# DEPFLOW：用于缓解抑郁症检测中语义偏差的解耦语音生成

预印本

**Yuxin Li[1*], Xia ngyu Zhang[2*]、Yifei Li[1], Zhiwei Guo[1],Haoyang Zhang[3],Eng Siong Chng[1],Cuntai Guan[1] ,[4]**

[1]计算与数据科学学院，南洋理工大学，新加坡[2]电气工程与电信学院，新南威尔士大学，悉尼，澳大利亚[3]软件与微电子学院，北京大学，中国北京[4]医学人工智能中心（C-AIM），李光前医学院，南洋理工大学，新加坡[*]共同贡献 {yuxin.li, YIFEI024, ZHIWEI004, ASESChng, ctguan}@ntu.edu.sg xiangyu.zhang2@unsw.edu.au; zhang.haoyoung@stu.pku.edu.cn

arXiv:2601.00303v1 [cs.CL] 1 Jan 2026

2026年1月5日

**摘要**

语音是一种可扩展、无创的生物标志物，可用于早期心理健康筛查。然而，DAIC-WOZ 等广泛使用的抑郁症数据集存在语言情感与诊断标签之间的强相关性，促使模型学习语义捷径。因此，在现实场景中，模型鲁棒性可能受到影响，例如伪装性抑郁患者尽管处于潜在抑郁状态，仍会使用积极或中性的社交语言。为缓解这种语义偏差，我们提出 DepFlow，这是一种分三个阶段的抑郁状态条件化文本转语音框架。首先，抑郁症声学编码器通过对抗训练学习与说话人和内容无关的抑郁嵌入，在有效实现解耦的同时保留对抑郁症的判别能力（ROC-AUC：0.693）。其次，采用 FiLM 调制的流匹配文本转语音模型将这些嵌入注入合成过程，从而能够控制抑郁严重程度，同时保留内容和说话人身份。第三，基于原型的严重程度映射机制可在抑郁连续谱上实现平滑且可解释的操控。我们利用 DepFlow 构建了一个面向伪装性抑郁的数据集 CDoA，该数据集将抑郁声学模式与情感分层文本库中的积极/中性内容配对，从而生成自然数据中占比较低的声学-语义不匹配。在三种抑郁症检测架构上的评估结果显示，CDoA 分别将宏平均 F1 提高了 9%、12% 和 5%，且在抑郁症检测中始终优于传统增强策略。除了增强鲁棒性，DepFlow 还为对话系统和基于模拟的评估提供了可控的合成平台；在这些场景中，真实临床数据仍受伦理与覆盖范围限制。

关键词 语音抑郁症检测 · 语音合成 · 伪装性抑郁 · 掩饰性抑郁 · 语义偏差

## 1 引言

抑郁症会以规律的方式改变语音生成，包括发声能量降低和发音不清晰 Cummins 等 [2015], Low 等 [2010]。这些变化与自主神经功能失调有关，并反映了精神运动迟滞、情感表达能力下降等核心临床症状 Bandyopadhyay Prasanta 等[2014], McCorry [2007], Almaghrabi 等 [2023]。重要的是，这类声学标志往往在人们意识到或报告自身抑郁症状之前就已出现 Morales-Luque 等 [2025], Brown [2025]。因此，语音是一种可扩展、无创且相对难以故意操纵的生物标志物，可用于早期和

被动式心理健康筛查，尤其适用于需要确保不同表达风格下均具备稳健性的真实世界监测场景。

语音通过多个相互作用的层面传递信息，包括声学线索、语言内容和说话人身份 Scherer [2003]。尽管这三者都会影响情感表达，但我们的分析发现，语言情感是广泛使用的 DAIC-WOZ 数据集中的一个关键混杂因素。具体而言，通过考察 DAIC-WOZ 中近一万个话语，我们发现，负面情感在抑郁语音中的出现频率显著高于健康语音。这种由数据集诱导的耦合表明，语言情感并不独立于抑郁状态，这意味着在此类数据上训练的模型自然会倾向于依赖情感线索，而非抑郁相关声学标记 Geirhos 等 [2020], Pearl [2009], Schölkopf 等 [2021]。

当语言内容与声学表达出现分歧时，这种风险尤为明显，例如，抑郁症患者可能仍保持积极或中性的社交语言，同时表现出抑郁语音模式。这与临床上的伪装性抑郁概念相一致：患者可能有意识或无意识地掩饰症状，以符合社会期望 Jansson 等 [2016], Brown [2025], Shetty 等 [2018],。在这些情况下，声学-语义不匹配并非异常现象，而可能在特定人群中系统性地出现。因此，当语义内容不再反映内在情感状态时，将语言情感视为抑郁症代理指标的模型很容易失效，从而导致其在实际部署中的稳健性下降。

为缓解语义偏差并减少捷径学习，我们提出在训练过程中让模型明确接触声学信息与语义信息之间受控的冲突。为此，我们提出 DepFlow，这是一种波形级、抑郁症条件语音生成框架，能够合成伪装式抑郁导向增强（CDoA）数据集；在保留说话人身份的同时，可独立操控抑郁相关声学表现力与语言内容。我们并非生成任意的合成数据，而是专门构建用于模拟伪装性抑郁场景的数据集：对于抑郁症说话人，DepFlow 会生成额外语音，其中保留其原有的抑郁症声学模式，同时配以外部提供的积极语言内容；这种组合在真实数据集中极少出现。对于健康说话人，也会生成以积极内容为主的额外语音。通过提供此类系统构建的高保真语音，其中抑郁声学特征被刻意置于与积极语义相冲突的情境中，DepFlow 促使下游模型减少对语义极性的依赖，转而将决策边界调整至依赖不变的声学生物标志物。

DepFlow 通过三个核心组件实现这一目标。首先，抑郁声学编码器（DAE）学习抑郁相关声学模式的连续流形，同时通过对抗目标抑制说话人身份和语言信息。其次，条件流匹配文本转语音模型以这些抑郁嵌入为条件，合成波形级语音，从而在保留语言内容和说话人特征的同时，精确控制抑郁相关声学表现力。最后，基于原型的严重程度控制机制能够在推理时平滑且可解释地调节抑郁表达性，使 DepFlow 可以系统地生成面向伪装的训练样本，打破现有数据集中存在的情感-诊断相关性。

我们通过增强三种具有代表性的基于深度学习的抑郁症检测架构的训练数据来评估 DepFlow。在所有模型中，加入 DepFlow 生成的解耦语音后，宏平均 F1 分数均得到稳定提升，表明决策边界更加稳定、鲁棒。此外，DepFlow 的表现始终优于现有的数据增强策略，凸显了有针对性的声学—语义解耦相较于通用的基于扰动的方法所带来的优势。这些结果表明，让模型接触与非抑郁性语言内容配对的抑郁声学特征，能够有效减少语义捷径学习，并使模型将关注重点转向不变的声学生物标志物。

除了提升抑郁症检测的鲁棒性之外，DepFlow 还是一种面向抑郁症的语音数据生成工具，可支持心理健康建模领域新兴的研究范式。具体而言，基于对话的抑郁症交互式研究需要大规模、多样且可控的语音数据 Zhang 等人 [2023], Ji 等[2024], ，但由于伦理限制、隐私顾虑以及对表达多样性的覆盖有限，这类数据难以从真实世界的临床录音中获取 Cummins 等 [2015], Liet al. [2025a]。这种数据生成能力为抑郁症感知对话系统、交互建模和基于模拟的评估框架等未来研究铺平了道路，因为仅依赖真实数据不足以支持训练和分析。

本文其余部分安排如下。第二节回顾抑郁症检测和数据增强方面的相关工作。第三节介绍所提出的 DepFlow 框架，并详细说明其各个组成模块。第四节概述实验设置，包括数据集、预处理、实现细节、CDoA 数据集的构建、下游抑郁症检测架构以及评估指标。第五节展示实验结果并讨论主要发现。第六节总结全文。

## 2 相关工作

### 2.1 基于深度学习的语音抑郁症检测

该领域的大部分进展得益于 AVEC 挑战赛系列 Valstar 等人 [2013], Ringeval 等人[2019], ，该系列为多模态抑郁评估建立了标准基准。

早期系统依赖手工设计的声学特征，例如韵律线索（基频、能量、语速）、频谱描述符（梅尔频率倒谱系数、共振峰）和语音质量指标（抖动、微闪、谐噪比）Li 等人 [2025a]。这些方法捕捉到了与精神运动迟滞相关的临床指标，但需要人工设计，且鲁棒性不足。

端到端模型通过直接从原始音频或梅尔频谱图中学习，解决了这些局限。基于卷积神经网络的架构能够提取局部频谱模式 Chlasta 等人 [2019], Dubagunta 等人 [2019], Saidi 等人 [2020], Othmani 等人 [2021], Vázquez-Romero 和 Gallardo-Antolín [2020], Wang 等人 [2022], ，而循环神经网络对时间动态进行建模，Transformer 则捕捉长程依赖关系 Salekin 等人 [2018], Zhao 等人 [2020], Muzammel 等人 [2020]。

近年来，自监督学习（SSL）凭借从大量未标注语料库中学习丰富声学表征的能力，已成为主流范式。Wav2Vec 2.0 Baevski 等人 [2020], HuBERT Hsu 等人 [2021], WavLM Chen 等人 [2022a], 和 Whisper Radford 等人 [2023] 等模型展现出强大的迁移能力。基于自监督学习的特征始终优于人工设计的特征和传统的端到端方法，尤其是在低资源抑郁症检测任务中 Zhang 等人 [2021], Toto 等人 [2021], Chen 等人 [2022b], Ravi 等人 [2022a], Wu 等 [2023], Li 等 [2025b], Zhang 等人 [2025]。

