華東理工大學学报 自然科学版 Natural Science Edition
JOURNAL OF EAST CHINA UNIVERSITY OF SCIENCE AND TECHNOLOGY

基于联合对比学习策略的单样本语音转换模型

王宇航 陈宁

One-Shot Voice Conversion Model Based on Combined Contrastive Learning

WANG Yuhang, CHEN Ning

在线阅读 View online: http://doi.org/10.14135/j.cnki.1006-3080.20251114001

您可能感兴趣的其他文章

Articles you may be interested in

基于内容特征kNN回归的零样本口音转换模型

Zero-Shot Accent Conversion Model Based on the kNN Regression of Content Features

华东理工大学学报（自然科学版）. 2025, 51(4): 497-504 http://doi.org/10.14135/j.cnki.1006-3080.20241009001

基于双曲空间的多视图对比学习捆绑推荐模型

Hyperbolic Space-Based Multi-View Contrastive Learning Model for Bundle Recommendation

华东理工大学学报（自然科学版）. 2026, 52(1): 109-117 http://doi.org/10.14135/j.cnki.1006-3080.20250326002

基于异构感知和对比学习的图推理机制

Research on Graph Reasoning Mechanism Based on Heterogeneous Perception and Contrastive Learning

华东理工大学学报（自然科学版）. 2026, 52(1): 129-141 http://doi.org/10.14135/j.cnki.1006-3080.20250429001

基于二阶段对比学习的中文自动文本摘要方法研究

Automatic Chinese Text Summarization Based on Two-Stage Contrastive Learning

华东理工大学学报（自然科学版）. 2024, 50(4): 586-593 http://doi.org/10.14135/j.cnki.1006-3080.20230713001

平衡信息与动态更新的原型表示联邦学习

Balance Information and Dynamic Update in Prototype-Based Federated Learning

华东理工大学学报（自然科学版）. 2024, 50(6): 905-912 http://doi.org/10.14135/j.cnki.1006-3080.20231019001

多传感器融合与少样本学习的燃烧早期预警及分类方法

Combustion Early Warning and Classification Method Based on Multi-Sensor Fusion and Few-Shot Learning

华东理工大学学报（自然科学版）. 2025, 51(5): 645-656 http://doi.org/10.14135/j.cnki.1006-3080.20250215001

扫码关注公众号，获取更多信息！

文章编号：1006-3080(2026)04-0574-09 DOI: 10.14135/j.cnki.1006-3080.20251114001

基于联合对比学习策略的单样本语音转换模型

王宇航， 陈 宁
（华东理工大学信息科学与工程学院, 上海 200237）

摘要: 针对单样本 (one-shot) 语音转换任务中目标说话人内容保持与说话人特征迁移难以兼顾的问题，本文在变分推理语音合成 (Variational Inference Text-to-Speech, VITS) 框架的基础上，提出了基于联合对比学习策略的单样本语音转换模型。该模型在内容与说话人两个特征空间同时引入对比约束。针对内容特征提取，将 WavLM (Waveform Language Model) 第 23 层的输出经 K 最近邻 (K-Nearest Neighbors, KNN) 匹配后，与 WavLM 第 24 层的输出结合，构成样本对，以实现音素级对齐和语义一致性。针对说话人特征提取，通过时序扰动生成对比样本，并采用监督对比损失，提取稳定且与内容无关的说话人嵌入。生成器部分通过结合条件变分自编码器 (Conditional Variational Auto Encoder, CVAE)与对抗损失进行优化训练。实验结果表明，该模型在 VCTK (Voice Cloning Tool Kit) 语料库上获得的客观与主观评价指标均优于最新基线模型，验证了所提模型在语音自然度、内容保持以及说话人相似性方面的有效提升。

关键词: 单样本；语音转换；对比学习；K 最近邻算法；时序扰乱

中图分类号: TP391 **文献标志码:** A

对于单样本 (one-shot) 的语音转换 (Voice Conversion, VC) 模型而言，由于目标说话人信息极其有限，若同时实现对语音内容的保持和对说话人特征的准确迁移，则需提升模型对语音内容信息和说话人信息的解耦能力。为此，研究者分别从不同角度探索了语音表征解耦的有效途径。文献 [1] 提出的模型通过结合矢量量化 (Vector Quantization, VQ) 和 U 形网络 (U-shaped Network, U-Net) 结构，有效地从连续语音信号中提取离散内容表示，从而实现高保真的语音重建。文献 [2] 基于互信息对抗学习框架，通过最小化内容嵌入与说话人嵌入之间的互信息强化了解耦效果。文献 [3] 则通过引入文本转语音模型的知识迁移机制，将风格化语音合成模型的说话人风格建模能力迁移至 VC 任务，大幅提升了转换后语音的自然度。与此同时，文献 [4] 基于 β-变分自编码器 (β-Variational Autoencoder, β-VAE) 的可控变分自编码器策略，通过调节库尔贝克-莱布勒 (Kullback–Leibler, KL) 散度权重实现了解耦强度的灵活控制，并成功地将所提出的模型扩展至跨语言场景。文献 [5] 则通过引入三重自适应注意力归一化模块实现了多层次语音特征的动态融合，显著增强了模型在任意说话人 VC 中的性能。

近年来，自监督学习 (Self-Supervised Learning, SSL) 得到快速发展，并成为获取通用的富含信息的语音特征的重要手段。在 VC 领域，多项研究表明预训练的 SSL 特征在保留语音内容和韵律信息方面具有显著优势，因而逐步替代传统的音素后验概率图 (Phonetic PosteriorGrams, PPG) 与梅尔频谱 (Mel-Spectrogram) 等特征表示。文献 [6-7] 证明了 SSL 特征可以显著提升 one-shot VC 任务中的内容保持能力与转换后语音的自然度。文献 [8] 通过在 SSL 训练阶段引入说话人解耦机制，有效降低了说话人信息

收稿日期: 2025-11-14
基金项目: 国家自然科学基金面上项目 (61771196)
作者简介: 王宇航(2001—)，男，江苏人，硕士生，主要研究方向为语音转换。E-mail: 15380346091@163.com
通信联系人: 陈 宁，E-mail: chenning_750210@163.com
引用本文: 王宇航, 陈 宁. 基于联合对比学习策略的单样本语音转换模型 [J]. 华东理工大学学报(自然科学版), 2026, 52(4): 574-582.
Citation: WANG Yuhang, CHEN Ning. One-Shot Voice Conversion Model Based on Combined Contrastive Learning[J]. Journal of East China University of Science and Technology, 2026, 52(4): 574-582.

