情感感知前缀：迈向语音转换模型中的显式情感控制

Haoyuan Yang、*Mu Yang*、*Jiamin Xie*、*Szu-Jui Chen*、*John H.L. Hansen***

鲁棒语音系统中心（CRSS），德克萨斯大学达拉斯分校，美国{haoyuan.yang, mu.yang, jiamin.xie, szu-jui.chen, john.hansen}@utdallas.edu

arXiv:2603.09120v1 [eess.AS] 10 Mar 2026

摘要

近年来，零样本语音转换取得了显著进展，在情感控制方面展现出潜力，但由于表达能力有限，其性能仍不理想或不够稳定。我们提出情感感知前缀，用于在两阶段语音转换骨干网络中实现显式情感控制。在不损害说话人身份的前提下，我们显著提升了情感转换性能：情感转换准确率（ECA）从基线方法的 42.40% 翻倍至 85.50%，同时保持语言完整性和语音质量。消融研究表明，要合成鲜明的情绪，必须联合控制序列调制和声学实现。此外，比较分析验证了所提方法的泛化能力，并进一步揭示了声学解耦在维持说话人身份方面的作用。 [1]

索引词：情感语音转换、情感迁移、语音转换、语音生成

1. 引言

情感控制对于语音生成的自然度和生动性至关重要，因为它传达了说话人的感受、动机和个性 [1]。有效的情感控制是打造真正沉浸式人机界面的基石，可赋能富有表现力的配音、人机交互和说话人匿名化等应用 [2, 3, 4]。

近年来，零样本语音转换已成为一种强大的范式，天然具备实现情感转换的潜力。通过利用参考提示中的风格信息，这些模型将语言内容序列转换为富含风格信息的音频令牌序列，随后通过声学建模，重建具有目标身份和风格的语音 [5]。值得注意的是，VEVO提出了一种两阶段设计，将零样本语音转换分解为序列调制和声学实现，从而将高层级的时间与韵律结构同低层级的频谱合成分离开来 [6]。这与语音产生的多尺度组织方式相呼应：韵律模式的动态调制在相对稳定、因说话人而异的声学约束下运行，并由此实现情感表达 [7]。

然而，我们的研究发现，现有的零样本语音转换模型往往表现出不理想或不稳定的情感可控性。虽然它们能够模仿整体说话风格，却缺乏将源语句转换为特定、高强度目标情感所需的控制能力。这一局限源于模型在动态调制阶段缺少明确的情感控制，导致模型过度依赖声学提示提供的隐式线索（例如全局能量或平均基频）。

本文提出了情感感知前缀，用于在语音转换模型中实现明确的情感控制。针对情感语音转换（EVC）任务，我们的方法在保留语言内容、说话人身份和整体质量的同时，实现了明确的情感控制。我们提出的方法基于 VEVO [6]，这一两阶段零样本语音转换框架。如图1所示，我们在整个两阶段架构中利用提示语音中的情感线索。通过将内容不变风格特征与情感向量相结合，我们设计了情感感知前缀。该前缀通过深度前缀提示机制引导序列调制阶段，从而确保对生成的音频词元进行明确控制。同时，在声学实现阶段使用同一参考语音，以提供稳定的低层频谱信息。我们的主要贡献如下：

1.**提升情感可控性**。通过引入采用深度前缀提示的情感感知前缀，我们将 VEVO 的情感转换能力（以目标情感转换准确率（ECA）衡量）从 42.40% 提升至 85.50%，同时保持说话人身份和音质。

2. **理解情感控制的层次敏感性**。通过分阶段隔离情感提示，我们证明序列级调制是高层韵律意图的主要驱动因素，而与声学阶段进行联合控制则能使转换准确率获得显著的非加和提升。

3. **探究声学解耦的作用**。我们将所提方法应用于一种不含声学解耦的单阶段语音转换框架。比较分析揭示了所提方法在情感控制方面的效能，并阐明了声学解耦对于保持说话人身份的重要性。

2. 相关工作

情感语音转换研究为在保持说话人身份的同时操控情绪状态所需的机制提供了关键洞见。早期方法主要依赖显式的表示解耦，采用专用情感编码器 [8] 或互信息目标 [9]，将情感风格与说话人身份及语言内容分离。为减少对高质量平行数据 [10]和大量人工标注 [11], 的依赖，后续研究转向非平行 [12, 13] 和真实场景设置 [14]。更新近的工作进一步超越了僵化的类别式情感迁移，转向更细腻的表现力，包括可控

