引入预训练表示混合矢量量化和 CTC 的语音转换

王琳，黄浩

（新疆大学信息科学与工程学院，新疆 乌鲁木齐 830017）

摘　要：预训练模型通过自监督学习表示在非平行语料语音转换（VC）取得了重大突破。随着自监督预训练表示（SSPR）的广泛使用，预训练模型提取的特征中被证实包含更多的内容信息。提出一种基于 SSPR 同时结合矢量量化（VQ）和联结时序分类（CTC）的 VC 模型。将预训练模型提取的 SSPR 作为端到端模型的输入，用于提高单次语音转换质量。如何有效地解耦内容表示和说话人表示成为语音转换中的关键问题。使用 SSPR 作为初步的内容信息，采用 VQ 从语音中解耦内容和说话人表示。然而，仅使用 VQ 只能将内容信息离散化，很难将纯粹的内容表示从语音中分离出来，为了进一步消除内容信息中说话人的不变信息，提出 CTC 损失指导内容编码器。CTC 不仅作为辅助网络加快模型收敛，同时其额外的文本监督可以与 VQ 联合优化，实现性能互补，学习纯内容表示。说话人表示采用风格嵌入学习，2 种表示作为系统的输入进行语音转换。在开源的 CMU 数据集和 VCTK 语料库对所提的方法进行评估，实验结果表明，该方法在客观上的梅尔倒谱失真（MCD）达到 8.896 dB，在主观上的语音自然度平均意见分数（MOS）和说话人相似度 MOS 分别为 3.29 和 3.22，均优于基线模型，此方法在语音转换的质量和说话人相似度上能够获得最佳性能。

关键词：预训练表示；自监督学习；矢量量化；解耦；联结时序分类

源代码链接：https://github.com/wanglin-1998/SSPR-CTC-VQVC.git

中图分类号：TP391　　**文献标志码**：A　　**DOI**：10.19678/j.issn.1000-3428.0067272

Voice Conversion Combining Vector Quantization and CTC Introducing Pre-Trained Representation

WANG Lin，HUANG Hao

（School of Information Science and Engineering，Xinjiang University，Urumqi 830017，Xinjiang，China）

【Abstract】 Pre-trained models have achieved significant breakthroughs in nonparallel corpus Voice Conversion（VC）via Self-Supervised Pre-trained Representation（SSPR）. Features extracted by pre-trained models contain a significant amount of content information owing to the widespread use of SSPR. This study proposes a VC model based on the combination of SSPR Vector Quantization（VQ）and Connectionist Temporal Classification（CTC）. It uses the SSPR extracted from a pre-trained model as input to improve the quality of single VC. The effective decoupling of content and speaker representations has become a key issue in VC. Using SSPR as the initial content information，VQ is performed to decouple content and speaker representations from speech. However，performing only VQ discretizes only the content information，thus rendering it difficult to separate pure content representations from speech. To further eliminate speaker-invariant information from the content information，a CTC loss-guided content encoder is proposed. CTC not only serves as an auxiliary network to accelerate model convergence but also its additional text supervision can be jointly optimized with VQ to achieve complementary performance and learn pure-content representations. Speaker representations adopt style-embedding learning，and two representations are used as inputs for VC in the system. The proposed method is evaluated on the open-source CMU dataset and VCTK corpus. Experimental results show that the proposed method achieves an objective Mel-Cepstrum Distortion（MCD）of 8.896 dB，as well as subjective Mean Opinion Score（MOS）of speech naturalness and speaker similarity of 3.29 and 3.22，respectively，both of which are better than those of the baseline model. This method achieves the best performance in terms of VC quality and speaker similarity.