尽管取得了这些进展，大多数模型仍会隐式地学习训练数据中语言情感与抑郁标签之间的强相关性。这种语义偏差会导致模型在文本情感与声学情感不一致时变得脆弱。

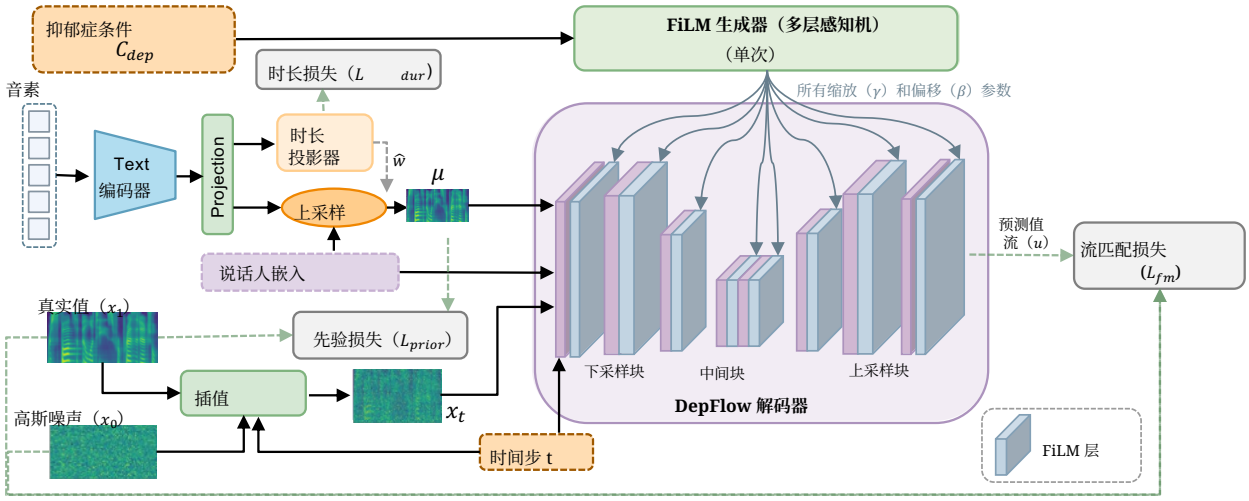



图1：**DepFlow 的训练流程**。DepFlow 将音素序列、说话人嵌入和抑郁条件嵌入$\mathbf{c}_{\text{dep}}$ 作为条件输入。文本编码器提取语言特征，时长模块将其与声学帧对齐，而条件流匹配解码器则通过基于FiLM的抑郁症条件控制生成梅尔频谱图。

### 2.2 基于语音的抑郁症检测的数据增强

数据增强方法可分为类内策略和类间策略 Lishi and Mak [2025]。

#### 2.2.1 类内增强

类内方法会扰动样本，但不会改变其诊断标签。传统方法采用音高偏移、速度变化或加性噪声。FrAUG Ravi 等 [2022b] 和 SpecAugment Park 等 [2019] 等更先进的技术则作用于时频表示，以提升鲁棒性。领域感知方法还会进一步操纵结构或音素级属性，以引入有针对性的变化，例如对话打乱 Wu 等[2023], 、元音级扰动 Feng 和 Chaspari [2023],或声谱图分辨率变化 Kumar 等 [2024]。

尽管有效，这些方法只能丰富类内多样性。它们既无法解决不同诊断类别之间的标签-特征纠缠，也无法缓解由情感–诊断相关性引发的语义偏差。

#### 2.2.2 类间增强

类间增强旨在生成能够体现不同诊断类别间转换的样本。研究者尝试过 Mixup 式插值，但其线性混合标签的假设在临床任务中往往不成立，从而产生临床上无效的样本 Lishi and Mak [2025]。基于生成对抗网络的方法 Goodfellow 等人 [2020] 尝试实现类别间转换，包括通过 CycleGAN 变体进行情感风格迁移 Yang 等人 [2020]。然而，生成对抗网络难以对抑郁严重程度进行稳定且细粒度的控制，并且通常在特征空间而非波形级进行操作。

反事实增强通过显式修改潜在表征，提供了一种更具理论依据的替代方案。Zuo 和 Mak Lishi and Mak [2025] 提出了基于 VQ 的反事实操控，将话语映射至相反诊断类别，并证明该方法在数据稀缺情况下能够提升鲁棒性。

随着神经语音合成技术的发展，波形级增强日益受到关注。DepressGEN Liang 等人[2025] 使用可控文本转语音生成连贯的抑郁特征样本，并结合大语言模型驱动的文本生成与韵律控制。然而，以往系统通常依赖韵律模仿，而非解耦的声学–语义操控，也未直接致力于纠正语义偏差。

## 3 方法

### 3.1 概述

图 1 展示了系统的整体架构，该系统由三个组件构成：抑郁声学编码器（DAE）、DepFlow 文本转语音模型，以及基于原型的严重程度映射机制。该框架将抑郁声学特征的学习与语音合成分离：DAE 首先学习一个不受说话人和内容影响的抑郁症声学模式流形，随后 DepFlow 使用这一嵌入来合成解耦语音，在保留说话人身份的同时，将抑郁症声学线索注入任意文本。最后，严重程度映射模块为抑郁表达性提供连续且可解释的控制，使系统能够生成面向伪装的语音样本，系统性地打破标准数据集中存在的情感-诊断相关性。

### 3.2 抑郁症声学编码器

为将抑郁症相关声学特征与语言和说话人因素分离，我们设计了一个抑郁声学编码器（DAE），它在抽象掉说话人身份和语言内容的同时，学习抑郁症声学模式的连续流形。这种解耦对于解耦语音生成至关重要：必须在不扭曲音素或改变说话人特征的前提下修改抑郁线索，否则合成样本在用于决策边界校正时将缺乏临床可靠性。DAE 的架构如图 2 所示。

#### 3.2.1 模型结构

一个预训练的 WavLM-Large 模型 Chen 等人 [2022a] 首先提取帧级上下文化特征 $\{\mathbf{x}_t\}_{t=1}^{T}$。每一帧通过带有 ReLU 和随机失活的线性层投影为 256 维嵌入表示，得到 $\mathbf{h}_t$。随后，基于注意力的统计池化层通过计算注意力权重 $\alpha_t = \mathrm{softmax}(f\mathrm{attn}(\mathbf{h}_t))$，并拼接注意力加权均值和标准差，将这些帧嵌入聚合为话语级表示。

由带有层归一化、SiLU 激活和随机失活的全连接层组成的后编码器，将这一池化嵌入转换为 32 维的抑郁相关声学嵌入 $\mathbf{d}$，该嵌入作为所有下游分支共享的表示。

在该嵌入之上连接了三个轻量级头部：

**序数回归头** 该头部预测 $K-1=4$ 个有序阈值，用于表示 $K=5$ PHQ-8 严重程度等级 Kroenke 等人 [2009]。给定 $\mathbf{d}$，一个小型多层感知机输出阈值逻辑值。

**说话人解耦头** 为实现说话人不变性，将说话人分类器应用于经 L2 归一化的˜嵌入 $\mathbf{d}$，并对 $S$ 个说话人进行预测。同一分类器也用于配备了一个

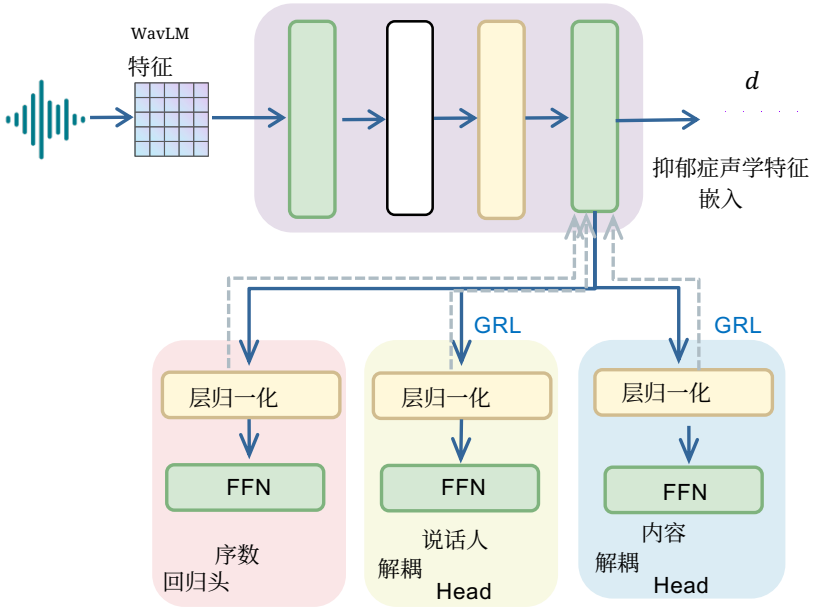



图 2：**抑郁声学编码器（DAE）的架构**。将帧级 WavLM 特征聚合为话语级表示，以生成抑郁症声学嵌入 $\mathbf{d}$；序数回归头用于估计基于患者健康问卷（PHQ）的严重程度，而基于梯度反转层（GRL）的说话人和内容解耦头则抑制说话人身份和语言信息。

梯度反转层（GRL）Ganin 等人 [2016], ，通过反转梯度来抑制共享嵌入中的说话人特定信息。

˜**语言解耦头**语言对抗分支可减少$\mathbf{d}$中的残余语音内容。预训练的 HuBERT 模型 Hsu 等人 [2021] 生成帧级伪音素标签 $\{c_t\}$，并通过多数投票将其合并为话语级类别。归一化后的嵌入经过另一个 GRL 和一个分类器 $f_{\text{con}}$，用于预测 $C$ HuBERT 派生单元。该对抗信号可抑制语言信息。