泄露对转换后语音质量的负面影响。文献 [9-10] 通过将不同层级的 SSL 特征分别用于内容与音色提取, 并结合梯度反转层和连接时序分类进行辅助。文献 [11] 把基于预训练 wav2vec2.0 模型的特征提取器与基于信息解耦的方法相结合, 以提高转换后语音的内容保持能力。近期研究表明, 在自监督特征基础上结合变分推理语音合成 (Variational Inference Text-to-Speech, VITS) 框架[12] 的模型, 可进一步提升语音特征解耦性能。例如, 文献 [13-14] 分别在 VITS 的基础上, 通过引入预训练的 WavLM 模型和基于 HuBERT的软单元[15], 实现了内容与说话人特征的有效解耦。文献 [16] 在 VITS 框架上将 HuBERT与轻量化的 SEANet 解码器结合, 完成了任意说话人的实时语音转换。

尽管上述模型具有先进的语义上下文建模能力, 但说话人嵌入与内容嵌入之间的解耦仍不够彻底, 从而导致转换后语音质量难以保持, 进而影响说话人相似度的问题。近期, 对比学习 (Contrastive Learning, CL) 作为一种新解耦策略被引入 VC 任务中[17]。文献 [18-19] 将 CL 与 VQ 结合, 引导内容编码器精确提取语音音素信息。文献 [20] 构造困难负样本以进一步增强模型的判别能力。文献 [21] 在变分自编码器 (Variational Autoencoder, VAE) 中把对比约束作用于说话人或内容潜变量以改善解耦性能。文献 [22] 则将 CL 与 SSL 结合以提高模型的鲁棒性与泛化性。

虽然以文献 [22] 为代表的基于 SSL 特征提取和基于 CL 的语音转换方法在内容保持和说话人特征迁移方面取得了显著进展, 但是特征解耦的有效性仍有提升空间。针对这一问题, 本文在 VITS 框架的基础上提出了一种联合 CL 策略, 在内容和说话人两个特征空间同时引入对比约束, 以实现有效的特征解耦, 进而提升转换语音的自然度。首先, 文献 [22] 虽然通过 CL 增强了自监督内容表征能力, 但对说话人特征的约束不足。主要原因是其所采用的基于音高变换的源语音数据增强方法只能去除说话人基音周期对内容特征提取的干扰, 无法去除说话人其他相关信息的影响[23-24]; 并且, 仅采用 WavLM (Waveform Language Model) 第 23 层提取内容特征, 虽然能够准确匹配音素, 但缺乏上下文语义, 容易匹配到发音相似但语义无关的帧。为此, 本文引入基于 K 最近邻 (K-Nearest Neighbors, KNN) 的样本对构造方法, 以尽可能去除说话人相关信息, 从而实现有效的音素聚类; 并且, 将 WavLM的第 23 层特征聚类后的结果(即离散化的特征)和 WavLM 的第 24 层特征结合构成样本对, 从而在保证音素对齐的同时实现高层语义层级的一致性。在说话人特征提取方面, 文献 [22] 需要通过分别提取自监督特征和梅尔谱特征构造样本对, 模型复杂度较高; 并且, 由于两者之间存在语义鸿沟, 不利于 CL 的实施。为此, 本文采用时序扰乱的样本对生成方法, 一方面降低了计算复杂度, 另一方面避免了语义鸿沟的影响。最后, 本文模型将条件变分自编码器 (Conditional Variational Auto Encoder, C VAE) 损失与生成对抗网络 (Generative Adversarial Network, GAN)损失优化生成器相结合, 使转换语音在自然度和说话人相似性上均得到提升。在 VCTK (Voice Cloning Tool Kit) 语料库[25] 上取得了客观指标与主观指标均优于基线模型的实验结果。模型的语音转换效果演示见 https://wyhnew.github.io/MSCL-VC-demo/。

1　算法描述

语音转换模型总体架构如图 1 所示。它在 VITS 框架的基础上进行了改进, 引入了联合 CL 策略。整体框架主要由 4 个核心模块组成: 内容特征提取器、说话人特征提取器、语音合成器以及后验编码器。其中, 后验编码器与内容编码器结构相同, 下文将不再详述。

1.1　内容特征提取器

1.1.1　预训练 WavLM　作为一种具有高质量语音表征能力的自监督模型, WavLM 能够从原始音频波形中学习深层语义信息。文献 [26] 表明: WavLM 的第 23 层输出与语音识别任务关联最为紧密; 第 24 层输出则更倾向于捕捉全局语义信息和高层上下文特征。因此, 本文在内容表征学习中, 将 WavLM 的第 23 层与第 24 层输出内容作为 CL 的样本对, 以丰富所提取的内容特征表示。

1.1.2　KNN 模块　为了构建高质量的 CL 样本, 本文借鉴文献 [27] 的思路, 引入了 KNN 算法。KNN 本质上是一种基于回归的匹配算法, 通过将输入样本与匹配集中最相似的样本进行匹配, 以预测输入样本的特征表示。鉴于 WavLM 的第 23 层能够提取与音素相关的特征信息, 本文采用该层对当前批次中其他说话人的语音进行特征提取, 并在这些特征上执行 KNN 操作以生成匹配集 M。因此, 相同音素在匹配集 M 中会被编码得更加接近。如图 2 所示, 来自 WavLM 的第 24 层特征样本 X_{query} 会通过预定义的匹配集寻找其 K 个最近邻特征。最终, 这 K 个最近邻特征的平均值 $\hat{X}$ 通过式(1)计算得到。

