的训练损失 $\mathcal{L}_{\text{Total}}$ 是 $\mathcal{L}_{\text{CE}}$ 和 $\mathcal{L}_{\text{MS}}$ 的加权和,即

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{MS}} + \delta \mathcal{L}_{\text{CE}} \quad (15)$$

式中 δ 为权重系数,通常取0.5。

1.4.2 扩散生成阶段优化目标

扩散生成阶段的训练目标是使得分网络预测 $\text{Score}_\epsilon(\bullet)$ 的结果逼近扰动核评分,有

$$\text{Exp}_{\tau, (x_0, y_{\text{predict}}, x_\tau), x_{\text{noise}}, x_\tau} = \left\| \nabla_{x_\tau} \log p_\tau(x_\tau | y_{\text{predict}}) - \text{Score}_\epsilon(x_0, y_{\text{predict}}, \text{aux}, \tau) \right\|_2^2 \quad (16)$$

由于 $\nabla_{x_\tau} \log p_\tau(x_\tau | y_{\text{predict}})$ 可以通过式(14)得分函数 $\text{Score}_\epsilon(x_0, y_{\text{predict}}, \text{aux}, \tau)$ 由模型得出,通过优化后最终的损失函数是模型输出和扰动核评分之间的非加权均方差,有

$$\text{Exp}_{\tau, (x_0, y_{\text{predict}}, x_\tau), x_{\text{noise}}, x_\tau} = \left\| \text{Score}_\epsilon(x_0, y_{\text{predict}}, \text{aux}, \tau) + \frac{x_{\text{noise}}}{\sigma(\tau)} \right\|_2^2 \quad (17)$$

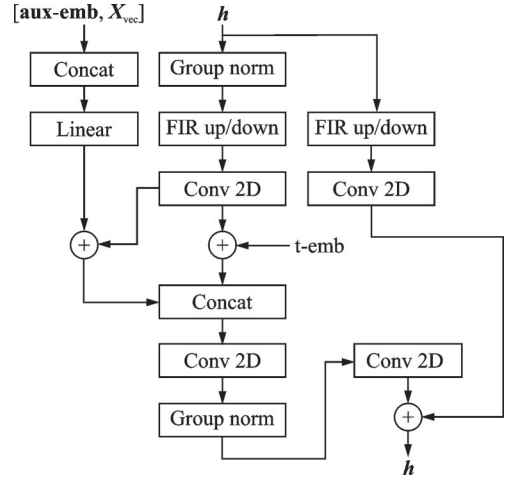


图4 参考语音特征融合残差块网络结构图
Fig.4 Architecture diagram of reference speech feature fusion residual block network

2 实验设置

2.1 数据集

在目标语音提取任务中,WSJ0-2mix-extr^[2]是最为广泛使用的数据集。WSJ0-2mix-extr数据集是WSJ0语料库的一个扩展版本,其中语音样本以8 000 Hz的采样频率进行录制。该数据集共分为3个部分,包含20 000句话的训练集、5 000句话的验证集以及3 000句话的测试集。

WSJ0-3mix-extr数据集^[42]也遵循类似于WSJ0-2mix-extr数据集的创建流程,它包含两个干扰说话者的语音,这些干扰语音与目标说话者的语音以相等的功率水平混合。

2.2 参数设置

(1) 输入配置:音频数据采样率为8 000 Hz,STFT变换窗口大小512,跳步128,使用汉明窗口,每个频谱图的长度为256个STFT时间帧序列。

(2) 扩散模型中,正向过程和反向过程采用相同的参数配置。刚度参数 $\gamma=1.5$,极端噪声系数设定为 $\sigma_{\max}=0.5$ 和 $\sigma_{\min}=0.05$,最小扩散时间 $\tau_{\min}=0.003$ 。在验证过程中,反向扩散过程 $N=100$ 个时间步,使用预测-校正方法,校正步长 $r=0.5$ 。

2.3 训练配置

网络使用Adam优化器^[10],学习率设置为0.000 1,Batchsize设置为8。使用网络权重的指数移动平均,衰减率为0.999。训练得分网络最多训练200个Epoch,验证集损失在连续10个Epoch中没有提升,则终止训练过程。

2.4 评价指标

为评估本方法对目标说话人语音的分离性能,本文采用了基于参考信号的客观评估指标,包括客观语音质量评估指标PESQ^[19]和增强型短时客观可懂度ESTOI^[20]。同时,为评估生成语音的主观听感,本文采用无参考语音质量评估指标NISQA^[43]。鉴于生成模型的目标是提升感知质量而非精确复刻参考信号,NISQA能更准确地反映生成语音的听觉真实性。