### 3.2.2 优化目标

训练过程联合优化抑郁症预测，以及从共享嵌入 $\mathbf{d}$ 中去除说话人和语言信息。序数回归头使用二元交叉熵进行训练，目标为单调序数目标：

$$\mathcal{L}_{\text{sup}} = -\sum_{k=1}^{K-1}\Big[t_k \log \sigma(o_k) + (1-t_k)\log\big(1-\sigma(o_k)\big)\Big], \tag{1}$$

并采用根据标签频率计算的类别平衡权重。

˜应用于 $\mathbf{d}$ 的说话人分类器使用交叉熵身份识别损失 $\mathcal{L}_{\text{id}}$，其对抗分支则通过对抗说话人损失 $\mathcal{L}_{\text{adv-spk}}$ 促使模型学习说话人不变嵌入。同样，语言对抗分支使用基于 HuBERT 派生标签的交叉熵损失 $\mathcal{L}_{\text{adv-con}}$ 进行训练，以抑制语言内容。

完整的训练目标是加权和：

$$\mathcal{L}_{\text{total}} = \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{spk}}\mathcal{L}_{\text{adv-spk}} + \lambda_{\text{con}}\mathcal{L}_{\text{adv-con}}. \tag{2}$$

### 3.3 抑郁症条件流匹配文本转语音（DepFlow）

DepFlow 基于 Matcha-TTS 架构（Mehta 等人） [2024] 进行改造，以支持由 DAE 得出的抑郁嵌入所条件化的波形级合成。该模型由文本编码器、时长预测器和条件流匹配（CFM）解码器组成，旨在满足三个关键要求：高保真合成、全局连贯性和精确可控性（Lipman 等人 [2022], ；Zhang 等人 [2024]）。

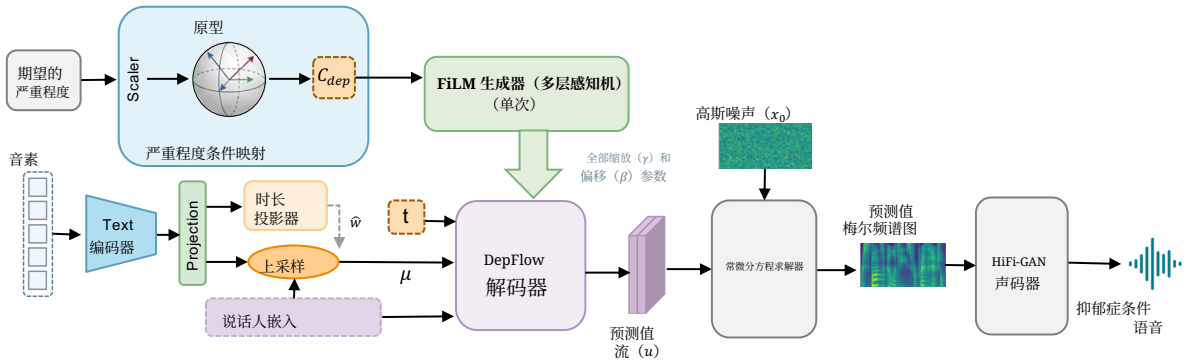



图 3：**DepFlow 的推理流程**。给定所需的严重程度，DepFlow 将其映射为抑郁条件嵌入 $\mathbf{c}_{\text{dep}}$，并结合音素和说话人输入，通过 FiLM 条件解码生成受严重程度控制的梅尔频谱图，最终使用 HiFi-GAN 将其转换为语音。

在合成过程中，CFM 解码器接收三个输入，即音素序列（文本）、说话人嵌入和由 DAE 得出的抑郁嵌入。抑郁嵌入不会改变语言或说话人表征，而是作为外部风格变量，用于塑造发声能量、韵律和发音。

### 3.3.1 基于 FiLM 的抑郁症条件化

解耦语音生成的一项关键要求是全局一致性：抑郁线索必须影响所有声学尺度，从帧级细节到长程韵律。为满足这一要求，DepFlow 采用基于 FiLM 的调制机制 Perez 等人 [2018]，而非特征拼接或注意力机制。

拼接提供的条件控制较弱，往往无法影响更深层的解码器层，而交叉注意力则会带来不必要的计算开销。相比之下，FiLM 对中间激活应用逐通道仿射变换，既高效，又能在整个解码层级中产生均匀影响。

给定一个抑郁控制嵌入 $\mathbf{c}_{\text{dep}}$，轻量级多层感知机会为所有 FiLM 调制模块生成缩放-平移对（$\gamma_i$，$\beta_i$）。每个解码器模块都会应用逐通道仿射调制：

$$\hat{\mathbf{h}}_i = \gamma_i \cdot \mathbf{h}_i + \beta_i. \tag{3}$$

其中，$\mathbf{h} \in \mathbb{R}^{B\times C\times T}$ 表示中间特征图。

FiLM 作为一种全局条件控制机制，通过缩放和平移特征图，有效调节语音整体的能量和节奏，同时不改变内容特征所编码的底层语言结构。

### 3.3.2 DepFlow 解码器

DepFlow 解码器采用 Matcha-TTS Mehta 等人 [2024], 的 U-Net 架构，包含下采样、瓶颈层和上采样阶段。每个阶段都结合残差网络和轻量级 Transformer 层，以捕捉短期和长期声学依赖关系。流匹配训练解码器预测速度场，将噪声转换为梅尔频谱图，从而实现稳定且可解释的连续时间生成过程。与依赖复杂弯曲去噪轨迹、且可能受到离散化敏感性影响的标准扩散模型不同，采用最优传输的流匹配能够学习平滑且确定性的传输路径。这种几何上的简洁性对我们解耦的严重程度控制目标尤为重要：它确保连续调整抑郁严重程度条件时，生成的声学轨迹能够平滑且稳定地偏移，同时保持稳定性和说话人身份，并避免非线性插值常引入的伪影。

### 3.3.3 优化目标

训练使用时长损失、高斯先验损失和流匹配损失，与 Matcha-TTS 中的做法相同：

**时长损失**

$$\mathcal{L}_{\text{dur}} = \text{MSE}(\log \mathbf{w}, \log \hat{\mathbf{w}}). \tag{4}$$

其中，$\mathbf{w}$ 和 $\hat{\mathbf{w}}$ 分别表示真实时长和预测时长。

**先验损失** A 高斯先验促使编码器输出与目标梅尔频谱图保持一致：

$$\mathcal{L}_{\text{prior}} = \frac{1}{|\mathcal{M}| \cdot n_f} \sum_{(i,j)\in\mathcal{M}} \frac{1}{2}\Big[(y_{i,j} - \mu_{i,j})^2 + \log(2\pi)\Big]. \tag{5}$$

**流匹配损失** 根据 CFM 公式，我们训练解码器来预测速度场：

$$\mathcal{L}_{\text{fm}} = \frac{1}{|\mathcal{M}| \cdot n_f} \sum_{(i,j)\in\mathcal{M}} \|\mathbf{u}_{i,j} - \hat{\mathbf{u}}_{i,j}\|^2 . \tag{6}$$

其中，目标速度为

$$\mathbf{u} = \mathbf{x}_1 - (1 - \sigma_{\min})\mathbf{z}. \tag{7}$$

**总损失**

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{dur}} + \lambda_p \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{fm}}. \tag{8}$$

其中 $\lambda_p$ 设为 1.0。

### 3.4 基于原型的严重程度条件映射

尽管 DepFlow 支持抑郁症条件生成，但有效的解耦语音合成需要一种既具有临床意义、又能够平滑控制的严重程度表示。原始 DAE嵌入虽然信息丰富，但可能存在说话人间变异或非线性几何结构，从而使直接操控变得复杂。为了获得稳定且可解释的严重程度轴，我们构建了一个原型库，用于表示每个 PHQ-8严重程度等级下的抑郁核心声学模式。在合成过程中，这些原型条件化的严重程度嵌入作为输入 DepFlow 的抑郁控制变量，从而实现对抑郁相关声学表现力的细粒度操控。

#### 3.4.1 受试者级嵌入聚合

与单个话语内部相比，抑郁严重程度在同一受试者的多段录音中表现得更加稳定。因此，我们通过对话语级 DAE嵌入进行平均，计算受试者级嵌入 $\mathbf{d}^{(\text{subj})}$。这能够降低说话人内方差，并得到与临床评估相一致的稳健抑郁描述符。

受试者按照 DAE训练期间使用的五个 PHQ-8严重程度分组进行归类。对于每个分组 $k$，我们计算一个原型：

$$\bar{\mathbf{p}}_k = \frac{1}{N_k} \sum_{j\in\mathcal{S}_k} \mathbf{d}_j^{(\text{subj})}, \quad \mathbf{p}_k = \frac{\bar{\mathbf{p}}_k}{\|\bar{\mathbf{p}}_k\|_2}, \tag{9}$$