【Key words】 pre-trained representation；Self-Supervised Learning（SSL）；Vector Quantization（VQ）；decouple；Connectionist Temporal Classification（CTC）

0　引言

语音转换（VC）旨在改变说话人身份而不改变说话内容的过程。根据训练数据的来源和类型可分为平行 VC 和非平行 VC。平行 VC 利用转换模型学习声级之间的映射关系从源说话人到目标说话人

收稿日期：2023-03-27　　**修回日期**：2023-06-25

基金项目：新疆维吾尔自治区重点实验室开放课题（2020D04047）。

通信作者 E-mail：wngln0722@163.com

的帧。一般使用的模型包括高斯混合模型[1]、深度神经网络[2]、循环神经网络[3]等。

实际上,收集这种平行语音语料库可能很困难。因此,研究工作越来越集中在非平行 VC 上。语音后验图(PPG)[4]是非平行 VC 的典型早期方法。从非特定人自动语音识别器中提取 PPG 特征帧级语言表示。在实际应用中,PPG 特征仍然包含一些说话人信息,往往会导致在转换阶段与预期说话人相似性方面的性能不令人满意。为了提高说话人相似度方面的性能,生成对抗网络(GAN)的变体通过应用鉴别器来模仿人类的听觉感知,如 CycleGAN[5]和 StarGAN[6]。然而,这些基于 GAN 的模型通常仍然难以训练。

变分自编码器(VAE)是这些方法针对非平行 VC 提出的易于训练的常用模型。基于 VAE 的 VC 通常采用编码器-解码器框架来学习内容和说话人表征的解纠缠。在传统的基于 VAE 的 VC[7-8]中,编码器从输入声学特征中提取说话人独立的隐变量,解码器从隐变量和给定的说话人表示中重建声学特征。然而,VAE 的建模局限于假设的高斯分布,而实际数据的分布可能要复杂得多。在 VAE 的基础上,通过对变分自编码器进行矢量量化(VQ)形成 VQ-VAE[9]以有效解决这一问题。VQ-VAE 是由 DeepMind 开发的一种强大的生成模型,也被应用于非平行 VC。与 VAE 相比,基于 VQVC 的优势在于它能够通过嵌入字典学习先验知识,而不是使用预定义的静态分布,如高斯分布。矢量量化中的嵌入字典学习完全表示不同的语音内容。虽然 VQVC 已被证明是一种很有前途的内容和说话者表示解纠缠的方法,但是现有基于 VQVC 的方法仍然存在一定局限性。在模型训练过程中主要选择梅尔语谱作为输入声学特征,依赖于声学特征的自构误差,不可避免地存在一些信息缺失。由于缺乏语言监督,因此内容表征很难被解开。为解决 VQVC 中存在的问题,本文提出 2 种策略:1)选用自监督预训练表示代替传统的梅尔语谱学习内容表征;2)使用联结时序分类(CTC)[10]文本监督辅助网络提纯内容信息。

最近,自监督学习(SSL)在语音处理方面也取得了显著的成果,基于 SSL 的表征可用于音素识别[11]和说话人分类[12-14],这表明 SSL 表示中包含的语音和说话人信息都是遗传的。因此,自监督预训练表示(SSPR)被越来越多应用于语音转换任务中。之前的一些工作试图将 SSPR 引入到 VC 任务。例如,FragmentVC[15]是 1 个 any-to-any VC 模型,它可以将任意源说话人的语音转换为任何目标说话人,甚至是在训练期间未见过的说话人。FragmentVC 从 SSPR 中提取语音信息和目标说话人信息,但它需要额外的交叉注意力模块来对齐源特征和目标特征。

本文将基于自监督预训练模型提取的特征作为系统最初的内容嵌入,送入内容编码器提取内容表征。模型经过内容编码器的输出将被 VQ 量化为嵌入词典,表示语言信息。解码器将量化的内容表示和说话人表征预测声学特征。VQVC 难以将只包含内容信息的表征与自监督预训练表示中声学特征分离开来,为了解决 VQVC 存在的问题,本文设计 CTC 辅助网络。通过引入强文本监督对内容编码器输出的表示进行建模,以提高模型的解缠能力。CTC 已成为当前端到端自动语音识别(ASR)任务中广泛使用的训练目标函数[16-17]。CTC 辅助网络能够引导内容编码器学习更纯的内容表示。CTC loss 的好处是让内容编码器的输出不受序列长度的影响。此外,它还考虑了整个特征序列中的全局信息。本文提出使用自监督预训练表征混合 VQ 和 CTC 算法的非平行语音转换,模型被命名为 SSPR-CTC-VQVC。