客观指标侧重于语音信号的特定技术参数,但并不直接反映听众对语音的整体感知质量。为了获得更全面的语音质量评价,本文结合客观评估和主观的人类感知评估。本文使用主观评价方法比较平均意见得分(Comparative mean opinion score, CMOS)^[44]验证反映听众的听觉体验,作为客观评价指标的补充和完善。

3 实验结果分析

3.1 WSJ0-2mix-extr数据集上的实验结果与分析

3.1.1 与现有方法对比

鉴于生成模型的优化目标是创造具有高自然度和可懂度的语音,其输出可能与参考信号存在逐点差异。因此,传统的信号保真度指标(如SDRi, SI-SNRi)无法全面衡量本文方法的性能。这些指标虽然能够反映信号之间的客观差异,但它们不能完全捕捉到语音的主观感知质量,特别是在评估生成语音的自然度和清晰度方面。因此,本文的评估重点放在了模型生成语音的主观听觉质量上,这包括语音的自然性、清晰度以及在处理扭曲或噪声干扰信号时的恢复能力。本文通过一系列的评价测试来衡量生成语音的质量,如PESQ、ESTOI和NISQA。这些测试能够更好地模拟人类听众对语音质量的感知,从而提供更全面的语音质量评估。

表1展示了本文模型DDFN与其他现有模型在目标说话人语音提取方面的结果对比。表1中,以“—”符号表示文献中的缺失数据,以“*”符号表示复现结果,并以粗体突出最优结果。结果表明,相较于传统的判别模型,本模型在无参考的NISQA指标上具有明显优势。此外,在参考度量指标如PESQ和ESTOI方面,本文模型也展现出与现有模型相当的性能,甚至在某些情况下略有超越。其中,ESTOI指标略有下降,可能是因为DDFN是一种生成模型,对于类似ESTOI这种侵入性客观评价指标时性能会有所下降。特别地,与基线模型SpEx+相比,DDFN模型在无参考NISQA上的表现更加显著,从3.22提升至3.58,增幅达到了11.18%。这一进步表明本文模型在提升语音的真实听觉感受方面取得了显著成效。尽管真实目标语音的NISQA值为3.53,而DDFN生成的语音NISQA值为3.58,略高于真实值。这一现象归因于真实目标语音中可能存在的非语音环境噪声,而本文模型通过有效的扩散噪声处理机制,能够进一步清除这些噪声,从而生成质量更优的语音样本。尽管判别模型SpEx+在PESQ和ESTOI等客观指标上表现优异,其NISQA得分却显著低于真实语音,相差0.33,这表明SpEx+在优化客观指标的过程中引入了伪影或失真,损害了最终输出的主观听感质量。

表1 DDFN方法与现有模型在WSJ0-2mix-extr数据集上的对比实验结果

模型	类型	PESQ	ESTOI	NISQA
混合语音	—	1.81	0.62	2.75
真实目标语音	—	—	—	3.55
SpeakerBeam ^[12]	判别	2.64	—	—
SBF-MTSAL-Concat ^[37]	判别	2.77	—	—
TseNet ^[45]	判别	3.14	—	—
SpEx ^[15]	判别	3.36	—	—
TCN-Conformer ^[46]	判别	3.41	—	—
SEF-Net ^[47]	判别	3.50	—	—
G-DiffSpEx ^[30]	扩散	3.05	—	3.55
SpEx+ ^[16] *	判别	3.48	0.92	3.22
DDFN (Ours)	判别-扩散	3.61	0.91	3.58

目前,仅依赖于扩散模型的目标语音提取方法尚未能够达到理想的性能水平。与之相对,DDFN模型通过融合判别模型的精确分离能力和扩散模型的生成能力,证明了其在目标语音提取任务中的有效性。DDFN不仅在无参考的NISQA指标上取得了提升,而且在保持参考度量指标稳定性方面也表现出色(相较于判别模型,PESQ略有提高,ESTOI有轻微下降)。这种性能的平衡是其他单一模型所难以实现的。这一发现突出了在目标语音提取任务中结合判别和生成模型的策略的重要性。通过这种方法,不仅能够生成在客观指标上表现良好的语音,更重要的是能够生成在主观听觉体验上更自然、更符合人类感知的高质量语音。

图5对比了SpEx+模型和DDFN模型估计的目标语音频谱图。从图5中可以明显看出,DDFN模型在谐波结构的清晰度上取得了显著的提升。这种改进不仅使得语音信号的质量接近真实目标语音,而且在某些方面甚至超越了原始语音。这一点在NISQA指标上的表现也得到了验证,其得分高于真实目标语音值,进一步强调了DDFN在语音质量提升方面的优势。

