文章编号：1006-3080(2025)04-0497-08 DOI：10.14135/j.cnki.1006-3080.20241009001

基于内容特征 kNN 回归的零样本口音转换模型

罗宜鑫[1]， 陈 宁[1]， 薛宇航[1]， 肖阳阳[2]
（1. 华东理工大学信息科学与工程学院, 上海 200237； 2. 中电万维信息技术有限责任公司, 兰州 730000）

摘要：口音转换 (Accent Conversion, AC) 旨在将源口音语音转换为目标口音语音，并保持源说话人音色和语音内容不变。现有的 AC 模型缺乏对训练数据分布以外的语音口音转换的泛化性。本文提出基于内容特征 k-邻近(kNN)回归的零样本 AC 模型。一方面，采用 WavLM 第 23 层提取源和目标口音语音的内容特征，并利用 kNN 回归将源口音语音内容特征置换为目标口音语音及其最邻近的内容特征以实现口音转换；另一方面，为了保持转换后语音中源说话人音色，构建多说话人声码器对含有目标口音的语音内容特征和源说话人音色特征进行融合，以合成目标口音语音。该模型无需源口音语音参与训练，即可实现多种源口音到目标口音的转换。实验结果表明，该模型取得了比并行或非并行 AC 模型更好的客观与主观评价结果。

关键词：口音转换；kNN 回归；零样本学习；语音转换；声码器

中图分类号：TP391 **文献标志码：**A

口音转换 (Accent Conversion, AC) 是语音转换 (Voice Conversion, VC) 的一个分支任务，旨在将源说话人带有源口音的语音转换为具有目标口音的语音，同时保持源说话人的身份信息和语音内容不变。该技术的应用场景广泛，包括个性化语音合成[1]、电影配音[2]，以及语言学习[3] 等。目前，口音转换面临的最大的难点在于，如何从源口音中重新构建出目标口音的发音规则。

根据训练阶段是否需要与源口音语音具有相同内容的目标口音语音(即并行数据用于发音建模)，现有的 AC 模型可以分为并行方法与非并行方法两类。并行 AC 模型在训练或推理阶段需要语音内容相同且包含目标口音的并行语音，因此这类方法被视为回归方法，即通过拟合源语音和目标语音之间的映射函数来实现口音转换。主流的并行 AC 模型均基于语音后验图 (Phonetic Posteriorgram, PPG)，通过匹配源口音语音和目标口音语音的 PPG 特征实现口音转换。文献 [4] 基于 PPG 特征对源口音和目标口音语音进行帧级特征匹配，用于训练概率模型以计算源语音和目标语音的梅尔频率倒谱系数 (Mel-Frequency Cepstral Coefficient, MFCC) 的联合概率分布，最后基于目标说话人的全局方差，采用最大似然估计计算转换后频谱参数的轨迹。文献 [5-6] 采用源口音语音的 PPG 特征训练一个保持源说话人音色的语音合成器，并在推理过程中输入目标语音的 PPG 特征用于改变源说话人的发音。文献 [7-8] 设计了一个 PPG转换模块，通过源口音语音的 PPG 特征预测转换后的目标口音语音的 PPG 特征，并通过转换后的 PPG特征合成梅尔谱 (Mel-Spectrogram) 以实现口音转换。虽然该方法在一定程度上提高了模型在不同口音下转换性能的泛化性，但依然需要在训练中采用并行数据对源口音 PPG 到目标口音 PPG 特征的转换进行建模。

然而，在实际应用场景中并行数据往往难以获取，并行方法的实际应用价值受限。近期，基于非并行数据训练的 AC 模型成为该领域的研究趋势。但对于非并行方法来说，由于缺少并行语音对正确的发音规则进行引导，实现源口音到目标口音转换的

收稿日期： 2024-10-09
基金项目： 国家自然科学基金面上项目 (61771196)
作者简介： 罗宜鑫(2000—)，男，江苏人，硕士生，主要研究方向为口音转换。E-mail: y30220940@mail.ecust.edu.cn
通信联系人： 陈 宁，E-mail: chenning_750210@163.com
引用本文： 罗宜鑫, 陈 宁, 薛宇航, 等. 基于内容特征 kNN 回归的零样本口音转换模型 [J]. 华东理工大学学报(自然科学版), 2025, 51(4): 497-504.
Citation: LUO Yixin, CHEN Ning, XUE Yuhang, *et al*. Zero-Shot Accent Conversion Model Based on the kNN Regression of Content Features[J]. Journal of East China University of Science and Technology, 2025, 51(4): 497-504.