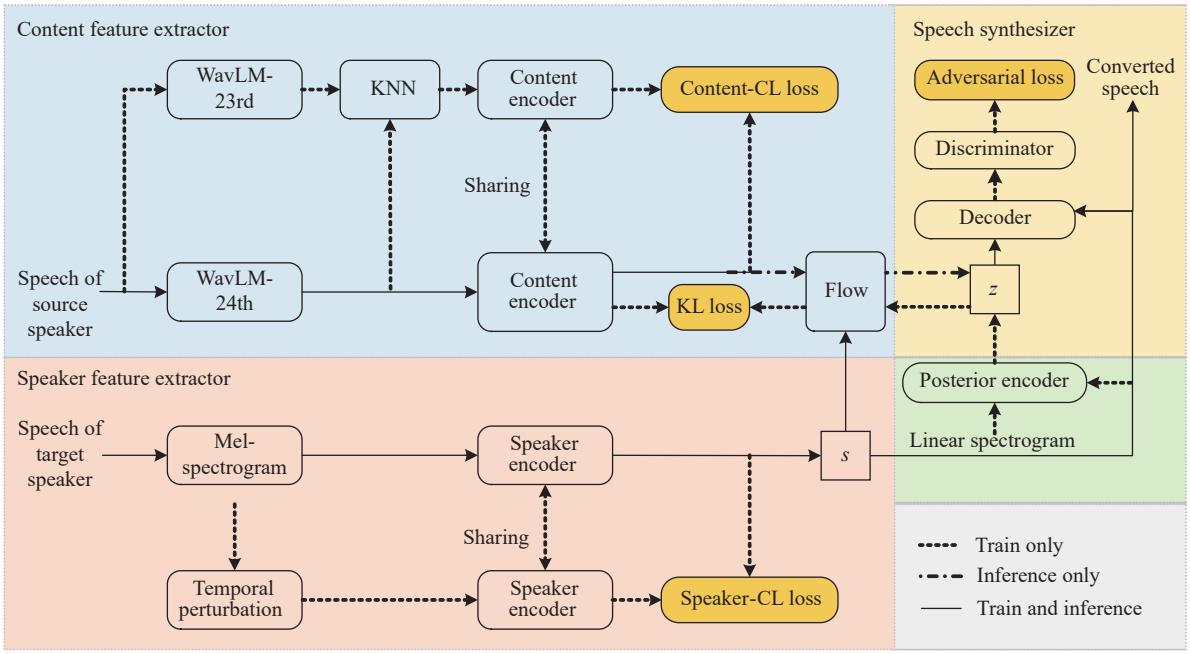


图 1 语音转换模型总体架构

Fig. 1 Overall architecture of voice conversion models

$$\hat{X} = \frac{1}{K}\sum_{j\in N_K} M_j \tag{1}$$

其中，N_K 表示样本 X_t 选取的 K 个最近邻特征的索引集合，计算式如式 (2) 所示。

$$N_K = \underset{S\subseteq\{1,\dots,B\}}{\arg\min}\sum_{j\in S} d_{\cos}(X_t,\ M_j),\ \ |S| = K \tag{2}$$

其中，X_t 表示 X_{query} 的第 t 帧，M_j 表示匹配集中的第 j 个特征，S 表示匹配集的索引集合，$d_{\cos}$ 表示余弦距离，B 为候选样本总数。

1.1.3 内容对比学习策略 为去除自监督特征样本 X_{query} 和 $\hat{X}$ 中与内容无关的信息，它们被输入至内容编码器中，并通过内容对比损失进行约束。该编码器由以下 3 部分组成：一维卷积层、堆叠的空洞卷积块和线性投影层。通过该编码器可分别获得原始内容嵌入 q 和预测内容嵌入 q_+，其映射过程可表示为内容编码器 E_c（式 (3)）。在 CL 过程中，q 与 q_+ 分别被视为查询样本和正样本。

$$q = E_c(X_{\text{query}}),\ \ q_+ = E_c\left(\hat{X}\right) \tag{3}$$

沿时间维度 t 计算查询样本 q 与正样本 q_+ 之间的对比损失 L_c，L_c 可通过 InfoNCE 损失[28]计算，具体如式 (4) 所示。

$$L_c = \frac{1}{T}\sum_{t=1}^{T}\log\frac{\exp(\text{sim}(q_+(t),q(t)))/\tau_0}{\sum_{t=1}^{B}\exp(\text{sim}(q(t),q_i(t)))/\tau_0} \tag{4}$$

其中，变量 q_i 表示负样本，T 表示序列长度，τ_0 为温度系数。此外，本文模型采用了 VITS 中设计的流模块，用于将后验分布映射到目标分布。该流模块由

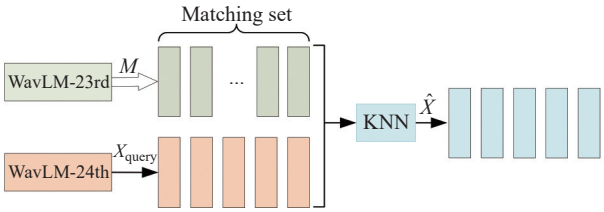


图 2 K-最近邻回归的结构

Fig. 2 Architecture of K-nearest neighbors regression

多个仿射耦合层组成，同时将说话人特征 s 作为条件因子进行引入，以增强先验分布的复杂性，从而提高模型的表达能力。先验分布 $p_\theta(z|q,s)$ 和后验分布 $q_\phi(z|x_{\text{lin}})$ 可表示为：

$$p_\theta(z|q,s) = N\left(z';\mu_\theta,\sigma_\theta^2\right) \tag{5}$$

$$q_\phi(z|x_{\text{lin}}) = N\left(z';\mu_\phi,\sigma_\phi^2\right) \tag{6}$$

其中，z 和 z' 分别为合成语音和高斯采样后的语音潜变量，x_{lin} 为后验编码器的输出，μ 为高斯分布均值，σ^2 为方差，θ 和 ϕ 分别为先验编码器和后验编码器参数。