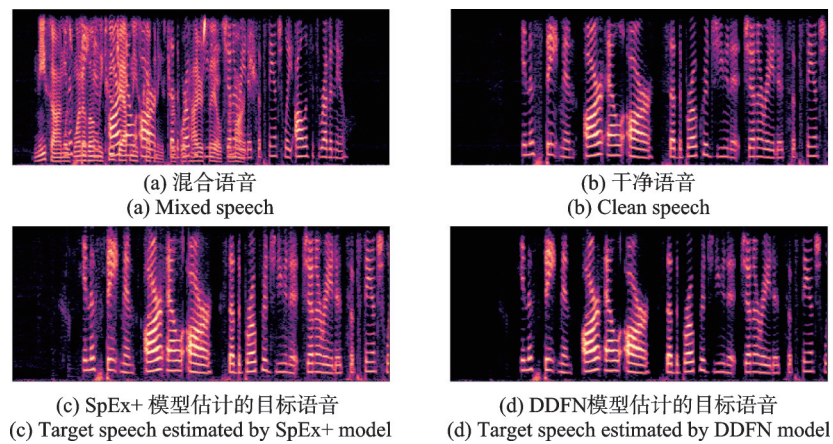


图5 不同方法提取两个说话人混合语音的目标语音频谱图

Fig.5 Spectrograms of target speech extracted by different methods from two-speaker mixed audio

3.1.2 参考语音特征融合方法对比

在问题提出的DDFN模型架构中,得分网络NC-SN++T特别包含了残差模块,该模块融合了说话人参考语音的特征,这对提升扩散再生网络的性能起到了重要作用。为了评估不同特征融合策略对网络性能的具体影响,在表2中展示了实验结果,这些实验包括4种不同的配置。方法1:得分网络NCSN++T中没有融合任何参考语音特征信息;方法2:得分网络NCSN++T中仅融合音色特征 X_{vec} ;方法3:得分网络NCSN++T中仅融合参考语音嵌入特征 aux-emb ;方法4:得分网络NCSN++T中融合音色特征 X_{vec} 和参考语音嵌入特征 aux-emb 。

表2 不同参考语音特征表示的实验结果
Table 2 Experimental results with different reference speech feature representations

输入条件	PESQ	ESTOI	NISQA
方法1	3.40	0.81	3.31
方法2	3.53	0.85	3.36
方法3	3.60	0.89	3.42
方法4	3.62	0.91	3.58

通过对这4组实验结果的分析,可以明确地看到,融合参考语音特征对于提高网络性能起到了显著的作用。当同时融合 X_{vec} 和 aux-emb 特征时,PESQ、ESTOI和NISQA相比于没有融合任何参考语音特征的方法1分别提升了0.22、0.10和0.27,提升比例达到了6.47%、12.6%和4.8%,DDFN模型在目标说话人提取任务上的表现达到了最优。这一结果突出了 X_{vec} 特征在增强模型对说话人身份识别能力方面的重要作用,并且同样突显了 aux-emb 特征在捕获说话人独特语音属性方面的贡献。相比于仅融合 X_{vec} 特征的方法2,方法3在融合 aux-emb 特征时采用了多尺度特征提取网络,该网络利用多尺度编码器整合了不同时间尺度的信息,从而提供了一个更为全面和深入的语音信号特征表示。因此,方法3在PESQ、ESTOI、NISQA三个指标上都优于方法2。方法2和方法3的性能指标大幅优于方法1,验证了将参考语音特征融入残差网络对模型性能提升的重要性,模型能够在多次采用过程中精确引导分离和生成。

3.1.3 逆采样步数对实验结果的影响

逆采样过程是从判别模型生成的样本中逐步逆向生成一个高质量的样本。逆采样的步数影响生成语音的质量和细节。通常来说,逆采样的步数越多,生成的语音样本就越清晰和逼真,减少噪声和伪影。图6展示了10组逆采样步数情况下DDFN模型的尺度不变信噪失真比(Scale-invariant signal-to-distortion ratio, SI-SDR)、PESQ、ESTOI和NISQA四个指标的变化情况。