**表示通讯作者。[1]演示请见：https://anonysample.github.io/vevoemoplus/

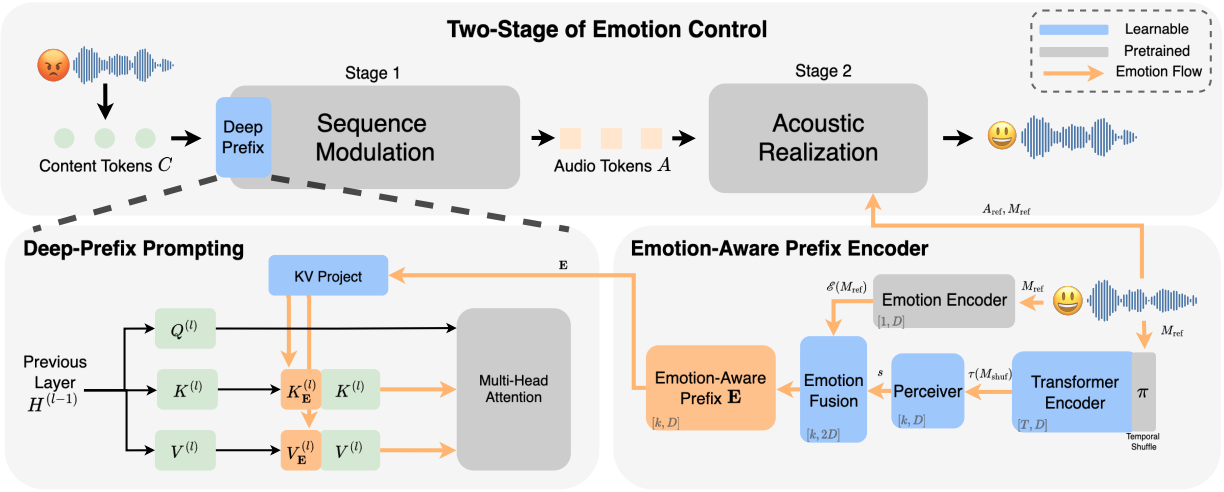


图1： 框架概述：序列调制阶段通过情感感知前缀的深度前缀提示执行序列级风格建模，而声学实现阶段则在参考语音的条件下对最终的语音产生进行建模。

语音情感生成 [15] 和细粒度强度调节 [16]。Kreuk等人 [17] 证明，情感可通过离散声学词元表示进行操控，表明情感属性可在自回归序列模型中被隐式编码并加以控制，从而为基于离散词元的语音转换模型中的情感控制铺平了道路 [5, 6]。

3. 方法

为在尽可能少地修改架构的情况下增强情感的表现力，我们在最先进的语音转换框架 VEVO [6]，的基础上，引入**情感感知前缀**和**深度前缀提示**。

3.1. 框架概述

如图1所示，我们所提方法遵循 VEVO 的两阶段语音建模流程：

阶段 1：序列调制。阶段 1 采用自回归（AR）Transformer 来预测离散且具有丰富风格信息的音频词元。AR Transformer 的提示由给定的内容词元和所提情感感知前缀$\mathbf{E}$共同组成，后者将在第 3.2 节中讨论。该过程可形式化为：

$$P(A|C,\mathbf{E}) = \prod_{t=1}^{T} P(a_t|a_{<t}, C, \mathbf{E}) \tag{1}$$

其中，$A = [a_1, a_2, ..., a_T]$ 是待预测的音频词元，$C = [c_1, c_2, ..., c_n]$ 表示从输入语音中提取的内容词元。

阶段2：声学实现。阶段2采用流匹配（FM）Transformer，根据预测的音频词元 A，并以参考音频词元 A_{ref} 及其对应的真实梅尔频谱 M_{ref}为条件，重建梅尔频谱$\hat{\mathrm{M}}$。最终波形由神经声码器生成。