难度很大。文献 [9] 通过训练针对源口音的语音识别 (Automatic Speech Recognition, ASR) 模型和基于目标口音语音训练的语音合成 (Text-to-Speech, TTS) 模型实现了 ASR-TTS 框架下的口音转换。文献 [10-11] 在目标口音数据集上训练 TTS 模型，通过迁移学习将源口音的语音特征映射到 TTS 编码的文本空间中，最终通过 TTS 解码合成目标口音语音。文献 [12] 将 ASR 预测的文本编码与口音标签结合用于控制目标口音发音，实现了多种口音之间的相互转换。文献 [13-14] 利用 ASR 识别源口音语音音素，并与目标文本音素进行比较，将不同的音素标记为口音音素并进行纠正，最终实现去除口音的语音合成。但本质上此类基于文本的口音转换模型的性能取决于 ASR 或 TTS 模型的性能，并且标注文本参与训练的成本较高。

需要注意的是，一方面，现有的基于并行或非并行数据的 AC 模型在训练过程中都需要源口音语音参与训练，但由于相同口音下不同说话人的发音具有一定差异，因此这类模型的性能在训练数据分布以外的口音转换任务上泛化性不佳。另一方面，基于 TTS 引导的非并行 AC 模型需要大量文本标注对 TTS 模型进行训练，并且性能受限于 TTS 模型本身。为了解决以上问题，本文提出了基于内容特征回归的零样本口音转换模型。首先，考虑到由大规模语音数据自监督训练的 WavLM[15] 模型的第 23 层特征含有丰富的音素信息及少量的说话人音色信息，本文模型采用该特征作为内容特征，一方面保证了内容特征的准确性，另一方面尽可能降低目标口音语音中掺杂的音色信息对源说话人音色合成的影响。其次，通过引入 k-邻近 (k-Nearest Neighbors, kNN)[16] 回归，将源口音语音的 WavLM-23 特征映射为具有相近音素信息的目标口音语音 WavLM-23 特征，从而实现源口音到目标口音的转换。最后，为了实现对源说话人音色的保留，引入说话人编码器并构建多说话人声码器以实现具有源说话人音色和目标口音的语音合成。实验结果表明，该模型在 zero-shot 场景下取得了比并行和非并行 AC 模型更好的客观与主观评价。

1 算法描述

本文提出的模型受到面向 VC 任务的 kNN-VC[17] 模型的启发，并针对 AC 任务进行了如下改进：

(1)考虑到文献 [17] 所采用的 WavLM 第 6 层特征与说话人信息高度相关[18]，为了尽可能避免特征中所包含的说话人特征的影响，本文采用了微调后的 WavLM 模型中与内容信息高度相关的第 23 层特征进行回归。

(2)为了适应 AC 任务对保持源说话人音色的要求，本文单独设计了说话人编码器以提取源说话人音色特征。

(3)为了更好地合成包含源说话人音色和目标口音的转换语音，本文在文献 [17] 采用 HiFi-GAN[19] 架构的基础上设计了多说话人声码器。

本文提出的基于内容特征 kNN 回归的零样本学习 (zero-shot) 口音转换模型 kNN-AC 的总体架构如图 1 所示。它包括内容特征提取、基于 kNN 回归的特征转换、说话人特征提取和基于多说话人声码器的语音合成共 4 个重要部分。

首先，采用基于 WavLM 第 23 层的语音特征编码提取源口音语音和目标口音语音的内容特征；接着，通过 kNN 回归算法，对源口音语音每一帧的内容特征，在目标口音语音特征中查询最邻近的 k 帧特征并计算平均值作为回归结果；最后，构建多说话人声码器，在目标口音内容特征中引入由说话人编码器提取的源说话人音色信息，合成口音转换后的语音。

1.1 内容特征提取

文献 [18] 表明，由 WavLM 第 23 层提取的自监

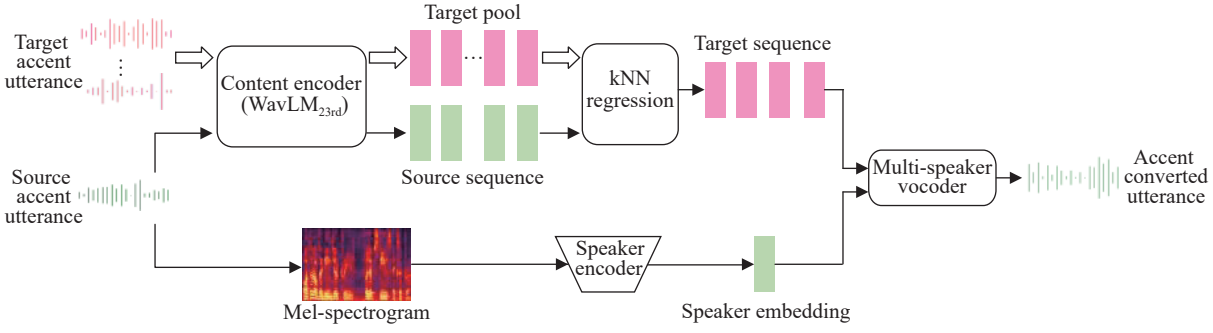

图 1 kNN-AC 模型总体架构

Fig. 1 Architecture of the kNN-AC model

督特征在相同音素之间的相似性远高于不同音素之间的相似性, 更适合与音素特征相关的下游任务, 因此本文采用由 WavLM 第 23 层提取的特征作为内容特征。特别地, 为了使特征中包含的音素信息更为准确, 本文采用了在 ASR 任务上微调过的 WavLM 模型权重。具体微调方式如下: 首先, 通过随机遮掩输入音频并预测被遮掩部分的自监督训练方法获取预训练权重。为了提升语音识别的准确性, 在已标注数据集 LibriSpeech[20] 上利用连接时序分类 (Connectionist Temporal Classification, CTC) 损失函数对预训练模型进行进一步的微调。