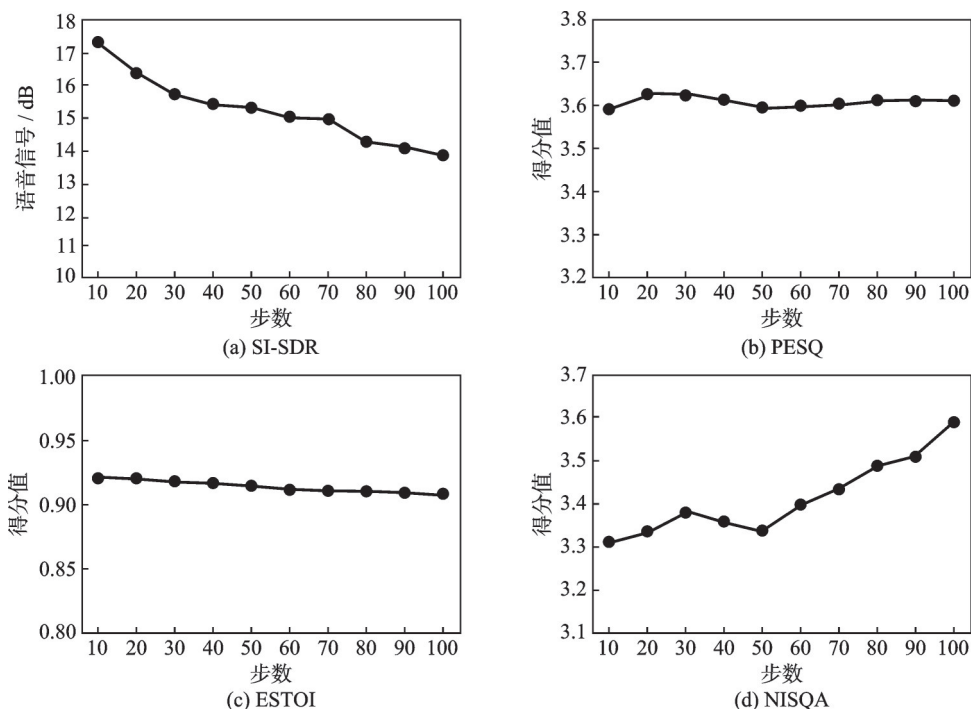


图6 不同逆扩散步数指标变化图

Fig.6 Metric variation diagram with different reverse diffusion steps

图6显示了随着逆采样步数的增加,SI-SDR值呈现下降趋势。这一现象的原因在于SI-SDR指标主要用于衡量信号间的差别,随着逆采样步数的增加,生成的语音与原始语音之间的差异也随之增大。然而,这种差异的增加并不直接意味着生成语音质量的恶化。在逆采样步数逐渐增加的过程中,基于参考度量的PESQ和ESTOI都表现得相对稳定,仅在小范围内波动。这表明尽管SI-SDR值下降,但语音的整体质量和可懂度并未受到显著影响。值得注意的是,NISQA值随着逆采样步数的增加而逐步提升。当逆采样步数达到100步时,NISQA值达到了3.6,这反映出人类主观感知到的语音质量和自然度实际上得到了改善。扩散模型能够捕捉数据分布的内在不确定性并生成符合该分布的有效样本,这种特性使得扩散模型在提升语音质量的感知指标方面具有潜在优势。然而,扩散模型在训练和推断阶段可能存在不匹配问题,随着采样步数的增加可能会导致生成的语音出现较为明显的逐点失真,这在某些情况下可能会损害其在传统参考度量指标上的表现。

实验结果还表明,即使在逆采样步数较少的情况下,例如仅10步,本文模型依然能够维持良好的性能。通过使用预测结果作为扩散模型的条件引导信息,可有效地提高目标生成的准确性,并减少扩散采样的步数。这不仅加快了模型的训练速度,也提升了生成目标语音的速度,从而提高了整个模型的效率和实用性。

3.1.4 主观听觉实验

客观评估指标虽然在语音质量分析中扮演着重要角色,但它们并不总能完全对应于人类的听觉感知。人类对语音质量的感知是一个复杂的心理物理过程,涉及到多种听觉感知特性,包括清晰度、自然度、响度以及语音的可懂度等。为了更准确地评估语音样本的质量,本文设计并实施了一个中等规模的CMOS听辨实验,共邀请了10名参与者加入。实验涉及10组数据,每组数据包含30对语音样本。在每一对样本中,样本A是使用基线模型SpEx+提取的结果,而样本B则是通过DDFN模型提取的。

实验过程中,每位参与者将分别听取这些成对的语音样本。在听完每一对样本之后,参与者需要根据自己的听觉感知,选择出他们认为音质更佳的样本,实验统计结果如表3所示。

表3 CMOS听辨实验结果
Table 3 CMOS listening test results

参与者	ID-1	ID-2	ID-3	ID-4	ID-5	ID-6	ID-7	ID-8	ID-9	ID-10
A个数	6	4	3	5	7	3	2	5	4	6
B个数	24	26	27	25	23	27	28	25	26	24
A占比/%	20.0	13.3	10.0	16.7	23.3	10.0	6.7	16.7	13.3	20.0
B占比/%	80.0	86.7	90.0	83.3	76.7	90.0	93.3	83.3	86.7	80.0

在对CMOS听辨实验的数据进行细致分析后,发现在300组比较的样本中,有45次评价倾向于基线模型提取的语音样本,占总样本数的15%。相比之下,有225次评价选择了DDFN模型提取的语音样本,占比高达85%。实验数据压倒性地表明,DDFN模型提取的语音在主观偏好上显著优于基线模型。实验结果再次支持DDFN模型在提升语音质量方面的有效性。通过这种直接的听众偏好比较,可以说明DDFN模型在改善语音的清晰度、可懂度和其他重要质量指标方面具有显著的优势。