1.2 说话人特征提取器

1.2.1 时序扰动模块 考虑到内容信息依赖于时序连续性，时序扰动可以有效削弱内容信息的影响[29]。本文通过对原始语音的梅尔频谱（S_{mel}）进行时序扰动构成对比样本 $\hat{S}_{\text{mel}}$。首先，将输入的梅尔语谱图沿时间轴切分为长度均匀的非重叠片段，这样既能保留完整的音素级声学纹理，又能有效切断语义联系；其次，对这些片段的顺序进行随机打乱；最后，将打乱后的片段进行无重叠的拼接，形成增强样本。这种时序扰动的增强方式既能在局部保留短时声学

特性, 又能有效地破坏全局的时序内容信息, 从而迫使说话人编码器专注于提取时不变的音色特征。

1.2.2　说话人对比学习策略　为了从语音中精确提取说话人特征, 本文将 S_{mel} 和 $\hat{S}_{\text{mel}}$ 分别输入说话人编码器进行编码, 该编码器由 3 层长短期记忆网络 (Long Short-Term Memory, LSTM) 和一层全连接层组成。编码器输出的 s 和 $\hat{s}$ 分别表示原始输入的说话人特征和增强输入的说话人特征。随后, 对这些表示应用监督对比损失 (Supervised Contrastive Loss, SupConLoss[30], SupConLoss 记为 L_{s}) 进行约束, L_{s} 计算式见式 (7)。

$$L_{\text{s}}=\frac{1}{N}\sum_{i=1}^{N}\frac{-1}{|P(i)|}\sum_{j\in P(i)}\lg\frac{\exp(s_i\cdot\hat{s}_j)/\tau_1}{\sum_{k\in A(i)}\exp(s_i\cdot s_k)/\tau_1} \tag{7}$$

其中, $P(i)$ 表示锚点样本 i 属于同一类别的正样本索引集合, $A(i)$ 表示除锚点样本外的所有样本集, τ_1 为温度系数, 下标 i、j、k 表示不同的样本。

1.3　语音合成器

本文采用基于 HiFi-GAN 的声码器[31]生成转换后的语音。判别器由 1 个时域判别器和 5 个周期判别器组成, 通过多尺度特征提取综合评估生成语音的质量与真实性。生成器记为 G, 判别器记为 D, 其训练损失函数如下:

$$L_{\text{G}}=L_{\text{adv}}(\text{G},\text{D})+\lambda_{\text{fm}}L_{\text{fm}}(\text{G},\text{D})+\lambda_{\text{mel}}L_{\text{mel}}(\text{G},\text{D}) \tag{8}$$

$$L_{\text{D}}=L_{\text{adv}}(\text{D},\text{G}) \tag{9}$$

其中, L_{adv}、L_{fm} 和 L_{mel} 分别表示对抗损失、特征匹配损失和梅尔频谱图损失, 参数 λ_{fm} 和 λ_{mel} 为权重因子。

1.4　训练损失

训练损失由 CVAE 损失、GAN 损失和 CL 损失组成。其中, CVAE 损失包括: 梅尔频谱重建损失 L_{recon} (式 (10)), 表示重建的梅尔频谱图与目标梅尔频谱图之间的距离 L_1; KL 散度损失 L_{KL}, 表示先验分布 $p_\theta(z|q,s)$ 与后验分布 $q_\phi(z|x_{\text{lin}})$ 之间的 KL 散度。

$$L_{\text{recon}}=\|y_{\text{mel}}-\hat{y}_{\text{mel}}\|_1 \tag{10}$$

其中, $\hat{y}_{\text{mel}}$ 表示重建语音的梅尔频谱图。因此, 训练损失可表示为式 (11):

$$L_{\text{loss}}=L_{\text{recon}}+L_{\text{KL}}+L_{\text{adv}}+L_{\text{fm}}+L_{\text{c}}+L_{\text{s}} \tag{11}$$

2　实验结果与分析

2.1　数据处理与训练设置

基于联合对比学习策略的单样本语音转换模型在 VCTK 数据集上进行训练与评估。VCTK 数据集包含来自 109 位说话人的语音数据, 其中包括 47 名男性和 62 名女性, 每位说话人大约包含 400 条语音样本, 总计超过 4 400 条音频数据。在本实验中, 从中选取 22 位说话人的语音数据作为测试集, 其余说话人的数据用于训练与验证。所有音频样本均被下采样至 16 kHz。线性频谱图与 80 维梅尔频谱图通过采用短时傅里叶变换计算得到, 参数设置为: 快速傅里叶变换点数 1 280、窗长 1 280、帧移 320。模型基于 PyTorch 构建, 并在单张 NVIDIA RTX 3090GPU 上训练 900 000 步, 批大小为 64, 最大语音片段长度为 128 帧。训练过程使用 AdamW 优化器[32], 初始学习率设为 2×10^{-4}。在式 (4) 中, 温度参数 τ_0 设为 0.5; 在式 (7) 中, 温度参数 τ_1 设为 0.1。

在内容特征提取模块 KNN 算法中, K 值的选择决定了对比样本的质量。若 K 值过小, 特征容易受源说话人残留噪声的影响; 若 K 值过大, 会导致特征的频谱包络被平均化, 进而使对比语音样本与源语音的内容相似度下降, 最终使得解耦效果下降[27,33]。为了论证这一结果, 本文对比了不同 K 值下的模型性能, 结果如图 3 所示。当 K=2^2 即 4 时, 转换语音的词错误率 (Word Error Rate, WER) 和梅尔倒谱失真 (Mel-Cepstral Distortion, MCD) 值最低, 因此, 本文选取 K=4 进行后续实验。