原型 $\mathbf{p}_k$ 表示每个严重程度类别的抑郁核心声学模式，并剔除了个体特征的影响。

#### 3.4.2 通过球面线性插值实现连续严重程度控制

推理期间，目标是生成与任意所需 PHQ-8 评分相对应的抑郁症嵌入，而不仅限于五个离散区间。给定目标评分 $s$，我们将其映射为一个归一化标量：

$$\alpha(s) = \text{clip}\left(\frac{s - 12}{12}, -1, 1\right). \tag{10}$$

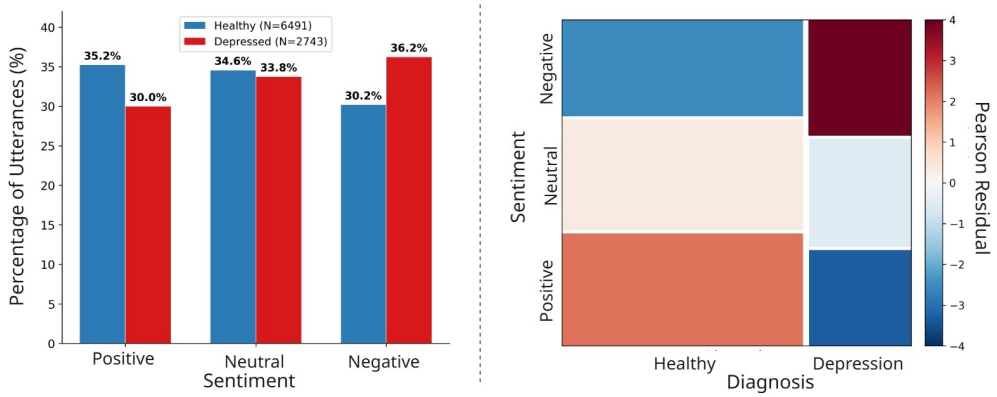
该标量表示其在严重程度谱中的相对位置。随后，我们确定相邻的两个原型 $\mathbf{p}_i$ 、$\mathbf{p}^{i+1}$，并计算插值权重 $\tau(s) \in [0, 1]$。最终的严重程度嵌入通过球面线性插值（SLERP）获得（Shoemake [1985]）：

$$\mathbf{c}_{\text{dep}}^{\text{infer}}(s) = \text{slerp}\big(\mathbf{p}_i, \mathbf{p}_{i+1}; \tau(s)\big). \tag{11}$$

球面线性插值在单位超球面上保留角度几何结构，使抑郁表达性平滑过渡，同时避免与线性插值相关的伪影。

表 1： **DAIC-WOZ数据集的人口统计学和诊断分布。**

| 性别 | 类别 | 数量 | PHQ-8评分 0-4 | 5-9 | 10-19 | 20-24 | PHQ评分统计 |
|---|---|---|---|---|---|---|---|
| 女性 | 对照组 | 56 | 38 | 18 | 0 | 0 | 3.4±3.09 |
| 女性 | 抑郁组 | 31 | 0 | 0 | 25 | 6 | 14.5±4.19 |
| Male | 对照组 | 76 | 48 | 28 | 0 | 0 | 6.8±5.92 |
| Male | 抑郁组 | 26 | 0 | 0 | 25 | 1 | 14.0±3.22 |




**(a) Sentiment Distribution by Diagnosis Group**

**(a) Mosaic Plot of Diagnosis vs. Sentiment with Pearson Residuals**

图4：DAIC-WOZ数据集中语义偏差的分析。（**a**）按诊断组划分的情感分布；（**b**）诊断与情感的马赛克图，按皮尔逊残差着色。红色方块表示过度代表（观测值 >期望值），蓝色方块表示代表不足。抑郁-负面交叉区域的深红色阴影凸显了抑郁标签与负面语言情感之间的显著相关性。

## 4 实验设置

### 4.1 数据集

我们的实验使用了两个数据集：CSTR VCTK 语料库 Yamagishi 等人 [2019] 以及 DAIC-WOZ 数据集 Gratch 等人 [2014]。CSTR VCTK 语料库包含 110 名英语说话者的录音，原始采样率为 48 kHz。DAIC-WOZ 数据集由 189 名英语参与者的临床访谈组成，其中 56 人经临床诊断患有抑郁症，音频采样率为 16 kHz。值得注意的是，根据 Bailey 和 Plumbley [2021][1] 报告的修正，DAIC-WOZ 训练分区中的 409 个样本被重新标记为 0 至 1。

在我们的流程中，CSTR VCTK 仅用于预训练文本转语音骨干网络（DepFlow基础模型），而 DAIC-WOZ 则用于（i）训练抑郁症声学编码器（DAE），（ii）通过抑郁症条件控制对 DepFlow 进行微调，以及（iii）训练和评估下游抑郁检测模型。

所有需要抑郁标签的操作均严格只使用 DAIC-WOZ 训练集和开发集，官方测试集仅保留用于最终评估。

### 4.2 情感分布分析

我们分析了 DAIC-WOZ训练集中语言情感与抑郁症诊断之间的关系。我们对 9,234 条受访者话语进行了情感分析，并根据时间戳进行说话人分离，同时排除了时长短于 1 秒的片段。我们使用 DeepSeek R1 大语言模型将每条话语分类为积极、中性或消极 [2025] [2]。我们分别计算了健康参与者和抑郁症参与者的情感分布。

[1]https://github.com/adbailey1/daic_woz_处理 [2]
https://github.com/deepseek-ai/DeepSeek-R1

图4(a) 显示两组之间存在明显的分布偏移。健康参与者说出的积极话语多于消极话语（35.2% 对 30.2%）；而抑郁症参与者则呈现相反模式，消极情感的比例高于积极情感（36.2% 对 30.0%）。

为量化这种关联，我们使用皮尔逊残差着色的马赛克图来可视化情感-诊断列联表（图4(b)）。对于每个单元格（$i$, $j$），皮尔逊残差定义为

$$r_{ij} = (O_{ij} - E_{ij})/\sqrt{E_{ij}}, \tag{12}$$

其中 $O_{ij}$ 和 $E_{ij}$ 表示 独立性假设下的观察频数和期望频数。这些模式在统计学上
具有显著性，这一点通过残差和独立性卡方检验的证实。
皮尔逊独立性卡方检验进一步证实了这种关联（$\chi^2 = 38.03$, $df = 2$, $p < 10^{-8}$）Pearson [1900]。总体而言，DAIC-WOZ 中的语言情感与诊断标签高度耦合，这表明存在一种分类模型可利用的语义捷径。

### 4.3 预处理与数据划分

在所有实验中，我们均采用 DAIC-WOZ 官方数据划分（训练集：107 名受试者，开发集：35 名受试者，测试集：47 名受试者），确保不同划分之间的受试者互不重叠。

对于 DAE 训练和下游抑郁症检测实验，我们均遵循 Speechformer Chen 等人[2022b] 的采样方案，在话语级进行类别平衡。训练期间，每个训练轮次都会从选定的每条话语中随机裁剪一段 10 秒的片段，以引入时间变化。评估时，使用每位测试受试者最长的 20 条话语，并通过对话语级输出进行多数投票，得到受试者级预测结果。

对于在 VCTK 上进行的 DepFlow 预训练，我们使用数据集提供的原始语料分段，并将音频从 48 kHz 下采样至 22.05 kHz，以匹配我们的声码器和声学特征处理流程。

对于在 DAIC-WOZ 上进行的 DepFlow 微调，我们使用训练集和开发集中的所有话语，但会移除时长短于一秒或以静音为主的片段。这些录音从 16 kHz 重采样至 22.05 kHz。训练或生成过程中不会使用 DAIC-WOZ 测试集中的任何信息。

所有由 DepFlow 生成的
扩充训练数
语音样本均严格仅使用 DAIC-WOZ训练集生成，且仅用于
据。绝不会将这些样本加入开发集或测试集。

### 4.4 实现细节

所有模型均在 NVIDIA
L40 GPU 上训
练。生成框架的训练分为三个阶段。

**阶段 1：抑郁症声学编码器** DAE 使用 AdamW Loshchilov 和 Hutter [2017] 进行优化，最多训练 500 轮次，批量大小为 64（学习率 $1 \times 10^{-4}$，权重衰减 $3 \times 10^{-3}$）。所有投影层和预测层均采用 0.2 的随机失活率。损失权重设置为 $\lambda_{\text{sup}} = 1.0$、$\lambda_{\text{id}} = 0.2$、$\lambda_{\text{spk}} = 0.2$ 和 $\lambda_{\text{con}} = 0.1$。根据 DAIC-WOZ 开发集上的表现采用提前停止。

**阶段 2 和 3：DepFlow 的预训练与微调** 首先，我们在 VCTK 上对基础 Matcha-TTS 模型进行预训练，不使用抑郁症条件控制，并遵循 Matcha-TTS 中的默认参数设置 Mehta 等人 [2024]，除非另有说明。随后，使用相同的超参数设置，在带有抑郁症条件和 FiLM 生成器的情况下，基于 DAIC-WOZ 对 DepFlow 进行微调。抑郁症条件控制嵌入的维度为 32。微调期间，同时更新基础文本转语音模型参数和 FiLM 生成器。FiLM 生成器的随机失活率设为 0.2。推理时，使用预训练的 HiFi-GAN 声码器 Kong 等人 [2020]，根据梅尔频谱图重建波形。

### 4.5 评估协议

#### 4.5.1 合成语音的可控性和有效性

DepFlow 的一个主要目标是在固定语言内容和说话人身份的同时，以可控的方式调节抑郁相关声学表现力。因此，我们从三个互补的角度评估其可控性和有效性。首先，我们评估提高目标严重程度是否会在学习到的抑郁症嵌入流形中产生单调偏移。其次，我们考察这些潜在空间偏移是否会转化为说话人内部具有系统性的变化

，并体现为临床相关的声学标志。第三，我们验证严重程度操控是否会损害文本转语音质量，包括语言内容的可懂度和说话人相似度。

本节中的所有分析均采用统一的严重程度控制合成方案，并应用于 DAIC-WOZ 测试集。对于每个话语，生成五种合成变体，分别对应从健康到重度抑郁症的五个临床严重程度等级。对于每个目标等级，PHQ 得出的严重程度值会转换为连续标量 $\alpha! \in![-1, 1]$，随后映射为抑郁条件嵌入$\mathbf{c}^{\text{dep}}$；该嵌入与目标说话人身份共同作为条件输入 DepFlow，以生成语音。