3.2. 情感感知前缀编码器

情感感知前缀编码器可提供话语级、内容不变的情感风格嵌入。该编码器由时序打乱 Transformer、Perceiver层和情感融合层组成。给定参考梅尔频谱 $\mathrm{M}_{\text{ref}} \in \mathbb{R}^{T\times D}$，这些模块按如下方式对 M_{ref} 进行编码：

时序打乱 Transformer 编码器。为减少内容泄漏，同时保留全局风格特征，我们对参考梅尔频谱的时间索引应用随机置换 $\pi(\cdot)$，得到 $\mathrm{M}_{\text{shuf}} = \pi(\mathrm{M}_{\text{ref}})$。此操作会扰乱音素与语言结构，同时保留与韵律和音色相关的帧级声学统计量。随后，将打乱后的帧输入轻量Transformer $\mathcal{T}$，以提取高层风格表示。

Perceiver层。受 GenVC[5],架构的启发，我们采用 Perceiver层将可变长度的潜在特征 $\mathcal{T}(\mathrm{M}_{\text{shuf}})$压缩为固定长度的风格嵌入 $s \in \mathbb{R}^{k\times D}$。具体做法是固定 k 个可学习的潜在标记 L，使其与 $\mathcal{T}(\mathrm{M}_{\text{shuf}})$进行交叉注意力。固定长度设计充当瓶颈，确保前缀注入自回归 Transformer 时维度一致，避免风格表征随话语时长增长。

情感融合层。为融入明确的情感信息，预训练情感编码器 $\mathcal{E}$ 从参考梅尔频谱 M_{ref}中提取嵌入。随后将其广播，并与风格嵌入 s 融合，再投影至语言模型的隐藏维度：

$$\mathbf{E} = \mathrm{Linear}(\mathrm{Concat}[s; \mathcal{E}(\mathrm{M}_{\text{ref}})]) \tag{2}$$

其中，Concat[;] 表示在特征维度上进行拼接。所得向量$\mathbf{E}$充当**情感感知前缀**，用于增强后续的音频标记生成。

3.3. 深度前缀提示

为确保在生成 token 的整个过程中，序列调制阶段的情感控制保持一致，我们采用了一种类似于 P-Tuning-v2 [18] 的 Deep-Prefix Prompting 机制。我们并非简单地将情感感知前缀 $\mathbf{E}$ 添加到输入序列前，而是将其注入为语言模型的逐层键值缓存。在每一层，我们使用独立的键和值投影，将情感感知前缀 $\mathbf{E}$ 映射到该层的潜在空间：

$$K_{\mathbf{E}}^{(l)} = \mathbf{E}\cdot W_K^{(l)}, \quad V_{\mathbf{E}}^{(l)} = \mathbf{E}\cdot W_V^{(l)} \tag{3}$$

表 1： 客观评估结果。我们所提方法在情感分类准确率方面显著优于 *VEVO* 骨干网络和其他最先进的基线方法，同时保持了说话人身份、语言内容和整体质量。

模型	说话人		质量与可懂度				情感	
	说话人中心相似度 ↑	等错误率 ↓	UT-平均意见分v2↑	DMOS		词错误率 ↓	情感相似度 ↑	情感分类准确率 ↑
				SIG	OVRL			
StarGANv2-VC-EVC	**0.565**	**3.24%**	2.318	3.330	3.037	5.87%	0.686	36.00%
StepAudioEditX（迭代 2）	0.425	6.50%	**3.395**	3.308	3.005	**4.43%**	0.634	20.90%
GenVC	0.238	20.87%	2.157	3.328	3.021	8.06%	0.681	32.48%
VEVO	0.476	5.40%	3.060	3.409	3.093	10.08%	0.696	42.40%
所提方法（无 Deep-Prefix）	0.503	4.42%	2.946	3.416	3.133	5.74%	0.842	83.50%
所提方法	0.500	4.50%	2.960	**3.431**	**3.152**	6.28%	**0.850**	**85.50%**

其中， $W_{(l)K}$和 $W_V^{(l)}$分别表示语言模型第 l 层的键和值投影矩阵。由此得到的前缀向量 $K^{(l)}{}_{\mathbf{E}}$ 和 $V^{(l)}{}_{\mathbf{E}}$，会分别添加到由前一层隐藏状态得到的标准键和值矩阵前面，形式为 $[K^{(l)}{}_{\mathbf{E}}; K^{(l)}]$ 和 $[V_{(l)}{}_{\mathbf{E}}; V_{(l)}]$,，随后执行标准的注意力计算。