3.1.5 模型复杂度与推理效率分析

尽管DDFN在语音质量与感知指标上表现出显著优势,但其两阶段结构(判别模型+扩散模型)也引入了更高的计算复杂度。为全面评估模型的实用性,本节对DDFN与基线模型SpEx+在参数量、计算量与推理时间等方面进行对比分析,结果如表4所示。从表4可以看出,DDFN的总参数量为 30.8×10^6 ,较SpEx+ (12.5×10^6)增加了约146%。计算量方面,DDFN的浮点运算数FLOPs为 111.1×10^9 ,是SpEx+的12.8倍。推理时间方面,DDFN单次推理需170 ms,其中扩散部分占128 ms,约为判别部分的3倍,这表明扩散模型的引入显著增加了计算负担。尽管DDFN的计算开销较高,但其在语音质量和主观听感方面的显著优势,使其成为对音质有严苛要求且对实时性不敏感的应用场景的理想选择。

表4 模型复杂度与推理效率对比
Table 4 Comparison of model complexity and inference efficiency

模型	参数量/ 10^6	FLOPs/ 10^9	推理时间/ms
SpEx+	12.5	8.7	42
DDFN(判别部分)	12.5	8.7	42
DDFN(扩散部分)	18.3	102.4	128
DDFN(总体)	30.8	111.1	170

3.2 WSJ0-3mix-extr数据集上的实验结果与分析

相比于盲语音分离,目标语音提取网络不需要提前确定混合语音中的说话人数量。因此,无论混合语音是由两位还是更多说话人的声音组成,目标语音提取网络都能够适应并完成目标语音提取任务,且无需对其网络结构进行任何调整。然而,从两个说话人到多说话人的目标语音提取任务,其难度会有所增加。这是因为在多说话人混合语音的情况下,存在更多的声源重叠和声音交互,这增加了分离和识别目标语音的复杂性。本文在WSJ0-3mix-extr数据集上训练和测试基线模型SpEx+和DDFN的性能,实验结果如表5所示。表5中,以“—”符号表示文献中的缺失数据,以“*”符号表示复现结果,

并以粗体突出最优结果。

表 5 DDFN 方法在 WSJ0-3mix-extr 数据集上的对比实验结果

Table 5 Comparative experimental results of DDFN method on WSJ0-3mix-extr dataset

模型	PESQ	ESTOI	NISQA
混合语音	1.49	0.44	1.99
真实目标语音	—	—	3.49
SpEx+ ^{[16]*}	2.39	0.75	2.71
DDFN (Ours)	2.51	0.82	3.10

实验结果显示,本文提出的DDFN方法在多个关键性能指标上的显著进步,与基线模型相比取得了明确的提升。在客观度量指标方面,PESQ、ESTOI分别实现了0.12、0.09的分数提升,相对于基线模型分别实现了5.1%、9.3%的相对增益。在主观感知指标上,NISQA得分达到了3.10,这比基线模型提高了14.4%。实验成果再次验证了DDFN方法在解决判别模型中出现的伪影等问题具有显著效果,并在提升语音的主观感知质量方面展现了有效性。

图7展示了基线模型和DDFN模型在处理3个说话人混合语音上估计目标语音的频谱图对比结果。可以发现,尽管基线模型能够从混合语音中提取目标语音,但与原始真实语音相比仍存在较大差异。这表明在处理复杂的多人混合语音场景时,基线模型仍面临较大的挑战,与实际应用需求之间存在明显差距。相比之下,DDFN模型虽然在提取目标语音方面与真实语音仍有差距,但相较于基线模型SpEx+,其性能有了显著提升,这在表5中也得到了验证。DDFN模型在提升谐波结构清晰度方面也有明显的效果,有效降低了非语音环境下的噪声干扰。综上所述,DDFN模型在目标语音提取的准确性和清晰度方面均优于基线模型。在处理多人混合语音的复杂场景中,DDFN模型展现出了更强的适应性和优越性。