该方案为每个话语生成五种严重程度受控的语音版本，共得到 4,700 个合成测试样本。该方案仅用于评估；训练模型时不包含任何合成样本，从而避免下游抑郁症检测实验发生任何形式的数据泄漏。

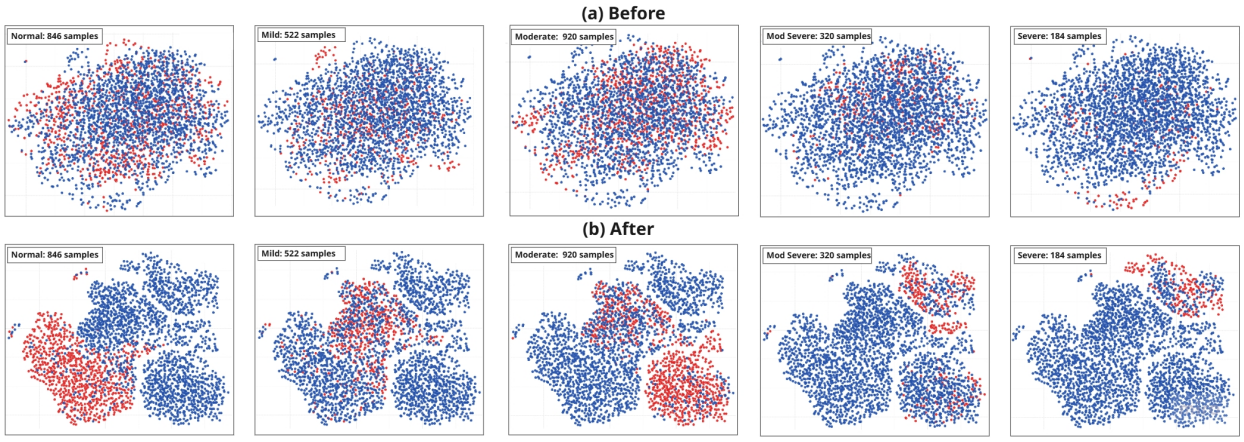



图5：DAE嵌入的 UMAP 可视化。(a) 从冻结的 WavLM-Large 模型中提取的嵌入表示；(b) 相同话语经训练后的 DAE 处理所得的嵌入表示。每个面板中的五个子图分别对应五个临床严重程度等级（健康、轻度、中度、中重度、重度）：红点表示来自目标严重程度等级的样本，蓝点表示其余所有样本。

#### 4.5.2 伪装型抑郁症导向增强（CDoA）数据集的构建

为评估语义-声学线索不一致的影响，我们通过结构化的三步流程生成 CDoA 数据集。

首先，我们构建一个基于情感的文本库。我们使用大语言模型 DeepSeek R1 Guo 等人 [2025]，对所有 DAIC-WOZ训练转录文本进行话语级标注（积极、中性、消极）。积极和中性话语归入良性文本库，而消极话语则构成抑郁文本库，从而能够明确控制语义极性。

遵循基线采样策略，我们确定了合成配额，以确保全局二元类别平衡，并覆盖所有患者健康问卷严重程度等级。最终数据集结合了原始数据与合成数据，共包含 5,760 个话语，实现了完全平衡的二元分布（2,880 个抑郁样本与 2,880 个非抑郁样本）。具体做法是针对每个严重程度组，为每位受试者合成固定数量的话语：13（正常）、34（轻度）、2（中度）、91（中重度）和 194（重度）。

最后，我们将每位受试者的患者健康问卷得分映射为连续标量 $\alpha \in [-1, 1]$，并将其投影到抑郁条件嵌入 $\mathbf{c}_{\text{dep}}$ 中，从而生成解耦样本。该嵌入为 DepFlow 提供条件信息，使其将抑郁声学特征注入良性（积极或中性）文本中，构造出受控的声学—语义不匹配：抑郁语音模式与非抑郁语言内容共存。

**基线抑郁症检测模型**所提出的增强方法在三个具有代表性的语音型抑郁检测模型家族上进行了基准测试，它们在输入特征和骨干架构方面各不相同：

- **DepAudioNet** Ma 等人 [2016]：一种基于对数梅尔频谱图的卷积循环模型。DepAudioNet 以对数梅尔频谱图为输入，使用多个卷积层，随后接入长短期记忆层来编码时间动态，并通过全连接层进行话语级抑郁症分类。

• **NUSD** Wang 等人 [2023]：一种基于帧级声学特征的说话人解耦时延神经网络模型。NUSD 将对数梅尔频谱图和基频特征作为输入，送入 ECAPA-TDNNDesplanques 等人 [2020] 骨干网络，并连接非均匀的说话人解耦分支，同时由主池化层和分类头预测话语级抑郁标签。

• **HAREN-CTC** Li 等 [2025b]：一种基于原始波形构建的自监督分层模型。HAREN-CTC 对从原始音频中提取的多层 WavLM 表征进行处理，并采用带有 CTC 分支和基于池化的分类头的分层编码器，实现受试者级抑郁症检测。

对于所有三种基线方法，我们都尽可能严格遵循原始实现和超参数设置，并采用相同的话语筛选和数据增强方案，以确保公平比较。

### 4.6 评估指标

我们从四个方面评估所提出的框架：（i）DAE 学到的抑郁条件嵌入的解耦效果；（ii）合成抑郁严重程度的可控性；（iii）DepFlow 文本转语音（TTS）的质量；（iv）合成 CDoA 数据集对下游抑郁症检测性能的影响。

**说话人和内容解耦指标** 为验证 DAE 学到的抑郁嵌入能够捕捉与严重程度相关的结构，同时抑制干扰属性，我们采用了说话人导向和内容导向指标。

说话人身份保留采用等错误率（EER）Hansen 和 Hasan [2015], 进行评估；这是一项标准的说话人验证指标：

$$\mathrm{EER} = \frac{\mathrm{FAR}(\tau^*) + \mathrm{FRR}(\tau^*)}{2}, \tag{13}$$

其中，FAR 和 FRR 分别表示误接受率和误拒绝率，$\tau^*$ 表示二者相交处的决策阈值。此外，还计算相似度差距，其定义为同一说话人配对与不同说话人配对之间平均余弦相似度的差值。

为评估残留的语言信息，我们训练了一个线性探针，以抑郁嵌入为输入，对池化后的内容表征（HuBERT特征 Hsu 等人 [2021]）进行回归。我们报告均方误差（MSE）Draper [1998],、决定系数（$R^2$）Draper [1998], 以及预测特征与真实特征之间的中心化核对齐（CKA）Kornblith 等人 [2019]。

**DepFlow文本转语音质量** 语言可懂度采用词错误率（WER）衡量，说话人一致性则采用说话人相似度指标 SIM-o Le 等人 [2023]评估。WER 使用 Whisper-Large-v3 Radford 等人 [2023], 计算，遵循 Anastassiou 等人 [2024]的评估方案。SIM-o 通过提取 WavLM-Large嵌入表示 Chen 等人 [2022c]，并计算合成话语与对应自然话语之间的余弦相似度得出。

**潜在空间中的严重程度可控性** 为了评估预期严重程度的提高是否会在学习到的抑郁流形中产生单调且全局一致的偏移，我们采用基于排序的指标。对于在五个临床严重程度等级下合成的每个话语组，我们计算：(i) 一致性指数（C-index）Harrell 等人 [1982], , 用于衡量预期严重程度较高的样本获得较高嵌入投影值的概率；以及 (ii) 斯皮尔曼秩相关系数（$\rho$）Spearman [1987], , 用于评估严重程度等级与嵌入投影之间的序数一致性。这两项指标侧重于单调排序，而非绝对回归准确率，因此适用于验证连续严重程度控制。

**用于提升抑郁症检测性能的数据增强数据集** 对于使用 DepFlow 数据增强训练和未使用 DepFlow 数据增强训练的下游抑郁症检测模型，我们报告临床场景下的受试者级灵敏度、特异度和宏平均 F1 Sokolova 和 Lapalme [2009]。受试者级预测通过对话语级决策进行多数投票得出。

## 5 实验结果与讨论

本节从四个互补的角度评估所提出的框架：(1) 学得的抑郁流形的质量与结构；(2) DepFlow合成语音的可控性与声学有效性；(3) DepFlow生成的不一致样本作为下游检测器数据增强手段的有效性；(4) 定性分析受控的严重程度操控如何塑造声学轨迹。

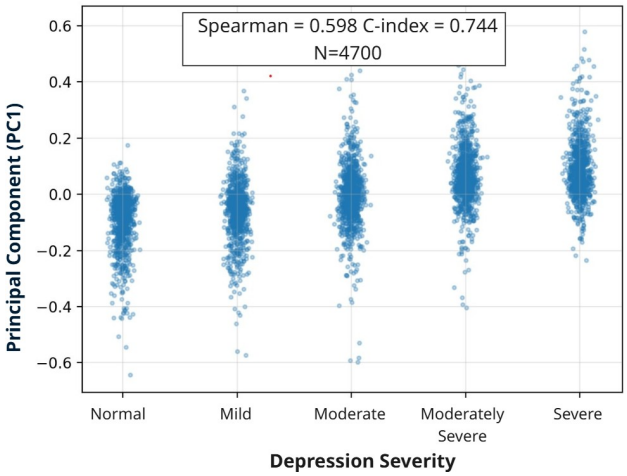



图 6：预期严重程度等级与潜在抑郁嵌入轨迹之间的对齐。

表 2：DepFlow合成语音的说话人内声学可控性（N = 4,700 个样本）。

| 特征 | 均值 $\rho$ | 中位数 $\rho$) |
|---|---|---|
| 共振峰F1均值 | 0.635 | 0.800 |
| 共振峰F2均值 | 0.604 | 0.800 |
| 静音–语音比 | 0.579 | 0.866 |
| 共振峰集中化 | -0.489 | -0.600 |
| 微闪 | 0.304 | 0.500 |
| HNR | -0.259 | -0.500 |
| 抖动 | -0.068 | -0.100 |
| $F_0$ 均值 | 0.064 | 0.100 |
| $F_0$ std | 0.058 | 0.100 |