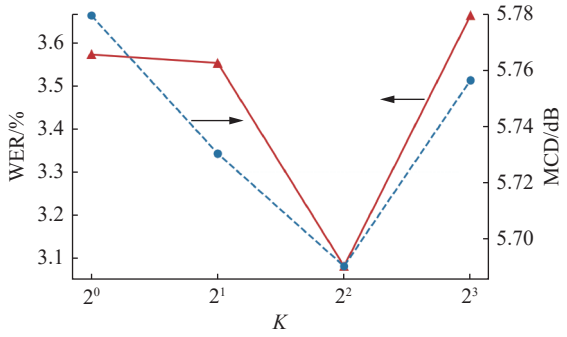


图 3　KNN 中的 K 值对模型性能的影响

Fig. 3　Impact of the K-value in KNN on model performance

2.2　基线模型与实验设置

在本实验中, 选择 7 个语音转换模型作为对比基线模型, 包括 BetaVAE-VC[4]、SRD-VC[2]、TriAAN-VC[5]、LIMI-VC[11]、Neural-VC[16]、Adapt-VC[34]、CLESSR-VC[22], 并对内容嵌入提取器与说话人嵌入提取器中的 CL 策略进行了消融实验。此外, 为了证明本文构建的对比策略优于文献 [22], 将本文构建的用于内容特征提取的对比策略记为 C-CL, 用于说话人特征提取的对比策略记为 S-CL, 分别替换文献 [22] 中基于音高变换 (Pitch-shift) 的对比策略和基于

梅尔特征嵌入 (Mel-embedding) 表示的对比策略, 对模型性能进行测试。除 SRD-VC 使用官方预训练模型外, 其余基线模型均基于官方开源代码并按照 2.1 节所述的语料划分方式重新训练。为了全面评估模型性能, 采用客观评估与主观评估两种方式。

2.3 客观评估

为了客观评估模型性能, 本文对 100 对转换语音的质量进行了测评, 采用的指标包括: WER, 字符错误率 (Character Error Rate, CER), MCD, 等错误率 (Equal Error Rate, EER), 非侵入式语音质量评估 (Non-Intrusive Speech Quality Assessment, NISQA)[35]。此外, 利用伪造检测工具包测试了转换语音的真实性, 并通过 t 分布随机邻域嵌入 (t-distributed Stochastic Neighbor Embedding, t-SNE)[36] 可视化展示了解耦性能。

在表 1 所示的指标中, WER 和 CER 根据文献 [37] 提供的自动语音识别 (Automatic Speech Recognition, ASR) 模型进行计算, 用于评估语音的可懂度; MCD 用于衡量梅尔倒谱系数的差异, 以评估模型的音质; NISQA 用于客观评估语音的自然度与清晰度。实验结果表明, 本文模型在大部分指标上均表现出优异的性能, 但在轻量化方面还有提升空间。首先, 文本模型的 WER 和 CER 最低, 相较于基线模型显著下降, 说明本文模型能准确地从源语音中解耦出内容信息。其次, 根据 NISQA、EER 和 MCD 结果, 本文模型生成的语音在质量与自然度上均优于基线模型。

表 1 本文模型与基线模型的客观性能对比

Table 1 Comparison of objective performance between the proposed model in this paper and baseline models

Model	WER/% (↓)	CER/% (↓)	NISQA (↑)	EER/% (↓)	MCD (↓)	Parameter/10^6 (↓)
BetaVAE-VC[4]	28.97	17.66	2.53	32.91 (0.51)	7.34	24.6
SRD-VC[2]	38.55	23.38	3.37	19.05 (0.64)	10.20	31.38
TriAAN-VC[5]	10.73	6.33	3.11	19.05 (0.65)	5.83	69.19
LIMI-VC[11]	15.24	8.33	2.78	31.65 (0.64)	6.92	**3.67**
Neural-VC[16]	6.66	3.68	4.18	10.08 (0.81)	5.69	28.3
Adapt-VC[34]	53.65	33.09	3.96	77.16 (0.61)	12.96	34.2
CLESSR-VC[22]	3.43	1.33	3.41	15.19 (0.66)	6.79	4.65
This paper	**3.08**	**1.16**	4.34	**10.02 (0.79)**	5.69	39.23
–w/o C-CL	5.03	1.79	**4.52**	15.00 (0.79)	6.09	—
–w/o S-CL	3.97	1.74	4.12	13.16 (0.80)	5.71	—
C-CL→Pitch-shift	13.79	10.29	3.50	17.27 (0.78)	**5.62**	—
S-CL→Mel embedding	5.46	2.12	4.05	11.00 (0.76)	6.36	—

↓—Lower is better; ↑—Higher is better

在伪造检测实验中, 使用基于 LSTM 的说话人编码器 Resemblyzer 工具包验证模型生成语音与真实语音的相似度。首先, 从目标说话人中随机选取 6 条真实语音作为基线; 然后, 从目标说话人中随机选择 4 条真实语音(real-n, n=0,1,2,3)和 6 条由不同模型生成的合成语音组成测试集。对每个测试样本计算其与基线样本的相似度得分(范围 0~1)。此外, 根据真实样本相似度平均值的 95% 置信下界生成阈值, 超过此阈值的样本被视为真实语音。如图 4 所示, 绿色与红色分别表示真实语音与转换语音的得分, 高于虚线的样本被预测为真实语音。本文模型在所有转换场景的转换语音中均表现最佳, 尤其在 M→M 和 F→M 转换场景中, 其得分超过阈值, 被判断为真实语音。

为了直观展示本文模型的特征解耦效果, 从测试集中随机选取 7 位说话人(P231, P236, P237, P347, P256, P376, P295), 每人 200 条语音, 使用 t-SNE 可视化模型提取内容特征、说话人特征与合成语音特征, 结果如图 5 所示。图 5(a) 中提取的内容特征完全混合, 表明说话人信息已被彻底解耦; 图 5(b) 所示的说话人特征中, 同一说话人的类内距离较小, 不同说话人的类间距离较大, 说明模型能有效区分说话人特征; 图 5(c) 所示的合成语音特征按说话人聚类, 进一步说明模型实现了内容与说话人特征的有