在 kNN-AC 模型中, 将从源口音语音中提取出的特征序列作为源序列(source sequence)。同时, 从多个目标口音语音中提取出特征序列作为目标池(target pool)。基于 WavLM 第 23 层的内容特征提取可以表示为式 (1):

$$\begin{cases} \text{source sequence} = \text{WavLM}_{23\text{rd}}(U^{\text{src}}) \\ \text{target pool} = \text{WavLM}_{23\text{rd}}\left(U_1^{\text{trg}}, \cdots, U_n^{\text{trg}}\right) \end{cases} \tag{1}$$

其中, U^{src} 表示一条源口音语音, U_n^{trg} 表示第 n 条目标口音语音, 即目标池由多条目标口音语音提取的特征组成。

1.2　基于 kNN 回归的特征转换

文献 [17] 的实验结果表明, WavLM 模型的不同特征层对于相关性高的音色、韵律等语音属性表现出了很好的聚类效果。然而, 由于口音转换任务中源口音和目标口音之间的发音差距较大, 而文献 [17] 中基于浅层的韵律回归会保留与源口音一致的发音方式, 无法直接用于 AC 任务, 因此需要基于音素特征回归并进行语音重构。为此, 本模型选用由 WavLM 第 23 层提取的特征作为内容特征, 相比于其他层提取的特征, 该层特征在包含音素信息的同时, 也携带少量的发音信息可用于语音重构。

为了实现从源口音到目标口音的转换, 我们对源序列中的每个特征向量进行 kNN 回归, 即对于 source sequence $= \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m\}$ 中的每一帧特征向量 $\boldsymbol{a}_i (1 \leqslant i \leqslant m)$, 分别在 target pool $= \{\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_n\}$ 中查询距离其最近的 k 个特征向量 $\boldsymbol{b}_j (1 \leqslant j \leqslant n)$。这里, 本文模型采用余弦相似度衡量两个特征向量之间的相似性, 计算方式见式 (2)。

$$1 - \cos(\theta_{i,j}) = 1 - \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|} \times \frac{\boldsymbol{b}_j}{\|\boldsymbol{b}_j\|} \tag{2}$$

查询到最邻近的 k 个特征向量 $\boldsymbol{b}_j$ 后, 用它们的平均值替换查询特征向量 $\boldsymbol{a}_i$, 得到目标序列 target sequence $= \{\boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_m\}$。基于内容特征 kNN 回归的口音特征转换的一大优势在于不需要训练任何参数即可实现源口音语音特征到目标口音语音特征的直接转换。

1.3　说话人特征提取

说话人语音中所包含的音色信息作为一种全局特征, 在说话持续期间保持不变。残差网络 (Residual Networks, ResNets)[21] 允许网络在不同层次上整合信息, 在网络深层提取全局特征的同时, 保留了网络浅层的局部特征对全局特征进行补充, 已在大量说话人相关任务[22-24] 上得以验证。因此, 本文模型采用如图 2 所示的由 4 个残差模块 (Residual Blocks, ResBlock) 构成主体的说话人编码器, 从源说话人语音的梅尔谱中提取说话人特征。