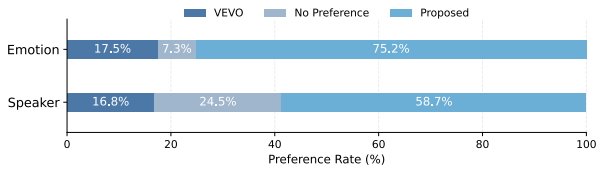


图2： 主观偏好：*VEVO* 与所提方法

4. 实验

4.1. 数据集

我们在情感语音数据集（ESD）[10]上训练所提模型。我们选取了涵盖全部五种情绪（中性、开心、悲伤、愤怒和惊讶）的全部 10 位英语说话人（5 男 5 女）。对于每位说话人和每种情绪，数据集均提供 350 条平行话语。我们使用前 300 条话语进行训练，其余 50 条中，20 条用作参考提示，30 条用于测试。

4.2. 实现细节

主框架基于 VEVO 构建，并遵循其分词器、自回归 Transformer、FM Transformer 和声码器的原始配置。我们的情感感知前缀编码器由一个 6 层时序打乱 Transformer 和一个包含 $k = 32$ 个潜在令牌的 Perceiver 层组成。我们使用预训练的 Emotion2Vec+ Large 模型 [19] 提取单个情感嵌入以进行情感融合，如第 3.2 节所述。在微调过程中，所有骨干网络参数均保持冻结。情感感知前缀编码器完全可训练，并将 LoRA [20]（秩为 $r = 32$）应用于自回归 Transformer，以实现轻量级适配。使用 AdamW 以 2×10^{-5} 的学习率进行 46k 步微调。

4.3. 基线系统

我们将结果与四种突出的基线系统进行比较：**(1)VEVO**，一种先进的零样本语音转换系统，以高保真风格模仿著称，也是所提方法的骨干网络；**(2) GenVC**，一种利用 Perceiver 编码器将说话人风格与内容解耦、无需监督标签的零样本语音转换模型；**(3) StarGANv2-VC-EVC**：我们定制实现的 StarGANv2-VC [21], 情感语音转换变体，遵循 StarGAN-EVC 的原则，实现多对多情感迁移；**(4)StepAudioEditX** [22]：一款拥有 30 亿参数的大型音频语言模型（LALM），作为基于基础模型的表现力音频编辑基准；我们报告其第二次编辑迭代的结果，记为 iter 2。

4.4. 客观评估

我们从三个主要维度评估性能：**说话人**。我们使用 ECAPA-时延神经网络 [23]提取说话人嵌入。根据观察到的同一说话人在不同情绪条件下的相似度可能有所变化这一现象 [24],，我们利用训练数据为每位说话人计算说话人级跨情绪嵌入质心，并将其用作身份锚点。我们报告说话人质心相似度（Spk-Cent SIM），其定义为转换后的话语与目标说话人质心之间的余弦相似度；同时，按照标准说话人验证任务报告等错误率（EER）。

质量与可懂度。我们使用 UTMOSv2 [25] 评估自然度，并使用 DNSMOS P.835 [26] 评估整体感知质量。对于可懂度，我们使用 Whisper-v3-Large [27] 转录生成的语音，并与真实文本进行比较，计算词错误率（WER）。

情感。我们使用 Emotion2Vec+ Large [19] 获取预测的情感标签和嵌入。我们报告情感转换准确率（ECA），即比较转换后的话语的 top-1 预测情感标签与对应的真实标签；此外还报告情感相似度得分（Emo SIM），其定义为转换后的话语与目标话语的嵌入之间的余弦相似度。

4.5. 主观评估

邀请了十名人类受试者，对 VEVO 基线方法和我们所提方法生成的样本进行语音质量、说话人相似度和情感相似度评估。每位受试者总共听取了 120 条话语。

质量。参与者使用 1（较差）至 5（优秀）的评分量表，评估转换样本的整体质量。我们报告平均意见得分（MOS）。

情感偏好。主观情感相似度通过 ABX偏好测试进行评估。每次试验中，听众会听到一条参考语句和两个转换后的样本（基线方法和所提方法）。他们选择情感更接近参考语句的那个样本，