### 5.1 抑郁条件嵌入的结构与序数性

我们评估了所提出的抑郁声学编码器（DAE）是否能够学习具有结构性且具临床意义的抑郁表征。如图5所示，该嵌入空间呈现出相对紧凑且连贯的连续体，从健康状态逐渐过渡到更重度的抑郁状态；相比之下，冻结的 WavLM 嵌入表示出较弱的聚类特性和有限的序数组织。这些观察结果表明，DAE 倾向于沿着以严重程度为导向的流形组织语音信号，而不仅仅是将样本归入离散的诊断类别。

表 III 证实了这些定性观察结果。较高的说话人验证错误率（EER）和较低的语言相关性分数（低 $R^2$ 和 CKA）表明，干扰属性已得到成功解耦。关键的是，该嵌入仍保留了诊断效用，在线性探测中取得了稳健的 ROC-AUC（0.693），证明即使说话人和内容信息受到抑制，严重程度信息仍得以保留。

总体而言，这些结果表明，DAE 构建了一个序数抑郁流形，同时在实证上削弱了与说话人和内容相关的成分，从而为下游解耦语音生成提供了一个实用且具有临床相关性的条件空间。

### 5.2 合成语音的可控性与有效性

本节所有分析均基于按照第 4.5.1 节所述流程生成的严重程度控制合成测试集。

#### 5.2.1 嵌入空间可控性

接下来，我们考察随着严重程度增加，DepFlow 是否会在所学习的抑郁嵌入空间中产生系统且有序的变化。对于每个合成的话语，我们从预训练的 DAE 中提取抑郁条件嵌入 $\mathbf{z}_{u,\ s}$

Tabl图 3：所提出的 DAIC-WOZ 上 DAE 嵌入解耦特性的评估 .

| Type | 评估指标 | 结果 |
|---|---|---|
| 说话人解耦 | EER | 0.355 |
| | 相似度差距（正例-负例） | 0.27 |
| 内容解耦 | MSE | 2.83 |
| | $R^2$ | 0.21 |
| | CKA相似度 | 0.014 |
| 抑郁症分类 | 准确率 | 0.638 |
| | 宏平均 F1 | 0.591 |
| | ROC-AUC | 0.693 |

并对所有合成嵌入表示应用主成分分析（PCA）。随后，将第一个主成分用作共享的严重程度坐标，并使用一致性指数和斯皮尔曼秩相关评估其与预设的五个临床严重程度等级之间的单调性。

如图6所示，投影后的嵌入表示分布从健康到重度抑郁症逐渐向较高值偏移。定量而言，我们得到的一致性指数为0.744，Spearman相关系数为0.598，表明在不同说话人和语言内容中，预期严重程度与合成语音的潜在轨迹之间存在强而一致的有序关系。这表明，DepFlow并未引入任意变化，而是在DAE学习到的抑郁流形中沿着连贯的方向驱动生成。

表4：受试者级 在DAIC-WOZ数据集上采用不同数据增强策略时的抑郁症检测性能
数据集。结果以 宏平均 F1 / 灵敏度 / 特异度报告。粗体表示各模型的最佳结果。

| 模型 | 增强方法 | 宏平均 F1 | 灵敏度 | 特异度 |
|---|---|---|---|---|
| DepAudioNet Ma 等人 [2016] | / | 0.482 (0.042) | 0.346 (0.252) | 0.667 (0.194) |
| | FrAUG | 0.518 (0.047) | 0.629 (0.137) | 0.495 (0.593) |
| | Mixup | 0.489 (0.056) | **0.657 (0.078)** | 0.418 (0.110) |
| | Instruct-TTS (CDoA) | 0.459 (0.064) | 0.457 (0.229) | 0.539 (0.258) |
| | DepFlow (CDoA)（我们的方法） | **0.526 (0.023)** | 0.417 (0.157) | **0.674 (0.159)** |
| NUSD Wang 等人 [2023] | / | 0.514 (0.046) | 0.300 (0.137) | **0.745 (0.102)** |
| | Mixup | 0.486 (0.052) | 0.357 (0.134) | 0.636 (0.156) |
| | Instruct-TTS (CDoA) | 0.485 (0.038) | 0.400 (0.193) | 0.606 (0.183) |
| | DepFlow (CDoA)（本文方法） | **0.577 (0.035)** | **0.471 (0.081)** | 0.697 (0.057) |
| *HAREN*-CTC Li 等 [2025b] | / | 0.525 (0.035) | 0.386 (0.081) | 0.673 (0.062) |
| | SpecAugment | 0.477 (0.067) | 0.386 (0.041) | 0.600 (0.114) |
| | Mixup | 0.526 (0.070) | 0.375 (0.090) | **0.712 (0.072)** |
| | Instruct-TTS (CDoA) | 0.510 (0.033) | **0.429 (0.033)** | 0.606 (0.033) |
| | DepFlow (CDoA)（本文方法） | **0.551 (0.028)** | 0.400 (0.081) | 0.709 (0.090) |

### 5.2.2 说话人内声学可控性

我们进一步考察低层次声学标记层面的可控性。对于每个合成样本，我们提取一组临床相关特征，并计算五种合成变体中预期严重程度等级与这些特征之间的说话人内斯皮尔曼相关。

表 3 总结了结果。若干临床认可的抑郁症相关指标呈现出明显且一致的单调趋势：

- 静音–语音比呈现最强的正相关，这与抑郁症患者言语活动减少和精神运动迟滞的相关报告一致。
- 微闪和谐噪比等嗓音质量指标的变化方向符合预期，这表明随着严重程度增加，发声会变得愈发不稳定、气声也愈发明显。
- 共振峰特征（第一共振峰、第二共振峰及中央化）会随严重程度发生系统性偏移，反映了抑郁言语中常见的元音中央化和构音不足。

表5：自然语音和DepFlow合成语音的TTS质量客观评估。对于自然DAIC-WOZ语音，由于SIM-o按定义表示相似度上限（100%），因此未报告该指标。

| 数据类型 | 严重程度 | 词错误率（%）↓ | SIM-o↑ |
|---|---|---|---|
| DAIC-WOZ | 健康 | 13.76 (0.182) | / |
| | Mild | 14.77 (0.176) | / |
| | 中度 | 15.20 (0.175) | / |
| | 中重度 | 11.41 (0.193) | / |
| | 重度 | 15.15 (0.208) | / |
| | **平均值** | **14.06 (0.150)** | / |
| 合成 | 健康 | 13.94 (0.225) | 57.13 (0.095) |
| | Mild | 14.03 (0.228) | 57.07 (0.095) |
| | 中度 | 13.81 (0.227) | 57.03 (0.095) |
| | 中重度 | 13.93 (0.224) | 56.88 (0.094) |
| | 重度 | 13.92 (0.230) | 56.73 (0.094) |
| | **平均值** | **13.93 (0.227)** | **56.97** (0.095) |

这些发现表明，DepFlow 并非只是生成感知上有所不同的变体，而是沿着具有临床意义且可通过声学特征解读的维度调节语音。值得注意的是， $F_0$ 统计结果显示相关性较弱，这表明 DepFlow 主要控制频谱—构音特征和音质线索，而非仅依赖音高操控。结合嵌入层面分析，这些结果证明 DepFlow 能够对抑郁严重程度进行平滑、有结构且有临床依据的控制。

### 5.2.3 文本转语音质量和说话人相似度

接下来，我们评估严重程度操控是否会损害语言可懂度或说话人身份。表5报告了自然 DAIC-WOZ 话语和 DepFlow 合成语音在五个严重程度等级下的词错误率（WER）和说话人相似度（SIM-o）结果。

如表5所示，合成语音的词错误率与自然录音非常接近（13.93% 对比 14.06%），这证实了尽管对副语言声学特征进行了大幅操控，DepFlow 仍能保留语言可懂度。

不同严重程度等级下的说话人相似度也保持稳定，这表明 DepFlow 主要改变表达和韵律维度，而不会扭曲音色身份。换言之，施加的抑郁线索改变的是受试者说话的方式，而不是说话者是谁。

总之，这些结果表明，DepFlow 能够以可控的方式注入与抑郁症相关的声学模式，同时保持语音质量和说话人特征，从而避免可能干扰下游抑郁症检测模型的伪影。

## 5.3 CDoA数据集对下游抑郁症检测的有效性

接下来，我们研究 DepFlow 生成的 CDoA样本是否能提升下游检测性能。这些实验旨在评估，接触语义不一致样本是否会促使模型形成更加均衡且稳健的决策边界。

### 5.3.1 Co与其他数据增强方法的比较 方法

表4展示了在受试者层面与传统增强策略（FrAUG Ravi 等人 [2022b], SpecAugment Park 等 [2019], Mixup Zhang 等人 [2017]）及 Instruct-TTS 基线方法的性能比较。DepFlow 在三种不同架构上始终优于所有基线方法。值得注意的是，与基于 CosyVoice-2 的 Du 等人 [2024]语音克隆 TTS 基线相比，我们的方法表现更优。该基线同样生成伪装导向的增强数据，但依赖通用语音克隆，未学习抑郁感知声学流形。

这一比较验证了我们提出的框架：通用文本转语音虽然能够生成多样的韵律，却无法合成具有临床意义的抑郁标记。显著的性能差距证实，鲁棒性提升具体源于有针对性地注入抑郁相关声学线索，而非单纯的数据扩充。