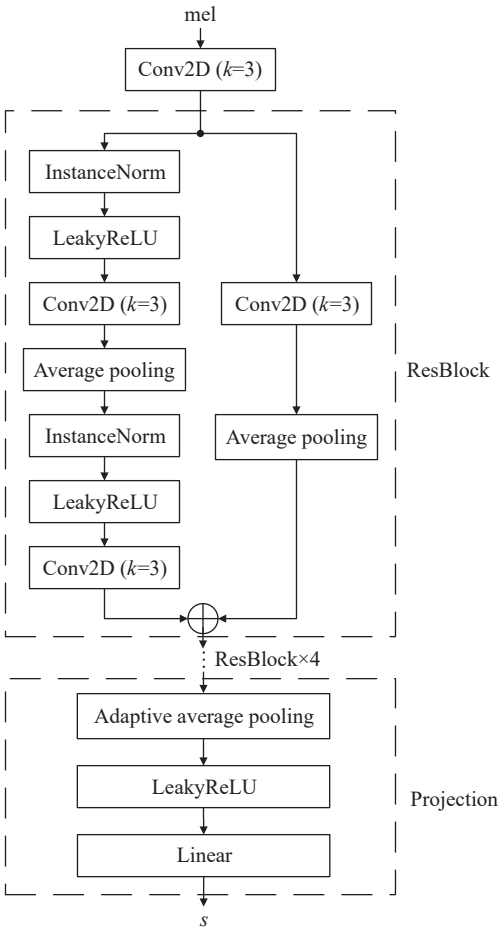

图 2　说话人编码器结构

Fig. 2　Architecture of the speaker encoder

图 2 中所示的每个 ResBlock 在连接前引入了时间维度上的平均池化操作, 旨在减少时间变化对说话人特征不变性的影响。最后通过自适应池化对时间维度进行收缩降维, 并对说话人特征维度进行映射。说话人编码器通过输入梅尔谱 ($\boldsymbol{m}$) 提取出说话人特征 $\boldsymbol{s}$, 其计算过程如式 (3) 所示, 其中: h_0 为第 1 个 ResBlock 的输入, h_{i-1} 为当前 ResBlock 的输入,

h_i 为输出。

$$\begin{cases} h_0 = \mathrm{Conv2D}(\boldsymbol{m}) \\ h_i = \mathrm{ResBlock}(h_{i-1}) \quad (1 \leqslant i \leqslant 4) \\ \boldsymbol{s} = \mathrm{Projection}(h_4) \end{cases} \tag{3}$$

从真实音频的 $\boldsymbol{m}$ 中提取说话人特征 $\boldsymbol{s}$($\boldsymbol{s}=\mathrm{SpeakerEncoder}(\boldsymbol{m})$),从合成语音梅尔谱 ($\widehat{\boldsymbol{m}}$) 中提取说话人特征 $\widehat{\boldsymbol{s}}$($\widehat{\boldsymbol{s}}=\mathrm{SpeakerEncoder}(\widehat{\boldsymbol{m}})$), $\boldsymbol{s}$ 和 $\widehat{\boldsymbol{s}}$ 的说话人循环一致性损失记为 L_{cyc} (式(4))。说话人编码器后续将与声码器一起进行训练。

$$L_{\mathrm{cyc}} = \left\| \boldsymbol{s} - \widehat{\boldsymbol{s}} \right\|_1 \tag{4}$$

1.4 基于多说话人声码器的语音合成

多说话人声码器的架构如图 3 所示。由于目标序列的每一帧特征是由目标池中的 k 个邻近特征取平均值得到的,因此源说话人的音色信息随之转换为 target pool 特征携带的少量说话人信息。为了在合成语音中保持源说话人音色,本文模型在 HiFi-GAN[19] 声码器架构的基础上引入了自适应实例归一化 (Adaptive Instance Normlization, AdaIN)[25] 以及相应的残差模块 AdaIN ResBlock。AdaIN ResBlock 与前述 ResBlock的主要区别在于归一化方式有所不同。与 ResBlock采用的实例归一化 (Instance Normlization, IN)[26] 相比,AdaIN 在内容特征的基础上引入了风格特征进行融合,如式 (5) 所示:

$$\mathrm{AdaIN}(\boldsymbol{c},\boldsymbol{s}) = L_{\sigma}(\boldsymbol{s})\frac{\boldsymbol{c}-\mu(\boldsymbol{c})}{\sigma(\boldsymbol{c})} + L_{\mu}(\boldsymbol{s}) \tag{5}$$

其中,$\boldsymbol{c}$ 表示输入的语音内容特征, $\mu(\boldsymbol{c})$ 和 $\sigma(\boldsymbol{c})$ 分别表示内容特征的均值和标准差, $L_{\sigma}(\boldsymbol{s})$ 和 $L_{\mu}(\boldsymbol{s})$ 分别表示基于说话人特征 $\boldsymbol{s}$ 学习的线性变换计算的自适应增益和偏差。

通过 AdaIN ResBlock 在目标序列中引入说话人信息,并直接通过生成器(Generator)进行语音合成,训练过程中采用判别器对声音质量进行优化。

对于真实语音 $\boldsymbol{x}$ 和生成语音 $\widehat{\boldsymbol{x}}$,判别器 D 和生成器 G 之间的对抗训练损失 $L_{\mathrm{adv}}(D;G)$ 和 $L_{\mathrm{adv}}(G;D)$ 可分别表示为式 (6) 和式 (7):

$$L_{\mathrm{adv}}(D;G) = (D(\boldsymbol{x})-1)^2 + (D(\widehat{\boldsymbol{x}}))^2 \tag{6}$$

$$L_{\mathrm{adv}}(G;D) = (D(\widehat{\boldsymbol{x}}))^2 \tag{7}$$

为了稳定训练并加速收敛,对于真实语音 $\boldsymbol{x}$ 对应的 $\boldsymbol{m}$ 和生成语音 $\widehat{\boldsymbol{x}}$ 对应的 $\widehat{\boldsymbol{m}}$,进一步采用式 (8) 和式 (9) 中的梅尔损失 (L_{mel}) 和特征匹配损失 (L_{FM}) 进行约束。其中特征匹配损失通过计算判别器中真实语音与生成语音每一层的平均绝对误差来保持真实语音与生成语音间的一致性。