表 2: 情感控制中的分层敏感性分析：基线方法 *VEVO* 与所提方法。

控制	模型	等错误率 ↓	情感相似度 ↑	情感分类准确率 ↑
序列	VEVO	3.98%	0.606	12.50%
	所提方法	3.00%	0.751	47.00%
声学	VEVO	4.50%	0.660	32.70%
	所提方法	3.53%	0.691	34.50%
联合	VEVO	5.40%	0.696	42.40%
	所提方法	4.50%	**0.850**	**85.50%**

情感更接近参考音频。

说话人偏好。主观说话人相似度采用 5 分制评分量表进行评估（1：非常不相似，5：高度相似），并以参考语句为参照。每次试验中，通过比较基线方法样本与所提方法样本的说话人相似度评分，得出偏好统计结果。

5. 结果与分析

5.1. 情感感知前缀的影响

客观评估。表 1 展示了所提方法与选定基线系统的性能对比。无 *Deep-Prefix* 的所提方法表示将前缀嵌入添加到输入序列中，而非添加到 KV 缓存中的所提方法。即使没有所提方法的 Deep-Prefix，该方法在情感转换方面已取得显著提升，情感分类准确率达到 83.50%，接近 VEVO 骨干网络 42.40% 情感分类准确率的两倍。Deep-Prefix 机制（**所提方法**）进一步将情感分类准确率提升至 85.50%，同时取得最高的 0.850 情感相似度，展现了其在最先进的语音转换系统中实现情感细粒度控制的有效性。

值得注意的是，尽管所提方法显著提升了情感转换效果，但其说话人身份、音质和可懂度仍达到与 VEVO 骨干网络相当或更高的水平。虽然自然度略有下降，但其他所有指标均有所提升，这表明我们提出的情感控制不会破坏底层语音转换性能。

主观评估。在平均意见分方面，所提方法的得分略高于 VEVO（4.018 vs. 3.878）。图2展示了情感和说话人相似度的偏好结果，显示感知相似度方面的差距显著更大。所提方法在说话人偏好（58.7% vs. 16.8%）和情感偏好（75.2% vs. 17.5%）两方面均显著优于 VEVO。这一明显差异不仅表明情感转换性能有所提升，也说明更准确的情感呈现增强了说话人身份的感知一致性。

5.2. 在何处以及如何控制情感?

为了解单个阶段的情感控制敏感度，我们将情感提示分别限制在两个阶段，并尝试以下设置：**序列控制**，其中我们分别向阶段 1 和阶段 2 提示 [目标情感、源情感]；**声学控制**，其中我们提示相反的 [源情感、目标情感]；**联合控制**，其中我们向 [目标情感、目标情感]。结果如表 2 所示。我们有以下观察：

表 3： 比较分析：采用和不采用情感感知前缀的 *GenVC*。

模型	等错误率 ↓	情感相似度 ↑	情感分类准确率 ↑
GenVC	**20.87%**	0.681	32.48%
GenVC 配合 EAP	44.51%	**0.760**	**58.35%**

序列调制的主导作用。在序列控制设置下，情感分类准确率从 VEVO 的 12.50% 提升至所提方法的 47.00%，表明序列级别的情感建模显著改善。此外，对于所提方法，序列控制（47.00%）与声学控制（34.50%）之间的情感分类准确率差距表明，阶段 1 是成功实现情感转换的主要驱动因素。这可能归因于阶段 1 对更高层次韵律意图的建模。

声学实现中的情感兼容性。对于 VEVO，声学控制达到 32.70%，显著优于序列控制（12.50%），这表明基线方法主要通过声学阶段表达情感。

联合控制的协同效应。所提方法通过联合控制获得了最高的情感分类准确率 85.50%，展现出明显的非加和式提升。一旦情感在序列级别得到有效编码，冻结的声学阶段便能更忠实地实现情感特征，这凸显了两阶段架构中情感建模的层次性与相互依赖性。