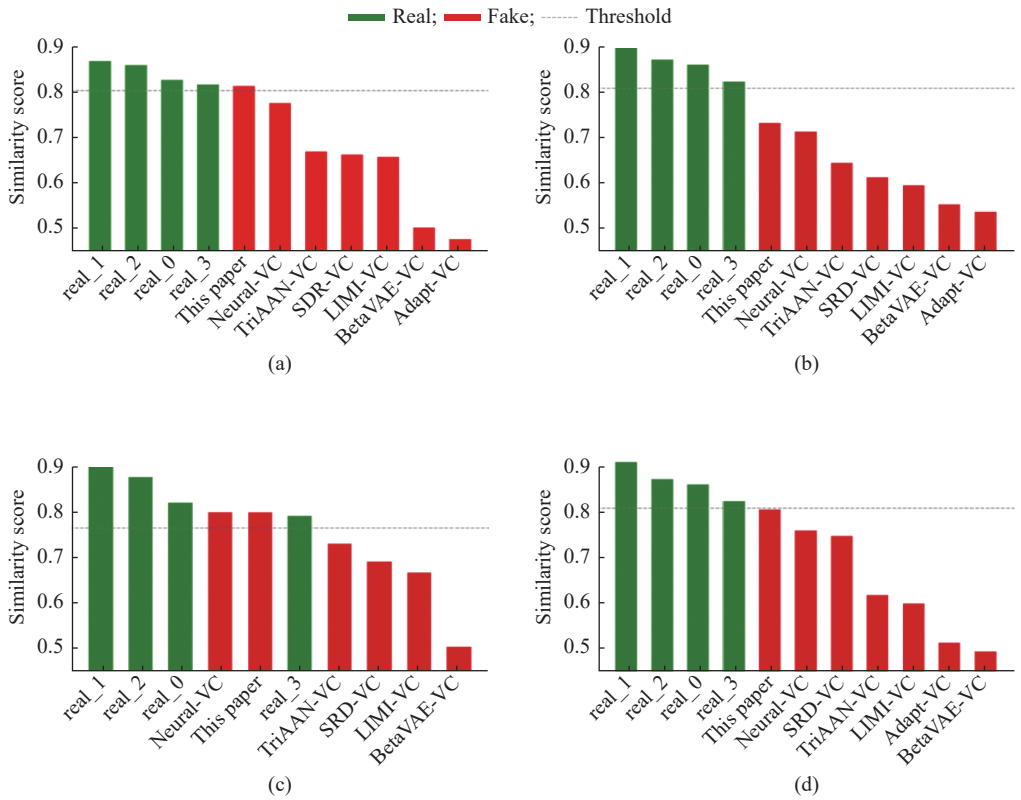


图 4　假语音检测结果: (a) M→M, (b) F→F, (c) F→M, (d) M→F (M: 男性; F: 女性)

Fig. 4　Fake speech detection results: (a) M→M, (b) F→F, (c)F→M, (d) M→F (M: Male; F: Female)

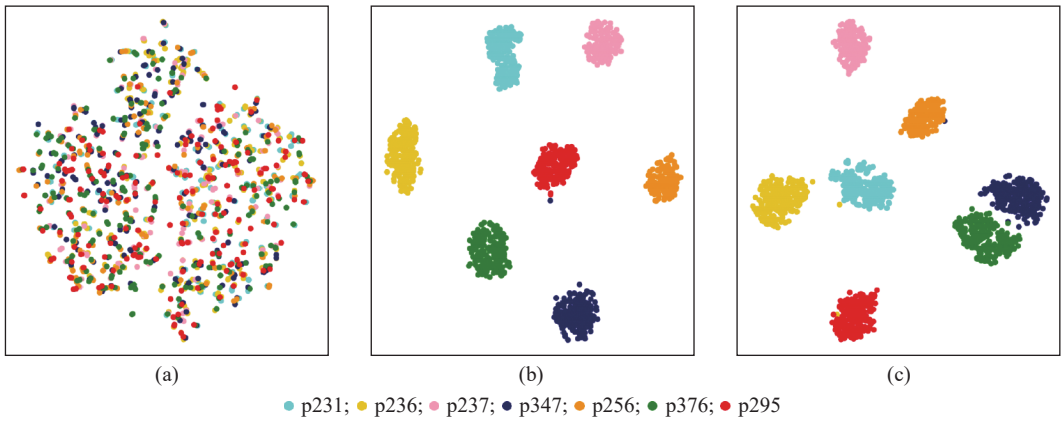


图 5　说话人特征的 t-SNE 可视化: (a) 内容嵌入, (b) 说话人嵌入, (c) 融合嵌入

Fig. 5　t-SNE visualization of speech representations: (a) Content embedding, (b) Speaker embedding, (c) Fused embedding

效分离。

2.4　消融实验

为独立验证模型中 CL 机制在内容与说话人特征上的作用, 进行了两组消融实验。由表 1 结果可知, 当移除 C-CL(即−w/o C-CL)时, WER 与 CER 显著上升, 而 NISQA 下降, 这说明内容特征上的 CL 有效过滤了内容嵌入中的无关信息。同样, 移除 S-CL(即−w/o S-CL)会导致 NISQA 下降, 表明该策略有助于提取更干净的说话人特征, 从而提升合成语音的质量与相似度。并且, 将 C-CL 替换为基于 Pitch-shift 的 CL 策略, 或将 S-CL 替换为基于 Mel-embedding 的对比策略, 均会导致大部分指标的恶化, 这说明本文引入的对比策略更有助于内容特征和说话人特征的解耦。