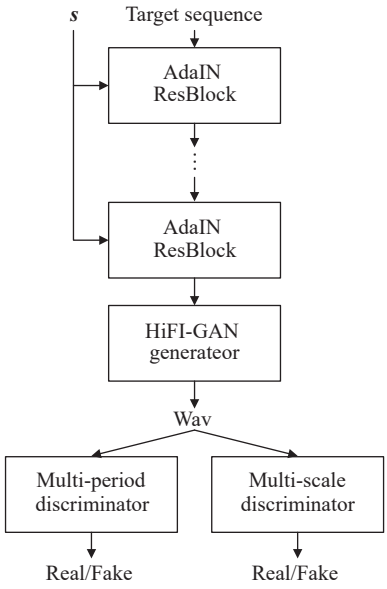

图 3 多说话人声码器架构

Fig. 3 Architecture of the multi-speaker vocoder

$$L_{\mathrm{mel}}(G) = \|\boldsymbol{m} - \widehat{\boldsymbol{m}}\|_1 \tag{8}$$

$$L_{\mathrm{FM}}(G;D) = \sum_{i=1}^{T}\frac{1}{N_i}\|D^i(\boldsymbol{x}) - D^i(\widehat{\boldsymbol{x}})\|_1 \tag{9}$$

最终,整个多说话人声码器的训练损失可分为生成器损失 (L_G) 和判别器损失 (L_D),分别如式 (10) 和式 (11) 所示,其中 λ 表示计算损失所赋予的权重系数。

$$L_G = L_{\mathrm{adv}}(G;D) + \lambda_{\mathrm{fm}}L_{\mathrm{FM}} + \lambda_{\mathrm{mel}}L_{\mathrm{mel}} \tag{10}$$

$$L_D = L_{\mathrm{adv}}(D;G) \tag{11}$$

2 实验结果与分析

2.1 数据集

实验所涉及的数据集包括 LibriSpeech[20], L2-Arctic[27], Arctic[28] 以及 IndicTTS[29]。其中 LibriSpeech 作为目标口音语音数据集,用于 kNN-AC 中说话人编码器和多说话人声码器的训练。Arctic 包含具有标准英文口音的语音数据,而 L2-Arctic 包含具有与 Arctic 语音内容相同而口音不同的并行语音数据。Arctic 和 L2-Arctic 常用于并行 AC 模型的训练。需要说明的是,首先,为了测试模型对不同口音语音的口音转换效果,选用 L2-Arctic 中的印度口音和阿拉伯口音语音作为测试对象,并且所有基线模型均采用与 kNN-AC 相同的数据进行性能测试。此外,我们还在 IndicTTS 数据集上对印度口音转换进行了测试,进一步验证本文模型的泛化性。需要明确的是,kNN-AC 的口音转换性能是在零样本学习的

条件下测试, 即在模型训练阶段没有采用任何源口音语音数据。

2.2　基线模型与实验设置

为了验证 kNN-AC 模型的性能, 实验分别选取了基于 PPG 特征的并行模型 PPG-AC[8], 基于 TTS 的非并行 AC 模型 ASR-TTS[9], 非并行多口音转换 AC 模型 Multi-AC[12], 以及基于内容特征回归的 VC 模型 kNN-VC[17] 作为基线模型, 并将各模型的转换语音与源语音数据进行比较。在各个数据集上, 本文模型的测试数据均与所有基线模型的测试数据保持一致, 各模型设置分别如下:

(1) PPG-AC[8]: 采用并行的 Arctic 和 L2-Arctic 语音数据进行训练, 对每个说话人, 训练数据包括除了最后 100 条语音的 1 032 条语音。

(2) ASR-TTS[9]: 考虑到基于 TTS 的非并行 AC 模型均未公开代码, 本文分别利用了基于 Whisper[30] 的 ASR 模型和基于 FastSpeech2[31] 的 TTS 模型复现文献 [9] 的架构。测试时通过 Whisper 识别源口音语音文本并通过 FastSpeech2 合成目标口音语音。

(3) Multi-AC[12]: 可通过口音标签实现不同口音之间的转换。我们对该模型提供的基于 IndicTTS 数据集实现的印度口音英语转换为标准口音英语语音进行比较测试。