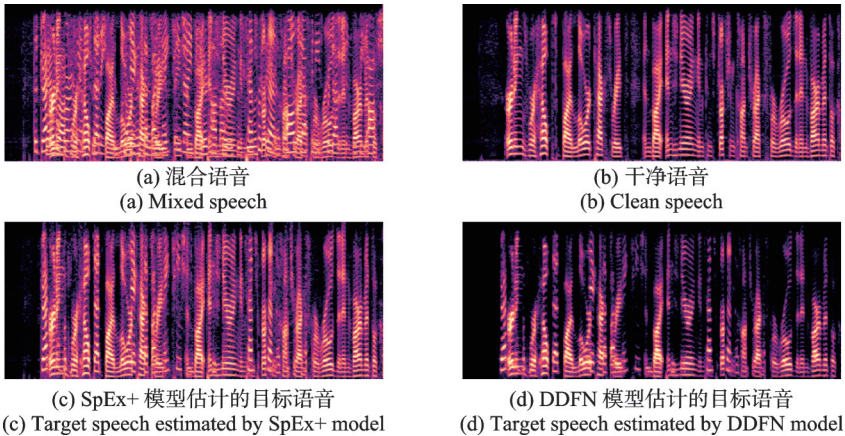


图 7 不同方法提取 3 个说话人混合语音的目标语音频谱图

Fig.7 Spectrograms of target speech extracted by different methods from three-speaker mixed audio

4 结束语

为解决深度学习在单通道语音分离中面临的主观听觉质量瓶颈(如伪影与语音损伤),本文提出了一种融合确定性判别模块与随机性生成模块的扩散模型框架 DDFN。相较于传统判别模型,本文方法

通过判别-生成联合架构实现生成模型性能突破,并基于扩散框架的迭代优化显著提升语音自然度与清晰度,有效抑制了频谱不连续现象。在 WSJ0-2mix-extr 与 WSJ0-3mix-extr 基准测试中,DDFN 的 PESQ、NISQA 指标均显著超越基线模型。DDFN 在 NISQA 指标上的显著提升源于其独特的双模块协同机制,确定性模块通过残差卷积网络精确建模语音谱结构,而随机性模块利用扩散过程逐步修正相位细节,二者的迭代优化有效抑制了传统方法中常见的频谱不连续现象。然而,该方法的实际部署仍面临挑战:扩散模型的迭代采样过程导致推理延迟约为传统模型的 4 倍,且两阶段训练范式对计算资源(特别是显存)提出了更高要求。

参考文献:

- [1] XIAO X, CHEN Z, YOSHIOKA T, et al. Single-channel speech extraction using speaker inventory and attention network [C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2019: 86-90.
- [2] DAVIES M E, JAMES C J. Source separation using single channel ICA[J]. Signal Processing, 2007, 87(8): 1819-1832.
- [3] NASSIF A B, SHAHIN I, ATILI I, et al. Speech recognition using deep neural networks: A systematic review[C]//Proceedings of IEEE Access. [S.l.]: IEEE, 2019: 19143-19165.
- [4] ZHAO F, LI H, ZHANG X. A robust text-independent speaker verification method based on speech separation and deep speaker[C]//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2019: 6101-6105.
- [5] 黄雅婷, 石晶, 许家铭, 等. 鸡尾酒会问题与相关听觉模型的研究现状与展望[J]. 自动化学报, 2019, 45(2): 234-251.
HUANG Yating, SHI Jing, XU Jiaming, et al. Research advances and perspectives on the cocktail party problem and related auditory models[J]. Acta Automatica Sinica, 2019, 45(2): 234-251.
- [6] 李建琛, 韩纪庆. 基于深度学习的说话人确认方法研究现状及展望[J]. 数据采集与处理, 2024, 39(5): 1062-1084.
LI Jianchen, HAN Jiqing. State of the art and prospects of deep learning-based speaker verification[J]. Journal of Data Acquisition and Processing, 2024, 39(5): 1062-1084.
- [7] VELURI B, ITANI M, CHEN T, et al. Look once to hear: Target speech hearing with noisy examples[C]//Proceedings of the CHI Conference on Human Factors in Computing Systems. New York, USA: ACM, 2024: 1-16.
- [8] ZMOLIKOVA K, DELCROIX M, OCHIAI T, et al. Neural target speech extraction: An overview[J]. IEEE Signal Processing Magazine, 2023, 40(3): 8-29.
- [9] 徐亮, 王晶, 杨文镜, 等. 基于 Conv-TasNet 的多特征融合音视频联合语音分离算法[J]. 信号处理, 2021, 37(10): 1799-1805.
XU Liang, WANG Jing, YANG Wenjing, et al. Multi feature fusion audio-visual joint speech separation algorithm based on Conv-TasNet[J]. Journal of Signal Processing, 2021, 37(10): 1799-1805.
- [10] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]//Proceedings of International Conference on Learning Representations. San Diego, USA: [s.n.], 2015: 1-15.
- [11] WANG D, CHEN J. Supervised speech separation based on deep learning: An overview[C]//Proceedings of IEEE/ACM Transactions on Audio, Speech, and Language Processing. [S.l.]: IEEE, 2018: 1702-1726.
- [12] ŽMOLÍKOVÁ K, DELCROIX M, KINOSHITA K, et al. SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures[J]. Proceedings of IEEE Journal of Selected Topics in Signal Processing, 2019, 13(4): 800-814.
- [13] GABRYŚ A, HUYBRECHTS G, RIBEIRO M S, et al. Voice filter: Few-shot text-to-speech speaker adaptation using voice conversion as a post-processing module[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2022: 7902-7906.
- [14] XU C, RAO W, CHNG E S, et al. Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2019: 6990-6994.
- [15] XU C, RAO W, CHNG E S, et al. SpEx: Multi-scale time domain speaker extraction network[J]. IEEE/ACM Transactions