1 相关工作

1.1 SSPR 学习

受预训练 SSL 表示和瓶颈特征的启发,越来越多的文章使用 SSPR 作为初始语言信息,并将其输入到语音转换系统中。内容编码器通过限制瓶颈特征最终获得与说话人相关的内容信息。将预训练模型 WavLM[18]作为语音转换系统初始的内容嵌入。WavLM 模型结构如图 1 所示,模型的结构与文献[18]的结构相同。

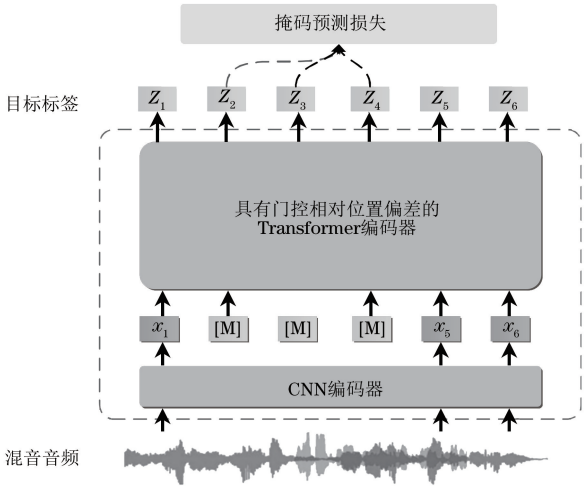

图 1 预训练模型 WavLM 结构

Fig.1 Structure of pre-trained model WavLM

本文采用自监督预训练模型 WavLM 通过 Transformer 的最后一层输出作为内容编码器系统输入，用来提取初阶的内容表示。

1.2 矢量量化

VQVC 通过嵌入字典学习先验来表示完全不同的语音内容。矢量量化器在矢量量化嵌入字典的帮助下，将从内容编码器输出的连续信息编码为一组离散的嵌入。文献[19]等工作表明，VQ 嵌入字典中的每个嵌入都与语音音素有密切的关系，即 VQ 嵌入字典具有对音素集建模的能力。VQ-VAE 结构如图 2 所示。嵌入字典正式写为 $A=[a_1,a_2,\cdots,a_N]$，$A\in\mathbb{R}^{N\times D}$，有 N 个不同的码本。在前向计算时，由 SSPR 得到的声学特征 x 送入内容编码器中获得 $z_e(x)$，映射到码本中最近嵌入 a_i 来离散化。$z(x)$和说话人嵌入 s 将连接在一起送入解码器。最后，利用解码器 D 对声学特征进行重构 $\hat{x}=D(z(x),s)$。VQ 训练的目标函数可以表示为：

$$L_{\mathrm{VQ}}=\|\mathrm{sg}[z_e(x)]-a_i\|_2^2+\gamma\|z_e(x)-\mathrm{sg}[a_i]\|_2^2 \tag{1}$$

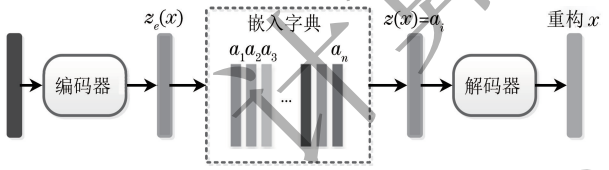

图 2 VQ-VAE 结构

Fig.2 The structure of VQ-VAE

在矢量量化的损失中添加承诺成本，以鼓励每个 $z_e(x)$向所选码本靠近[19]。

1.3 联结时序分类

联结时序分类是一种损失函数，常用于解决输入特征与输出标签之间对齐不确定的时间序列映射问题。对于给定的由矢量量化得到的 $z(x)$ 嵌入，其输入序列记为 $I=[i_1,i_2,\cdots,i_T]$，其中 T 是输入序列的长度。CTC 损失函数的任务是最大化概率分布 $P(m\mid I)$ 为对应的音素标号序列 m 的长度。实验中使用的 2 个语料库均为英文，选择开源的 CMU 英语词典和音素集，CMU 音素集有 39 个音素，加上 CTC 的额外空白标签，音素数为 40。CTC 损失函数定义为：

$$L_{\mathrm{CTC}}=P_{\max}(m\mid I)=\sum_{\omega\in\varphi(m)}P(\omega\mid I) \tag{2}$$