## 6 结论

总之，DepFlow 提供了一种波形级的抑郁条件语音生成框架，通过将抑郁症声学线索与语言情感解耦，同时保留语音质量和说话人身份，构建伪装抑郁症导向增强数据集（CDoA数据集）。现有语料库中很少包含这些语义不一致但具有临床意义的条件，而通过有针对性的数据增强来提升下游抑郁症检测模型的鲁棒性，这些条件至关重要。

实证结果表明，DepFlow能够生成连贯的严重程度流形，实现对抑郁症声学线索的平滑且可解释控制，同时保留语音可懂度和说话人身份。作为一种增强来源时，DepFlow生成的CDoA数据集能够持续提升多个下游抑郁症检测架构的性能。尽管这些结果尚不足以构成因果证明，但在灵敏度、特异度、宏平均 F1和决策边界稳定性方面观察到的提升提供了相互印证的证据，表明受控的声学—语义不匹配能够减少模型对语言捷径的依赖，并促使模型关注与抑郁症相关的声学生物标志物。因此，DepFlow为利用可控生成建模来正则化抑郁症检测器并增强鲁棒性提供了一种有原则的方法。

仍存在若干局限性。尽管所学习的严重程度轴在内部具有一致性，并且与已知的抑郁症声学相关因素相吻合，但尚未通过临床医生评分进行验证，也尚未在更广泛的人群和数据集上得到验证。从伦理角度看，合成具有抑郁症声学特征的语音可能带来潜在风险，包括被滥用于伪造、无意中强化关于抑郁症的文化刻板印象，以及造成心理伤害。因此，负责任的使用需要透明的治理、知情同意、对敏感语音数据的谨慎处理，并考虑采取来源追踪或水印等保障措施。未来的工作还将探讨对合成线索开展感知研究，将方法扩展至多语种和人口统计特征多样的说话人，并研究更广泛的部署框架，以在基于语音的心理健康评估中平衡鲁棒性提升与可靠性、可解释性及社会影响。

## 参考文献

Nicholas Cummins、Stefan Scherer、Jarek Krajewski、Sebastian Schnieder、Julien Epps 和 Thomas F Quatieri。基于语音分析的抑郁症与自杀风险评估综述。*Speech communication*，71:10–49，2015年。Lu-Shih Alex Low、Namunu C Maddage、Margaret Lech、Lisa B Sheeber 和 Nicholas B Allen。家庭互动期间青少年语音中临床抑郁症的检测。*IEEE*生物医学工程汇刊，58(3): 574–586，2010年。S Bandyopadhyay Prasanta、R Forster Malcolm、Elsevier Oxford、H Barkow Jerome、Cos mides Leda、Tooby John、Bechtel William、C Richardson Robert、T Beck Aaron、Rush A John 等。美国精神病学会，《精神障碍诊断与统计手册》：DSM-5，华盛顿特区，American Psychiatric Publishing，2013年。Ananth Mahesh，《为健康的进化概念辩护：自然、规范与人类生物学》，英格兰奥尔德肖特，Ashgate。*Philosophy*，39(6):683–724，2014年。Laurie Kelly McCorry。自主神经系统的生理学。*American Journal of Pharmaceutical Education*，71(4):78，2007年。Shaykhah A Almaghrabi、Scott R Clark 和 Mathias Baumert。抑郁症的生物声学特征：综述。*Biomedical Signal Processing and Control*，85:105020，2023年。Carmen Morales-Luque、Laura Carrillo-Franco、Manuel Víctor López-González、Marta González-García 和 Marc Ste- fan Dawid-Milner。描绘声音与自主神经功能之间的神经生理学联系：范围综述。*Biology*，14(10):1382，2025年。Seth Brown。伪装抑郁症。*Discover Mental Health*，5(1):71，2025年。Klaus R Scherer。情绪的声音交流：研究范式综述。*Speechcommunication*，40(1-2): 227–256，2003年。Robert Geirhos、Jörn-Henrik Jacobsen、Claudio Michaelis、Richard Zemel、Wieland Brendel、Matthias Bethge 和 Felix A Wichmann。深度神经网络中的捷径学习。*NatureMachine Intelligence*，2(11):665–673，2020年。Judea Pearl。因果关系。剑桥大学出版社，2009年。Bernhard Schölkopf、Francesco Locatello、Stefan Bauer、Nan Rosemary Ke、Nal Kalchbrenner、Anirudh Goyal 和 Yoshua Bengio。迈向因果表征学习。*Proceedings of the IEEE*，109(5):612–634，2021年。Lennart Jansson、Julie Nordgaard 等。用于鉴别诊断的精神科访谈，第270卷。Springer，2016。

Prasad Shetty、Akshata Mane、Sourabh Fulmali 和 Ganesh Uchit。理解掩饰性抑郁：临床案例。《印度精神病学杂志》，60(1):97–102，2018年。Dong Zhang、Shimin Li、Xin Zhang、Jun Zhan、Pengyu Wang、Yaqian Zhou 和 Xipeng Qiu。Speechgpt：赋予大语言模型内在的跨模态对话能力。*arXiv*预印本 *arXiv:2305.11000*，2023年。Shengpeng Ji、Yifu Chen、Minghui Fang、Jialong Zuo、Jingyu Lu、Hanting Wang、Ziyue Jiang、Long Zhou、Shujie Liu、Xize Cheng 等。Wavchat：口语对话模型综述。*arXiv*预印本 *arXiv:2411.13577*，2024。Yuxin Li、Sinchana Kumbale、Yanru Chen、Tanmay Surana、Eng Siong Chng 和 Cuntai Guan。基于文本和音频的自动抑郁症检测：系统综述。*IEEE*生物医学与健康信息学杂志，2025a。Michel Valstar、Björn Schuller、Kirsty Smith、Florian Eyben、Bihan Jiang、Sanjay Bilakhia、Sebastian Schnieder、Roddy Cowie 和 Maja Pantic。Avec 2013：持续性视听情绪与抑郁识别挑战赛。载于第*3*届*ACM*国际视听情绪挑战赛研讨会论文集，第3–10页，2013年。Fabien Ringeval、Björn Schuller、Michel Valstar、Nicholas Cummins、Roddy Cowie、Leili Tavabi、Maximilian Schmitt、Sina Alisamir、Shahin Amiriparian、Eva-Maria Messner 等。Avec 2019研讨会与挑战赛：心理状态、利用人工智能检测抑郁症及跨文化情感识别。载于第*9*届国际视听情绪挑战赛与研讨会论文集，第3–12页，2019年。Karol Chlasta、Krzysztof Wołk 和 Izabela Krejtz。利用深度卷积神经网络自动语音筛查抑郁症。《计算机科学程序汇刊》，164:618–628，2019年。S Pavankumar Dubagunta、Bogdan Vlasenko 和 Mathew Magimai Doss。学习与声源相关的信息以检测抑郁症。载于*ICASSP 2019-2019 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第6525–6529页。IEEE，2019年。Afef Saidi、Slim Ben Othman 和 Slim Ben Saoud。用于高效抑郁症检测系统的混合CNN-SVM分类器。载于*2020*年第*4*届国际先进系统与新兴技术会议（*IC_ASET*），第229–234页。IEEE，2020年。Alice Othmani、Daoud Kadoch、Kamil Bentounes、Emna Rejaibi、Romain Alfred 和 Abdenour Hadid。迈向用于语音情感与抑郁识别的鲁棒深度神经网络。载于《模式识别》。*ICPR*国际研讨会与挑战赛：虚拟活动，*2021*年*1*月*10–15*日，论文集，第二部分，第5–19页。Springer，2021年。Adrián Vázquez-Romero 和 Ascensión Gallardo-Antolín。利用集成卷积神经网络自动检测语音中的抑郁症。《熵》，22(6):688，2020年。Dong Wang、Yanhui Ding、Qing Zhao、Peilin Yang、Shuping Tan 和 Ya Li。基于ECAPA-TDNN的临床语音抑郁症检测。载于*Interspeech*，第3333–3337页，2022年。Asif Salekin、Jeremy W Eberle、Jeffrey J Glenn、Bethany A Teachman 和 John A Stankovic。用于检测社交焦虑与抑郁症的弱监督学习框架。《*ACM*交互、移动、可穿戴与泛在技术汇刊》，2(2):1–26，2018年。Ziping Zhao、Zhongtian Bao、Zixing Zhang、Nicholas Cummins、Haishuai Wang 和 Björn Schuller。用于语音抑郁症评估的层次化注意力迁移网络。载于*ICASSP 2020-2020 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第7159–7163页。IEEE，2020年。Muhammad Muzammel、Hanan Salam、Yann Hoffmann、Mohamed Chetouani 和 Alice Othmani。Audvowelconsnet：基于音素级的临床抑郁症诊断深度CNN架构。《机器学习与应用》，2:100005，2020年。Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：用于自监督语音表征学习的框架。《神经信息处理系统进展》，33:12449–12460，2020年。Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman 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》，29:3451–3460，2021年。Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao 等。WavLM：用于全栈语音处理的大规模自监督预训练。*IEEE*《信号处理精选主题杂志》，16(6):1505–1518，2022a。Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别。载于国际机器学习会议，第28492–28518页。PMLR，2023年。Pingyue Zhang、Mengyue Wu、Heinrich Dinkel 和 Kai Yu。Depa：用于抑郁症检测的自监督音频嵌入。载于论文集