- on Audio, Speech, and Language Processing, 2020, 28: 1370-1384.
- [16] GE M, XU C, WAN H L, et al. SpEx+: A complete time domain speaker extraction network[C]//Proceedings of the Annual Conference of the International Speech Communication Association. [S.l.]: INTERSPEECH, 2020: 1406-1410.
 - [17] YANG Xue, BAO Changchun, ZHOU Jing. et al. Target speaker extraction by directly exploiting contextual information in the time-frequency domain[C]//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2024: 10476-10480.
 - [18] SHANNON C E. A mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27 (3): 623-656.
 - [19] ITU-T. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs: ITU-T Recommendation P.862[S]. [S.l.]: [s.n.], 2001.
 - [20] WANG H, HAI J, LU Y J, et al. SoloAudio: Target sound extraction with language-oriented audio diffusion Transformer [C]//Proceedings of ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2025: 7509-7513.
 - [21] MITTAG G, NADERI B, CHEHADI A, et al. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets[C]//Proceedings of Interspeech 2021. [S.l.]: [s.n.], 2021: 2127-2131.
 - [22] VINCENT P. A connection between score matching and denoising autoencoders[J]. Neural Computation, 2011, 23(7): 1661-1674.
 - [23] 刘美琴, 陈虹宇, 周一鸣, 等. 低码率生成式无人机视频编码算法[J]. 数据采集与处理, 2025, 40(2): 320-333.
LIU Meiqin, CHEN Hongyu, ZHOU Yiming, et al. Low bit rate generative drone video compression[J]. Journal of Data Acquisition and Processing, 2025, 40(2): 320-333.
 - [24] LU Y J, WANG Z Q, WATANABE S, et al. Conditional diffusion probabilistic model for speech enhancement[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2022: 7402-7406.
 - [25] SAITO K, MURATA N, UESAKA T, et al. Unsupervised vocal dereverberation with diffusion-based generative models [C]//Proceedings of ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023: 1-5.
 - [26] 刘安安, 苏育挺, 王岚君, 等. AIGC 视觉内容生成与溯源研究进展[J]. 中国图象图形学报, 2024, 29(6): 1535-1554.
LIU Anan, SU Yuting, WANG Lanjun, et al. Review on the progress of the AIGC visual content generation and traceability [J]. Journal of Image and Graphics, 2024, 29(6): 1535-1554.
 - [27] WANG H, VILLALBA J, MORO-VELAZQUEZ L, et al. Noise-robust speech separation with fast generative correction [C]//Proceedings of INTERSPEECH 2024. [S.l.]: [s.n.], 2024.
 - [28] MA H, CHEN R, JING R, et al. Enhancing Intelligibility for generative target speech extraction via joint optimization with target speaker ASR[J]. IEEE Signal Processing Letters, 2025, 32: 2309-2313.
 - [29] WANG H, HAI J, LU Y J, et al. SoloAudio: Target sound extraction with language-oriented audio diffusion Transformer [C]//Proceedings of ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2025: 7509-7513.
 - [30] NGUYEN T, SUN G, ZHENG X, et al. Conditional diffusion model for target speaker extraction[C]//Proceedings of ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2024.
 - [31] JAYARAM V, THICKSTUN J. Source separation with deep generative priors[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2020: 4724-4735.
 - [32] LEMERCIER J M, RICHTER J, WELKER S, et al. StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation[C]//Proceedings of IEEE/ACM Transactions on Audio, Speech and Language Processing. [S.l.]: IEEE, 2023: 2724-2737.
 - [33] VARIANI E, LEI X, MCDERMOTT E, et al. Deep neural networks for small footprint text-dependent speaker verification [C]//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2014: 4052-4056.