5.3. 比较分析：声学解耦在身份保持中的作用

为考察声学解耦的结构作用，我们将情感感知前缀应用于 GenVC（记为 GenVC 配合 EAP），并进行了比较分析。如表 3 所示，尽管情感转换显著改善（情感分类准确率 32.48% → 58.35%），说话人身份却严重退化（等错误率 20.87% → 44.51%），即使在使用同一说话人的提示时也是如此。这与我们将 EAP 应用于 VEVO 时观察到的情况不同（见表 1 和表 2），表明声学解耦对于保持说话人身份可能至关重要。VEVO 保留了一个独立的声学实现阶段，并使用冻结的 FM Transformer（声学解耦）；而 GenVC 则将序列级表示直接输入微调声码器（声学耦合）。在不提示或调整声码器、仅提供来自序列调制阶段的情感提示表示的情况下，GenVC 能够生成情感语音，但代价是说话人身份崩塌。相比之下，借助专门的声学实现阶段，VEVO 能够生成兼具情感表现力且保持说话人身份的语音。这一发现表明，声学解耦可能对于在稳定的情感控制中保持说话人身份起着关键作用。

6. 结论

本文提出了情感感知前缀，以增强语音转换模型中的显式情感控制。通过利用深度前缀提示，我们将基线方法的情感分类准确率（ECA）从42.40%提升至85.50%，同时保持内容、质量和说话人身份。研究结果表明，要获得鲜明的情感表达，需要在两个阶段进行联合控制。此外，比较分析证实了该模块的可移植性，并表明解耦的声学阶段对身份保持至关重要。