2.5　主观评估

主观评估包括相似度平均意见得分 (Similarity Mean Opinion Score, SMOS)、自然度平均意见得分 (Naturalness Mean Opinion Score, NMOS) 以及交叉 AB 听感测试 (Cross-AB Test, XAB)。SMOS 越高表示转换语音与目标说话人语音的相似度越高, NMOS 越高则代表语音的自然度越好, 两者分值范围均为 1~5。本实验中共有 33 名具有语音处理背景的研究人员作为受试者参与主观评测。表 2 所示为本文模

表 2 本文模型与基线模型的主观性能对比

Table 2 Comparison of subjective performance between the proposed model in this paper and baseline models

Model	SMOS					NMOS				
	M→F	M→M	F→M	F→F	Average	M→F	M→M	F→M	F→F	Average
BetaVAE-VC[4]	3.25	2.71	3.16	2.63	2.94	2.92	2.21	2.75	2.77	2.66
SRD-VC[2]	3.71	2.58	3.26	3.62	3.29	3.27	2.14	3.01	3.23	2.91
TriAAN-VC[5]	3.68	3.08	3.24	3.50	3.38	3.16	2.49	2.79	3.26	2.93
LIMI-VC[11]	2.79	3.11	3.31	3.08	3.07	2.67	2.61	3.15	2.64	2.78
NeuralVC[16]	**4.21**	3.90	**4.22**	4.03	4.09	3.97	4.24	**4.23**	3.98	4.11
AdaptVC[34]	1.77	2.67	2.00	2.96	2.35	2.02	1.74	1.72	1.97	1.86
CLESSR-VC[22]	4.08	3.86	3.80	3.96	3.93	3.97	3.47	3.81	3.63	3.72
This paper	4.19	**3.95**	4.14	**4.22**	**4.13**	**3.99**	**4.30**	4.07	**4.10**	**4.11**

型与基线模型的主观性能对比结果，由表 2 可知，在同性别与跨性别的 4 种转换场景中，本文模型在大部分情况下均取得最高得分。

XAB 测试用于评估转换语音与参考语音之间的感知相似性。如图 6 所示，本文模型在所有与基线模型的对比中均表现最佳，在语音自然度方面表现最优，且在说话人相似度上与目标说话人最为接近。

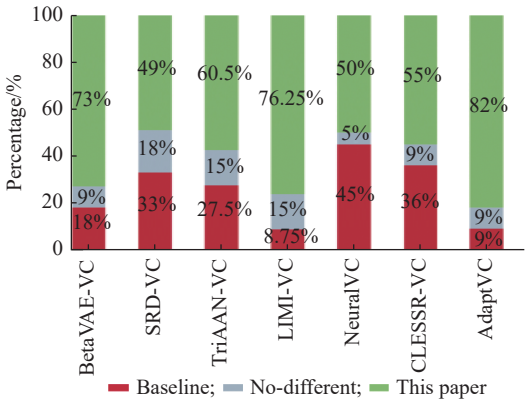


图 6 身份说话人相似度的 XAB 测试结果

Fig. 6 XAB test results for the speaker similarity

3 结束语

本文针对语音转换任务中现有自监督学习方法存在的特征解耦不彻底和说话人特征表征单一等关键问题，提出了一种基于联合对比学习策略与分层特征融合的高保真语音转换模型。该模型在变分推理语音合成框架的基础上，通过在内容与说话人两个特征空间同时引入对比约束，实现了更高性能的特征解耦与重构。所提模型在语音自然度、可懂度及说话人相似度方面均达到了比基线模型更优的性能。未来的研究将进一步探索更高效的对比学习策略和更轻量化的解决方法，以进一步提升模型的整体性能，并同时减少模型的训练参数量。

参考文献：

[1] WU D Y, CHEN Y H, LEE H Y. VQVC+: One-shot voice conversion by vector quantization and u-net architecture[C]// Proceedings of the Annual Conference of the International Speech Communication Association. Shanghai: ISCA, 2020: 4691-4695.

[2] YANG S C, TANTRAWENITH M, ZHUANG H, *et al*. Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion[C]// Proceedings of the Annual Conference of the International Speech Communication Association. Incheon: ISCA, 2022: 2553-2557.

[3] LI Y A, HAN C, MESGARANI N. Styletts-VC: One-shot voice conversion by knowledge transfer from style-based tts models[C]// IEEE Spoken Language Technology Workshop. Doha: IEEE, 2023: 920-927.

[4] LU H, WANG D, WU X, *et al*. Disentangled speech representation learning for one-shot cross-lingual voice conversion using β-vae[C]// IEEE Spoken Language Technology Workshop. Doha: IEEE, 2023: 814-821.

[5] PARK H J, YANG S W, KIM J S, *et al*. Triaan-VC: Triple adaptive attention normalization for any-to-any voice conversion[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023: 1-5.

[6] LIN J, LIN Y Y, CHIEN C M, *et al*. S2VC: A framework for any-to-any voice conversion with self-supervised pretrained representations[C]// Proceedings of the Annual Con-

ference of the International Speech Communication Association. Brno: ISCA, 2021: 836-840.

[7] LIN Y Y, CHIEN C M, LIN J H, *et al*. FragmentVC: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021: 5939-5943.

[8] QIAN K, ZHANG Y, GAO H, *et al*. Contentvec: An improved self-supervised speech representation by disentangling speakers[C]// Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022: 18003-18017.

[9] HUANG H, WANG L, YANG J, *et al*. W2VC: WavLM representation based one-shot voice conversion with gradient reversal distillation and CTC supervision[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2023, 2023(1): 45.

[10] GUO Y, LI Z, LI J, *et al*. Vec2wav 2.0: Advancing voice conversion via discrete token vocoders[EB/OL]. (2024-09-03). https://arxiv.org/abs/2409.01995.

[11] HUANG L, YUAN T, LIANG Y, *et al*. Limi-VC: A light weight voice conversion model with mutual information disentanglement[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023: 1-5.

[12] KIM J, KONG J, SON J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech[C]// International Conference on Machine Learning. Vienna: PMLR, 2021: 5530-5540.

[13] LI J, TU W, XIAO L. Freevc: Towards high-quality text-free one-shot voice conversion[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023: 1-5.