(4) kNN-VC[17]: 考虑到 kNN-AC 是受到 kNN-VC 模型的原理启发并适应 AC 任务的具体需要而构建的, 本文也采用了 kNN-VC 模型在 AC 任务上的结果作为测试对象, 实验设置与 kNN-AC 保持一致, 即在 LibriSpeech 上训练目标口音合成, 并在 L2-Arctic 上测试口音转换性能。并且, kNN 特征回归均设置 k=4 以及相同 25 min 的目标池进行推理。训练过程中 batch size 设置为 8, 采用 AdamW[32] 优化器, 初始学习率设置为 2×10^{-4}, 衰减率为 10^{-3}。

2.3　与基线模型的性能比较

本文分别采用客观评价和主观评价对 AC 模型的性能进行比较。

客观评价包括与语音可懂度有关的词错率 (Word Error Rate, WER) 和字错率 (Character Error Rate, CER), 以及说话人相似度识别的平均错误率 (Equal Error Rate, EER)。客观指标实验通过 ESPnet[33] 提供的 ASR 模型进行测定, 并采用 Resemblyzer[34] 测定 EER。

主观评价包括语音自然度 (Mean Opinion Score, MOS) 和口音度 (Accentedness)。主观评价实验随机挑选 10 条语音, 提供配对文本并请 10 名受试者进行打分。打分均采用 5 分制 (1~5), MOS 的打分越高表示语音质量越好; 口音度的打分越低, 表示口音转换效果越好。kNN-AC 模型的口音转换 demo 见 https://chiaki-luo.github.io/knnac/。

kNN-AC 模型与各基线模型的性能对比结果如表 1 所示, 其中"↓"表示数值越低越好, "↑"表示数值越高越好。客观性能对比结果表明, 首先, 在 L2-Arctic 数据集上, 与源口音语音相比, 各种 AC 模型均可以降低 WER/CER 指标, 表明各模型均可在一定程度上将源口音转换为目标口音。其次, 用于 VC 任务的 kNN-VC 模型不能达到任意一种 AC 模型的口音转换性能。最后, kNN-AC 取得了比并行模型 PPG-AC 和基于 TTS 的 ASR-TTS 模型更低的 WER/CER, 表

表 1　本文模型与基线模型的性能对比

Table 1　Performance comparison between kNN-AC and the baseline models

Dataset	Baseline	WER/% (↓)	CER/% (↓)	EER (Threshold)/% (↓)	Score	
					MOS (↑)	Accentedness(↓)
L2-Arctic	Ground truth	20.01	12.12	— (—)	4.76±0.05	4.71±0.04
	PPG-AC [8]	13.31	8.18	10.00(0.65)	3.63±0.09	2.48±0.14
	ASR-TTS [9]	14.52	6.89	10.00(0.67)	3.25±0.12	1.90±0.11
	kNN-VC [17]	18.71	10.68	8.33(0.69)	**3.99±0.13**	3.74±0.12
	kNN-AC	**12.32**	**6.83**	**6.67(0.61)**	3.77±0.15	**1.67±0.11**
IndicTTS	Ground truth	15.08	5.46	— (—)	4.61±0.08	4.50±0.12
	Multi-AC [12]	10.39	3.03	**6.55(0.64)**	4.41±0.07	1.93±0.11
	ASR-TTS [9]	10.38	4.00	10.00(0.67)	3.37±0.11	1.58±0.07
	kNN-VC [17]	12.96	3.13	10.00(0.62)	4.28±0.08	3.44±0.13
	kNN-AC	**9.05**	**2.19**	7.00(0.65)	**4.44±0.06**	**1.29±0.08**

明 kNN-AC 取得了更好的口音转换性能。另外，在 IndicTTS 数据集上，kNN-AC 转换语音的可懂度也超过了包含 Multi-AC 在内的其他 AC 模型，但说话人编码器对音色特征的提取在该数据集上的表现略有下降，使得 EER 略有升高。值得一提的是，与其他 AC 模型不同，kNN-AC 的性能是在训练阶段未见源口音语音数据的情况下达成的，因此 kNN-AC 具有更好的泛化性。

主观性能对比结果表明，首先所有 AC 模型均在一定程度上影响了语音自然度，但相对而言 kNN-AC 的影响最小，然而说话人特征的引入使得 kNN-AC 的声码器相对于 kNN-VC 的声码器更难训练，从而降低了语音质量。其次，kNN-VC 对于口音度几乎没有改善，而 kNN-AC 的改进显著降低了口音度并取得了比 PPG-AC、ASR-TTS 以及 Multi-AC 模型更好的性能表现。

2.4 目标语音时长对模型性能的影响

为了测试构建 target pool 的语音时长对 kNN-AC 转换语音的可懂度和音色保持度的影响，本实验对在不同时长下所获得的 WER、CER 及 EER 进行测试，结果如图 4 所示。可以看出：(1)当目标语音时长大于 3 min 时，由 kNN-AC 转换的语音能够获得比源口音语音更高的可懂度。(2)随着目标语音时长增加，转换语音的可懂度明显上升。可能的原因是，时长的增加使目标池中所包含的音素更多，包含与源语音音素更相似的语音因素的可能性更大，使得语音内容回归的准确性随之上升，从而有利于提升口音转换性能。(3)与文献 [17] 不同，目标语音时长基本不影响 EER 指标，这说明说话人编码器可准确提取源说话人信息，同时多说话人声码器有效融合了源说话人信息，从而保留了转换语音中源说话人的音色。