7. 生成式人工智能使用披露

在本研究的撰写过程中，所有（共同）作者仅使用生成式人工智能工具对语法和用词进行审阅与修正，且所有（共同）作者对论文内容承担全部责任。

8. 参考文献

[1] K. R. Scherer，《情感的声音交流：研究范式综述》，《*Speech Communication*》，第 40 卷，第 227–256 页，2003 年。[2] Y. Zhao、R. Liu 和 G. Cong，《通过多尺度多模态上下文交互实现表现力视频配音》，载于*IEEE ICASSP* 会议论文集，2025 年，第 1–5 页。[3] K. Zhou、B. Sisman、R. Rana、B. W. Schuller 和 H. Li，《情感语音转换中的情绪强度及其控制》，《*IEEE*情感计算汇刊》，第 14 卷，第 1 期，第 31–48 页，2023 年。[4] J. Yao、H. Liu、E. S. Chng 和 L. Xie，《EASY：通过因子化蒸馏实现情感感知说话人匿名化》，载于*Interspeech* 会议论文集，2025 年。[5] Z. Cai、H. L. Xinyuan、A. Garg、L. P. Garc´ıa-Perera、K. Duh、S. Khudanpur、M. Wiesner 和 N. Andrews，《GenVC：自监督零样本语音转换》，载于*IEEE*自动语音识别与理解研讨会论文集，2025 年。[6] X. Zhang、X. Zhang、K. Peng、Z. Tang、V. Manohar、Y. Liu、J. Hwang、D. Li、Y. Wang、J. Chan、Y. Huang、Z. Wu 和 M. Ma，《VEVO：通过自监督解耦实现可控的零样本语音模仿》，载于*ICLR* 会议论文集，2025 年。[7] K. R. Scherer，《声音情感表达：综述及未来研究模型》，《心理学公报》，第 99 卷，143，1986 2，第 p. .。[8] X. He、J. Chen、G. Rizos 和 B. W. Schuller，《用于情感语音转换的改进版 StarGAN：提升语音质量与数据增强》，载于*Interspeech* 会议论文集，2021 年，第 821–825 页。[9] Z. Du、B. Sisman、K. Zhou 和 H. Li，《表达性语音转换中的情感风格与说话人身份解耦》，载于*Interspeech* 会议论文集，2022 年，第 2603–2607 页。[10] K. Zhou、B. Sisman、R. Liu 和 H. Li，《情感语音转换：理论、数据库与 ESD》，《*Speech Communication*》，第 137 卷，第 1–18 页，2022 年。[11] C. Busso、R. Lotfian、K. Sridhar、A. N. Salman、W. Lin、L. Goncalves、S. Parthasarathy、A. R. Naini、S. Leem、L. Martinez-Lucas、H. Chou 和 P. Mote，《MSP-Podcast 语料库》，*CoRR*，第 abs/2509.09791 卷，2025 年。[12] G. Rizos、A. Baird、M. Elliott 和 B. W. Schuller，《用于情感语音转换的 StarGAN：通过端到端情绪识别的数据增强进行验证》，载于*IEEE ICASSP* 会议论文集，2020 年，第 3502–3506 页。[13] H. Zhu、H. Zhan、H. Cheng 和 Y. Wu，《采用半监督生成建模的情感语音转换》，载于*Interspeech* 会议论文集，N. Harte、J. Carson-Berndsen 和 G. Jones 编，2023 年，第 2278–2282 页。[14] H. Chou、Y. Lin、C. Sung、Y. Tsao 和 C. Lee，《ZSDEVC：采用解耦机制的零样本、基于扩散的情感语音转换》，载于*Interspeech* 会议论文集，2025 年。[15] D. Cho、H. Oh、S. Kim 和 S. Lee，《Emosphere++：通过情感自适应球面向量实现情感可控的零样本文本转语音》，《*IEEE*情感计算汇刊》，第 16 卷，第 3 期，第 2365–2380 页，2025 年。[16] A. P. Gudmalwar、I. D. Biyani、N. J. Shah、P. Wasnik 和 R. R. Shah，《Emoreg：用于基于扩散的语音转换中情绪强度正则化的方向性潜向量建模》，载于人工智能促进协会年会论文集，2025 年，第 23 960–23 968 页。[17] F. Kreuk、A. Polyak、J. Copet、E. Kharitonov、T. A. Nguyen、M. Rivi`ere、W. Hsu、A. Mohamed、E. Dupoux 和 Y. Adi，《使用离散且解耦的表示进行无文本语音情感转换》，载于自然语言处理实证方法会议论文集，2022 年，第 11 200–11 214 页。[18] X. Liu、K. Ji、Y. Fu、Z. Du、Z. Yang 和 J. Tang，《P-tuning v2：提示调优在不同规模和任务中均可普遍媲美微调》，*ArXiv*，第 abs/2110.07602 卷，2021 年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID:238857040[19] Z. Ma、Z. Zheng、J. Ye、J. Li、Z. Gao、S. Zhang 和 X. Chen，《emotion2vec：用于语音情感表示的自监督预训练》，载于*ACL* 论文集论文，2024 年，第 15 747–15 760 页。[20] E. J. Hu、Y. Shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang、L. Wang 和 W. Chen，《LoRA：大型语言模型的低秩适配》，载于*ICLR* 会议论文集，2022 年。[21] Y. A. Li、A. Zare 和 N. Mesgarani，《StarGANv2-VC：一种多样化、无监督、非平行的自然语音转换框架》，载于*Interspeech* 会议论文集，2021 年，第 1349–1353 页。[22] C. Yan、B. Wu、P. Yang、P. Tan、G. Hu、Y. Zhang、Xiangyu、Zhang、F. Tian、X. Yang、X. Zhang、D. Jiang 和 G. Yu，《Step-audio-editx 技术报告》，2025 年。[23] B. Desplanques、J. Thienpondt 和 K. Demuynck，《ECAPA-TDNN：基于时延神经网络的说话人验证中的增强通道注意力、传播与聚合》，载于*Interspeech* 会议论文集，2020 年，第 3830–3834 页。[24] Z. Du、J. Lu、K. Zhou、L. Kaushik 和 B. Sisman，《转换任何人的声音：基于条件扩散模型的端到端表达性语音转换》，载于*Odyssey* 会议论文集，2024 年，第 172–179 页。[25] K. Baba、W. Nakata、Y. Saito 和 H. Saruwatari，《VoiceMOS Challenge 2024 的 T05 系统：从深度图像分类器迁移学习，以预测高质量合成语音的自然度平均意见分》，载于*IEEE SLT* 会议论文集，2024 年，第 818–824 页。[26] C. K. A. Reddy、V. Gopal 和 R. Cutler，《DNSMOS P.835：一种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 ICASSP* 会议论文集，2022 年，第 886–890 页。[27] A. Radford、J. W. Kim、T. Xu、G. Brockman、C. McLeavey 和 I. Sutskever，《通过大规模弱监督实现稳健语音识别》，2022 年。