*gs of* 第 *29* 届 *ACM* 国际会议 *f*多媒体， pag第135–143页，2021年。

Ermal Toto、ML Tlachac 和 Elke A Rundensteiner。Audibert：用于抑郁症筛查的深度迁移学习多模态分类框架。载于 第*30*届*ACM*信息与知识管理国际会议论文集，第4145–4154页，2021年。Weidong Chen、Xiaofen Xing、Xiangmin Xu、Jianxin Pang 和 Lan Du。Speechformer：融合语音特征的分层高效框架。*arXiv*预印本 *arXiv:2203.03812*，2022b。Vijay Ravi、Jinhan Wang、Jonathan Flint 和 Abeer Alwan。通过说话人解耦，在检测抑郁症的同时保留说话人身份的一步探索。载于 *Interspeech*，第2022卷，第3338页，2022a。Wen Wu、Chao Zhang 和 Philip C Woodland。基于语音的抑郁症检测中的自监督表征。载于 *ICASSP 2023-2023 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第1–5页。IEEE，2023年。Yuxin Li、Eng Siong Chng 和 Cuntai Guan。用于从语音中进行抑郁症检测的分层自监督表征学习。*arXiv*预印本 *arXiv:2510.08593*，2025b。Xiangyu Zhang、Hexin Liu、Qiquan Zhang、Beena Ahmed 和 Julien Epps。Speecht-rag：利用语音时序信息，通过检索增强生成实现大语言模型中的可靠抑郁症检测。*arXiv*预印本 *arXiv:2502.10950*，2025年。ZUO Lishi 和 Man-Wai Mak。面向数据稀缺条件下基于语音的抑郁症检测的基于向量量化的反事实增强。*IEEE*生物医学与健康信息学杂志，2025年。Vijay Ravi、Jinhan Wang、Jonathan Flint 和 Abeer Alwan。FrAUG：一种基于帧率的语音信号抑郁症检测数据增强方法。载于 *ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第6267–6271页。IEEE，2022b。Daniel S Park、William Chan、Yu Zhang、Chung-Cheng Chiu、Barret Zoph、Ekin D Cubuk 和 Quoc V Le。SpecAugment：一种用于自动语音识别的简单数据增强方法。*arXiv*预印本 *arXiv:1904.08779*，2019。Kexin Feng 和 Theodora Chaspari。利用数据增强，从语音中进行抑郁症分类的知识驱动型基于元音的方法。载于 *ICASSP 2023-2023 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第1–5页。IEEE，2023年。Lokesh Kumar、Kumar Kaustubh、Shashaank Aswatha Mattur 和 SR Mahadeva Prasanna。使用对数梅尔频谱图进行抑郁症分类：基于窗口大小的数据增强与深度学习模型的比较分析。载于 *2024*年第*27*届*Oriental COCOSDA*语音数据库及评估技术协调与标准化国际委员会会议（*O-COCOSDA*），第1–6页。IEEE，2024。Ian Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。生成对抗网络。《*ACM*通讯》，63(11):139–144，2020年。Le Yang、Dongmei Jiang 和 Hichem Sahli。用于改进从语音信号中进行抑郁严重程度估计的特征增强网络。*IEEE Access*，8:24033–24045，2020年。Wenrui Liang、Rong Zhang、Xuezhen Zhang、Ying Ma 和 Wei-Qiang Zhang。DepressGEN：用于抑郁症检测的合成数据生成框架。载于 *Interspeech 2025*论文集，第464–468页，2025年。Kurt Kroenke、Tara W Strine、Robert L Spitzer、Janet BW Williams、Joyce T Berry 和 Ali H Mokdad。PHQ-8作为普通人群当前抑郁症状的测量工具。《情感障碍杂志》，114(1-3):163–173，2009年。Yaroslav Ganin、Evgeniya Ustinova、Hana Ajakan、Pascal Germain、Hugo Larochelle、François Laviolette、Mario March 和 Victor Lempitsky。神经网络的域对抗训练。《机器学习研究杂志》，17(59):1–35，2016。Shivam Mehta、Ruibo Tu、Jonas Beskow、Éva Székely 和 Gustav Eje Henter。Matcha-TTS：一种具有条件流匹配的快速文本转语音架构。载于 *ICASSP 2024-2024 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第11341–11345页。IEEE，2024。Yaron Lipman、Ricky TQ Chen、Heli Ben-Hamu、Maximilian Nickel 和 Matt Le。用于生成建模的流匹配。*arXiv*预印本 *arXiv:2210.02747*，2022。Xiangyu Zhang、Daijiao Liu、Hexin Liu、Qiquan Zhang、Hanyu Meng、Leibny Paola Garcia Perera、EngSiong Chng 和 Lina Yao。在小波域中说话：一种加速语音扩散模型的简单高效方法。载于*2024*年自然语言处理实证方法会议论文集，第159–171页，2024年。Ethan Perez、Florian Strub、Harm De Vries、Vincent Dumoulin 和 Aaron Courville。FiLM：使用通用条件层进行视觉推理。载于 *AAAI*人工智能大会论文集，第32卷，2018年。Ken Shoemake。利用四元数曲线实现旋转动画。载于 第*12*届计算机图形学与交互技术年度会议论文集，第245–254页，1985年。

Junichi Yamagishi、Christophe Veaux 和 Kirsten MacDonald。CSTR VCTK 语料库：供 CSTR 声音克隆工具包使用的英语多说话人语料库（版本 0.92）。说话人朗读的彩虹短文可在国际英语方言档案中找到：*(http://web. ku. edu/˜ idea/readings/rainbow. htm).*，2019年。Jonathan Gratch、Ron Artstein、Gale M Lucas、Giota Stratou、Stefan Scherer、Angela Nazarian、Rachel Wood、Jill Boberg、David DeVault、Stacy Marsella 等。人机访谈心理困扰分析语料库。载于*LREC*，第3123–3128页。雷克雅未克，2014年。Andrew Bailey 和 Mark D Plumbley。利用音频特征进行抑郁症检测中的性别偏差。载于*2021年第29届欧洲信号处理会议（EUSIPCO）*，第596–600页。IEEE，2021年。Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。DeepSeek-R1：通过强化学习激发大语言模型的推理能力。*arXiv*预印本 *arXiv:2501.12948*，2025年。卡尔·皮尔逊。X. 关于如下判据：在相关变量系统中，给定一组偏离其可能值的偏差，若可合理地认为这些偏差源自随机抽样，则该判据成立。《伦敦、爱丁堡和都柏林哲学杂志与科学期刊》，50(302):157–175，1900年。伊利亚·洛希洛夫和弗兰克·胡特。解耦权重衰减正则化。*arXiv*预印本 *arXiv:1711.05101*，2017 年。Jungil Kong 、 Jaehyeon Kim 和 Jaekyoung Bae。HiFi-GAN：用于高效、高保真语音合成的生成对抗网络。《神经信息处理系统进展》，33:17022–17033，2020年。Xingchen Ma、Hongyu Yang、Qiang Chen、Di Huang 和 Yunhong Wang。DepAudioNet：一种高效的基于音频的抑郁症分类深度模型。载于第六届视听情绪挑战赛国际研讨会论文集，第35–42页，2016。Jinhan Wang、Vijay Ravi 和 Abeer Alwan。通过原始语音信号进行抑郁症检测的非均匀说话人解耦。载于*Interspeech*，第2023卷，第2343页，2023年。Brecht Desplanques、Jenthe Thienpondt 和 Kris Demuynck。ECAPA-TDNN：基于时延神经网络的说话人验证中的强调通道注意力、传播与聚合。*arXiv*预印本 *arXiv:2005.07143*，2020年。John HL Hansen 和 Taufiq Hasan。机器与人类的说话人识别：教程式综述。《*IEEE*信号处理杂志》，32(6):74–99，2015年。NR Draper。应用回归分析。麦格劳-希尔公司，1998年。Simon Kornblith、Mohammad Norouzi、Honglak Lee 和 Geoffrey Hinton。重新审视神经网络表示的相似性。载于国际机器学习会议，第3519–3529页。PMLR，2019年。Matthew Le、Apoorv Vyas、Bowen Shi、Brian Karrer、Leda Sari、Rashel Moritz、Mary Williamson、Vimal Manohar、Yossi Adi、Jay Mahadeokar 等。Voicebox：大规模文本引导的多语言通用语音生成。《神经信息处理系统进展》，36:14005–14034，2023年。Philip Anastassiou、Jiawei Chen、Jitong Chen、Yuanzhe Chen、Zhuo Chen、Ziyi Chen、Jian Cong、Lelai Deng、Chuang Ding、Lu Gao 等。Seed-TTS：一系列高质量、通用的语音生成模型。*arXiv*预印本 *arXiv:2406.02430*，2024。Zhengyang Chen、Sanyuan Chen、Yu Wu、Yao Qian、Chengyi Wang、Shujie Liu、Yanmin Qian 和 Michael Zeng。用于自动说话人验证的大规模自监督语音表示学习。载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*），第6147–6151页。IEEE，2022c。Frank E Harrell、Robert M Califf、David B Pryor、Kerry L Lee 和 Robert A Rosati。评估医学检测的效用。《美国医学会杂志》，247(18):2543–2546，1982年。查尔斯·斯皮尔曼。两事物之间关联的证明与测量。《美国心理学杂志》，100(3/4):441–471，1987年。Marina Sokolova 和 Guy Lapalme。分类任务性能度量指标的系统分析。《信息处理与管理》，45(4):427–437，2009年。Hongyi Zhang、Moustapha Cisse、Yann N Dauphin 和 David Lopez-Paz。Mixup：超越经验风险最小化。*arXiv*预印本 *arXiv:1710.09412*，2017年。Zhihao Du、Yuxuan Wang、Qian Chen、Xian Shi、Xiang Lv、Tianyu Zhao、Zhifu Gao、Yexin Yang、Changfeng Gao、Hui Wang 等。CosyVoice 2：利用大语言模型实现可扩展的流式语音合成。*arXiv*预印本 *arXiv:2412.10117*，2024。