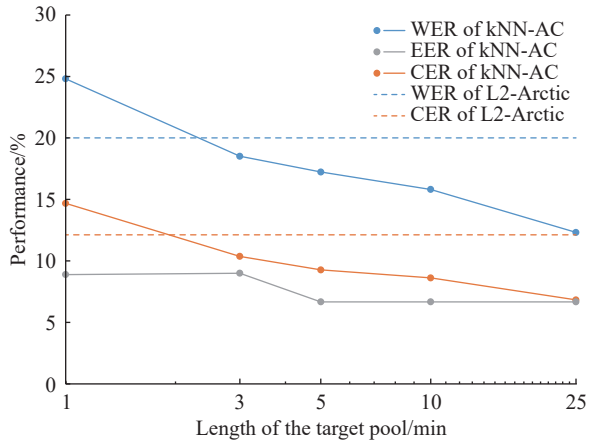

图 4 用于构建 target pool 的语音时长对模型性能的影响

Fig. 4 Influence of the length of speech adopted for the construction of the target pool on the performance of the proposed model

3 结 论

为了解决并行 AC 模型对数据采集的依赖，以及非并行 AC 模型泛化性的问题，本文提出一种 zero-shot AC 模型。采用 WavLM 第 23 层提取内容特征以保证内容准确性，通过对源口音语音内容特征进行 kNN 回归实现口音特征向目标口音的转换。最后，构建说话人编码器提取源说话人音色特征并通过多说话人声码器实现具有源说话人音色的目标口音语音合成。实验结果表明，该模型取得了比并行和非并行 AC 模型更好的性能。

参考文献：

[1] TÜRK O, ARSLAN L M. Subband based voice conversion[C]// Seventh International Conference on Spoken Language Processing. Denver, Colorado, USA: ICSLP, 2002: 137-140.

[2] OSHIMA Y, TAKAMICHI S, TODA T, *et al*. Non-native speech synthesis preserving speaker individuality based on partial correction of prosodic and phonetic characteristics[C]// Proceedings of Interspeech. Dresden, Germany: Interspeech, 2015: 299-303.

[3] YANG L F, FU K Q, ZHANG J S, *et al*. Non-native acoustic modeling for mispronunciation verification based on language adversarial representation learning[J]. Neural Networks, 2021, 142: 597-607.

[4] ZHAO G L, SONSAAT S, LEVIS J, *et al*. Accent conversion using phonetic posteriorgrams[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5314-5318.

[5] ZHAO G L, DING S J, GUTIERREZ-OSUNA R, *et al*. Foreign accent conversion by synthesizing speech from phonetic posteriorgrams[C]//Proceedings of Interspeech. Graz, Austria: Interspeech, 2019: 2843-2847.

[6] LI W J, TANG B L, YIN X, *et al*. Improving accent conversion with reference encoder and end-to-end text-to-speech [EB/OL]. (2020-05-19) [2024-10-10]. https://arxiv.org/abs/2005.09271.

[7] QUAMER W, DAS A, LEVIS J, *et al*. Zero-shot foreign accent conversion without a native reference[C]//Proceedings of Interspeech. Incheon, ROK: Interspeech, 2022: 4920-4924.

[8] HUANG W C, TODA T. Evaluating methods for ground-

truth-free foreign accent conversion[C]//2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference. Taipei, Taiwan, China: APSIPA ASC, 2023: 1161-1166.

[9] LIU S X, WANG D S, CAO Y W, *et al*. End-to-end accent conversion without using native utterances[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 6289-6293.

[10] ZHOU Y, WU Z Z, ZHANG M Y. TTS-guided training for accent conversion without parallel data[J]. IEEE Signal Processing Letters, 2023, 30: 533-537.

[11] CHEN X, PEI J K, XUE L M, *et al*. Transfer the linguistic representations from TTS to accent conversion with non-parallel data[C]//2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Coex, Seoul, ROK: IEEE, 2024: 12501-12505.

[12] JIN M M, SERAI P, WU J L *et al*. Voice-preserving zero-shot multiple accent conversion[C]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE, 2023: 1-5.

[13] TAN D X, DENG L Q, YEUNG Y T, *et al*. Editspeech: A text based speech editing system using partial inference and bidirectional fusion[C]// 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Virtual: IEEE, 2021: 626-623.

[14] TAN D X, DENG L Q, ZHENG N Z, *et al*. Correctspeech: A fully automated system for speech correction and accent reduction[C]//2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). Singapore: IEEE, 2022: 81-85.