- [34] HIRANO M, SAWATA R, MURATA N, et al. Diffusion-based signal refiner for speech enhancement and separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2026, 34: 1426-1440.
- [35] SONG Y, SOHL-DICKSTEIN J, KINGMA D P, et al. Score-based generative modeling through stochastic differential equations[C]//Proceedings of International Conference on Learning Representations (ICLR). [S.l.]: [s.n.], 2021.
- [36] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: Robust DNN embeddings for speaker recognition[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2018: 5329-5333.
- [37] KINGMA D P, WELING M. Auto-encoding variational Bayes[C]//Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014). Banff, Canada: [s.n.], 2014.
- [38] TENG X, TIAN X, POEPEL D. Testing multi-scale processing in the auditory system[J]. Scientific Reports, 2016, 6: 34390.
- [39] LU Y J, WANG Z Q, WATANABE S, et al. Conditional diffusion probabilistic model for speech enhancement[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2022: 7402-7406.
- [40] TOLEDANO D T, FERNÁNDEZ-GALLEGO M P, LOZANO-DIEZ A. Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on TIMIT[J]. PloS One, 2018, 13(10): e0205355.
- [41] HYVARINEN A, DAYAN P. Estimation of non-normalized statistical models by score matching[J]. Journal of Machine Learning Research, 2005, 6(4): 695-709.
- [42] XU C, RAO W, CHENG E S, et al. Time-domain speaker extraction network[C]//Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). [S.l.]: IEEE, 2019: 327-334.
- [43] WELKER S, RICHTER J, GERKMANN T. Speech enhancement with scorebased generative models in the complex STFT domain[C]//Proceedings of ISCA Interspeech. [S.l.]: [s.n.], 2022: 2928-2932.
- [44] STREIJL R C, WINKLER S, HANDS D S. Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives[J]. Multimedia Systems, 2016, 22(2): 213-227.
- [45] LUO Y, MESGARANI N. TasNet: Time-domain audio separation network for real-time, single-channel speech separation [C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2018: 696-700.
- [46] SINHA R, TAMMEN M, ROLLWAGE C, et al. Speaker-conditioning single-channel target speaker extraction using conformer-based architectures[C]//Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC). [S.l.]: IEEE, 2022: 1-5.
- [47] ZENG B, SUO H, WAN Y, et al. SEF-Net: Speaker embedding free target speaker extraction network[C]//Proceedings of the Annual Conference of the International Speech Communication Association. [S.l.]: INTERSPEECH, 2023: 20-24.

作者简介:



王刚(1987-),男,硕士研究生,研究方向:语音信号处理、航天数据智能处理,E-mail:2300436484@qq.com。



牟昭曦(1998-),男,博士研究生,研究方向:语音信号处理。



贾春波(1987-),通信作者,男,助理研究员,研究方向:多模信号处理,E-mail:wgcsn@163.com。



周立新(1987-),男,硕士研究生,研究方向:数据治理与深度学习。



顾保国(1979-),男,高级工程师,研究方向:航天测控数据智能处理。

(编辑:刘彦东)

Integrated Processing Method for Target Speech Extraction Based on Diffusion Model

WANG Gang¹, MU Zhaoxi², JIA Chunbo^{3*}, ZHOU Lixin¹, GU Baoguo¹

(1. Taiyuan Satellite Launch Center, Taiyuan 030027, China; 2. College of Computer, Xi'an Jiaotong University, Xi'an 710072, China; 3. College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract: This paper introduces a two-stage deep learning approach to enhance the performance of single-channel target speech extraction. Conventional discriminative models often over-optimize objective metrics, leading to unnatural artifacts, distortions, and compromised subjective auditory quality in the extracted speech. To address this limitation, we innovatively incorporate a diffusion model as a second-stage refinement module. This module, built upon a stochastic differential equation (SDE) framework, regenerates and refines the initial output from the discriminative model. This process not only reduces extraction errors but also significantly improves the naturalness and harmonic clarity of the target speech. Experimental results on the WSJ0-2mix-extr dataset show that our method achieves an 11.18% improvement (from 3.22 to 3.58) on the non-intrusive speech quality assessment (NISQA) metric, which simulates subjective auditory perception. This indicates a substantial enhancement in perceived speech quality and naturalness. Furthermore, subjective listening tests using the comparative mean opinion score (CMOS) confirm the effectiveness of our approach in boosting speech clarity and intelligibility.

Highlights:

1. A two-stage discriminative-diffusion fusion network (DDFN) is proposed, wherein diffusion-based regeneration refines discriminative model outputs to effectively mitigate artifacts and distortions arising from over-optimization of objective metrics, substantially enhancing the speech naturalness and harmonic clarity.
2. The NCSN++T score network is enhanced within a stochastic differential equation (SDE) framework by integrating X-vector timbre representations and multi-scale reference speech features into residual blocks, thereby guiding reverse diffusion sampling to achieve accurate target speech reconstruction with reinforced harmonic structures.
3. The efficacy of DDFN is validated through both objective and subjective evaluations: NISQA scores of 3.58 and 3.10 on WSJ0-2mix-extr and WSJ0-3mix-extr datasets, corresponding to relative improvements of 11.18% and 14.40% over the baseline; and an 85% preference rate in CMOS subjective listening tests.

Key words: speech separation; diffusion model; speech extraction; speaker information; generative model