[14] GUO H, LIU C, ISHI C T, *et al*. QuickVC: A lightweight vits-based any-to-many voice conversion model using istft for faster conversion[C]// IEEE Automatic Speech Recognition and Understanding Workshop. Taipei, China: IEEE, 2023: 1-7.

[15] VAN NIEKERK B, CARBONNEAU M A, ZAÏDI J, *et al*. A comparison of discrete and soft speech units for improved voice conversion[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022: 6562-6566.

[16] CAO D, ZHANG Z, ZHANG J. NeuralVC: Any-to-any voice conversion using neural networks decoder for real-time voice conversion[J]. IEEE Signal Processing Letters, 2024, 31: 2070-2074.

[17] LI T, LIU Y, HU C, *et al*. CVC: Contrastive learning for non-parallel voice conversion[C]// Proceedings of the Annual Conference of the International Speech Communication Association. Brno: ISCA, 2021: 1324-1328.

[18] TANG H, ZHANG X, WANG J, *et al*. VQ-CL: Learning disentangled speech representations with contrastive learning and vector quantization[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023: 1-5.

[19] TANG H, ZHANG X, WANG J, *et al*. AVQVC: One-shot voice conversion by vector quantization with applying contrastive learning[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022: 4613-4617.

[20] PRIHASTO B, LIN Y X, LE P T, *et al*. CNEG-VC: contrastive learning using hard negative example in non-parallel voice conversion[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023: 1-5.

[21] DENG Y, ZHANG X, WANG J, *et al*. CLN-VC: Text-free voice conversion based on fine-grained style control and contrastive learning with negative samples augmentation[C]// IEEE International Conference on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking. Wuhan: IEEE, 2023: 1143-1148.

[22] XUE Y, CHEN N, LUO Y, *et al*. Clessr-VC: Contrastive learning enhanced self-supervised representations for one-shot voice conversion[J]. Speech Communication, 2024, 165: 1-8.

[23] AHANGARAN M, DAWALATABAD N, KARJADI C, *et al*. Obfuscation via pitch - shifting for balancing privacy and diagnostic utility in voice - based cognitive assessment[J]. Alzheimer's & Dementia, 2025, 21(3): e70032.

[24] QIAN K, JIN Z, HASEGAWA-JOHNSON M, *et al*. F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020: 6284-6288.

[25] YAMAGISHI J, VEAUX C, MACDONALD K. CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)[R]. Edinburgh: University of Edinburgh, 2019.

[26] CHEN S, WANG C, CHEN Z, *et al*. WavLM: Large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518.

[27] BAAS M, VAN NIEKERK B, KAMPER H. Voice conversion with just nearest neighbors[C]// Proceedings of the Annual Conference of the International Speech Communication Association. Dublin: ISCA, 2023: 2053-2057.

[28] PARULEKAR A, COLLINS L, SHANMUGAM K, *et al*.

Infonce loss provably learns cluster-preserving representations[C]// The Thirty Sixth Annual Conference on Learning Theory. Bangalore: PMLR, 2023: 1914-1961.

[29] HWANG Y, CHO H, YANG H, *et al*. Mel-spectrogram augmentation for sequence to sequence voice conversion [EB/OL]. (2020-01-06). https://arxiv.org/abs/2001.01401.

[30] KHOSLA P, TETERWAK P, WANG C, *et al*. Supervised contrastive learning[C]// Advances in Neural Information Processing Systems. Virtual: Curran Associates Inc, 2020: 18661-18673.

[31] KONG J, KIM J, BAE J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis[C]// Advances in Neural Information Processing Systems. Virtual: Curran Associates Inc, 2020: 17022-17033.

[32] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C]// International Conference on Learning Representations. New Orleans: OpenReview, 2019: 1-8.

[33] HALDER R K, UDDIN M N, UDDIN M A, *et al*. Enhancing k-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications[J]. Journal of Big Data, 2024, 11: 113.

[34] KIM J, KIM J H, CHOI Y, *et al*. AdaptVC: High quality voice conversion with adaptive learning[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Hyderabad: IEEE, 2025: 1-5.

[35] MITTAG G, MÖLLER S. Deep learning based assessment of synthetic speech naturalness[C]// Proceedings of the Annual Conference of the International Speech Communication Association. Shanghai: ISCA, 2020: 1748-1752.

[36] MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579-2605.

[37] WATANABE S, HORI T, KARITA S, *et al*. ESPnet: End-to-end speech processing toolkit[C]// Proceedings of the Annual Conference of the International Speech Communication Association. Hyderabad: ISCA, 2018: 2207-2211.

One-Shot Voice Conversion Model Based on Combined Contrastive Learning

WANG Yuhang, CHEN Ning

(School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract: To address the challenge of balancing content preservation of the target speaker and speaker characteristic transfer in one-shot voice conversion tasks, this paper proposes a one-shot voice conversion model based on a joint contrastive learning strategy, built upon the Variational Inference Text-to-Speech (VITS) framework. The model introduces contrastive constraints in both the content and speaker feature spaces simultaneously. For content feature extraction, the output from the 23rd layer of Waveform Language Model（WavLM）is combined with the output from the 24th layer of WavLM after K-Nearest Neighbors (KNN) matching to form sample pairs, achieving phoneme-level alignment and semantic consistency. For speaker feature extraction, contrastive samples are generated via temporal perturbation, and supervised contrastive loss is adopted to extract speaker embeddings that are stable and content-independent. The generator is trained with the combination of a Conditional Variational AutoEncoder (CVAE) and adversarial loss optimization. Experimental results on the Voice Cloning Tool Kit (VCTK) corpus show that the proposed model outperforms state-of-the-art baseline models in both objective and subjective evaluation metrics, verifying its effectiveness in improving speech naturalness, content preservation, and speaker similarity.

Key words: one-shot; voice conversion; contrastive learning; KNN regression; temporal perturbation

（责任编辑：张欣）