[15] CHEN S Y, WANG C Y, CHEN Z Y, *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.

[16] FIX E, HODGES J L. Discriminatory analysis, nonparametric discrimination: Consistency properties[J]. International Statistical Institute, 1989, 57(3): 238-247.

[17] BAAS M, VAN NIEKERK B, KAMPER H. Voice conversion with just nearest neighbors[C]// Proceedings of Interspeech. Dublin, Ireland: Interspeech, 2023: 2053-2057.

[18] DUNBAR E, HAMILAKIS N, DUPOUX E. Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1211-1226.

[19] KONG J, KIM J, BAE J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis[J]. Advances in Neural Information Processing Systems, 2020, 33: 17022-17033.

[20] PANAYOTOV V, CHEN G, POVEY D, *et al*. Librispeech: An ASR corpus based on public domain audio books[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane, QLD, Australia: IEEE, 2015: 5206-5210.

[21] HE K M, ZHANG X Y, REN S Q, *et al*. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Juan, PR, USA: IEEE, 2016: 770-778.

[22] LI Y A, HAN C, MESGARANI N. Styletts-vc: One-shot voice conversion by knowledge transfer from style-based tts models[C]//2022 IEEE Spoken Language Technology Workshop (SLT). Doha, Qatar: IEEE, 2023: 920-927.

[23] JAKUBEC M, LIESKOVSKA E, JARINA R. Speaker recognition with resNet and VGG networks[C]//31st International Conference Radioelektronika (RADIOELEKTRONIKA). Brno, Czech Republic: IEEE, 2021: 1-5.

[24] JAKUBEC M, JARINA R, LIESKOVSKÁ E, *et al*. Deep speaker embeddings for Speaker Verification: Review and experimental comparison[J]. Engineering Applications of Artificial Intelligence, 2024, 127: 107232.

[25] HUANG X, BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 1510-1519.

[26] DUMOULIN V, SHLENS J, KUDLUR M. A learned representation for artistic style[C]//International Conference on Learning Representations. Toulon, France: ICLR, 2017.

[27] ZHAO G L, SONSAAT S, SILPACHAI A, *et al*. L2-arctic: A non-native English speech corpus[C]//Proceedings of Interspeech. Hyderabad, India: Interspeech, 2018: 2783-2787.

[28] KOMINEK J, BLACK A W. The CMU arctic speech databases[C]//Speech Synthesis Workshop. Pittsburgh, PA, USA: ISCA, 2004: 223-224.

[29] BABY A, THOMAS A L, NISHANTHI N L, *et al*. Resources for indian languages[C]//Community-Based Building of Language Resources. Brno, Czech Republic: Tribun EU, 2016: 37-43.

[30] RADFORD A, KIM J W, XU T, *et al*. Robust speech recognition via large-scale weak supervision[C]//International Conference on Machine Learning. Honolulu, Hawaii, USA: PMLR, 2023: 28492-28518.

[31] REN Y, HU C X, TAN X, *et al*. FastSpeech 2: Fast and high-quality end-to-end text to speech[C]//International Conference on Learning Representations. Virtual Event, Austria: ICLR, 2021.

[32] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C]//International Conference on Learning

Representations. New Orleans, LA, USA: ICLR, 2019.

[33] HAYASHI T, YAMAMOTO R, INOUE K, *et al*. Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 7654-7658.

[34] WAN L, WANG Q, PAPIR A, *et al*. Generalized end-to-end loss for speaker verification[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 4879-4883.

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

LUO Yixin[1], CHEN Ning[1], XUE Yuhang[1], XIAO Yangyang[2]

(1. School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China; 2. China Telecom Wanwei Information Technology Co. Ltd, Lanzhou 730000, China)

Abstract: Accent Conversion (AC) aims to convert speech from the source accent to the target accent while preserving the source speaker's timbre and the speech content at the same time. Existing AC models cannot achieve good generalization capability for AC on speech that does not follow the distribution of the training data, as limits their applications seriously. To this end, a zero-shot AC model based on the kNN regression of speech content features is proposed. On the one hand, the 23rd layer of WavLM is adopted as the content encoder to extract the content features from both source and target accented speech, and kNN regression is employed to replace the source accented content feature with its nearest neighbors in the pool constructed by the target accented content features to achieve accent conversion. On the other hand, to preserve the source speaker's timbre in the converted speech, a multi-speaker vocoder is constructed to fuse the obtained target accented content features with the source speaker's timbre feature extracted by the speaker encoder to synthesize the speech with the target accent. In the proposed model, no source accented speech is required at the training stage, so it can convert various source accented speech to the target accented speech. That is, the proposed model achieves good generalization ability. Experimental results demonstrate that the proposed model achieves better objective and subjective evaluation results than available parallel or non-parallel AC models.

Key words: accent conversion; kNN regression; zero-shot learning; voice conversion; vocoder

（责任编辑：张欣）