CTC 辅助网络如图 3 所示，包含 3 个 Conv1d 层、1 个双向长短时记忆(BLSTM)层和 1 个线性层。每个 Conv1d 层与批量归一化和 ReLU 激活层相结合，将 BLSTM 层和线性层的单元数分别设置为 128 和 40。

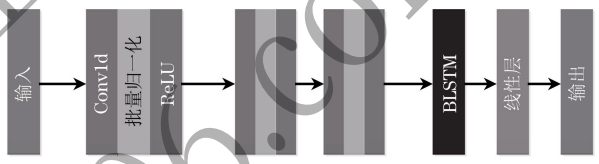

图 3 联结时序分类辅助网络结构

Fig.3 Structure of connectionist temporal classification auxiliary network

2 模型架构

本文提出的 SSPR-CTC-VQVC 包括 1 个内容编码器、1 个说话人编码器、1 个解码器和声码器。基于 WavLM 的 SSPR 送入内容编码器来提取内容表示。说话人嵌入信息是从说话人编码器得到的。VQ 将内容表示离散化，CTC 辅助网络用于去除内容表示中的说话人信息。将内容表示和说话人表示连接后送到解码器中以获得重构表示，并合成语音。SSPR-CTC-VQVC 模型总体架构如图 4 所示。

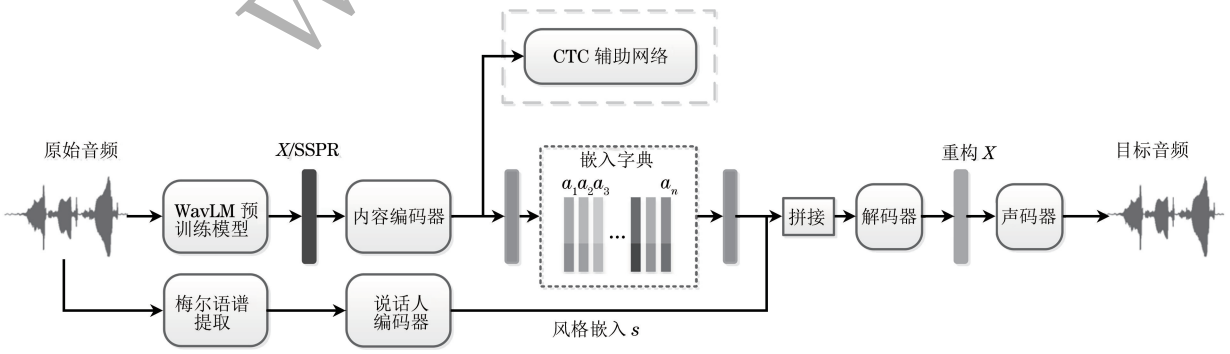

图 4 SSPR-CTC-VQVC 模型结构

Fig.4 The structure of SSPR-CTC-VQVC model

2.1 内容编码器

实例归一化(IN)[20]可以很好地分离说话人和内容表示。与文献[20]不同，本文用 4 个卷积块改进了其内容编码器，比改进前的多 1 个卷积块和 1 个瓶颈线性层。在内容编码器中使用实例归一化层对全局信息进行归一化，有效地提取内容表示。内容表示由 4 个卷积块和 1 个瓶颈线性层生成。每个卷积块包含 3 个具有跳跃连接的 Conv1d 层，每个

Conv1d 层与实例归一化和 ReLU 激活函数相结合。Conv1d 层的核大小、通道、步长分别设置为 5、256 和 1。内容编码器将 WavLM 学习的 SSPR 作为初步提取内容表示的输入，其结构如图 5 所示。

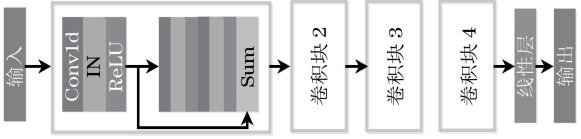

图 5 内容编码器结构

Fig.5 Structure of content encoder

2.2 说话人编码器

说话人编码器的目标是为不同的话语生成来自同一说话人相同的嵌入，但不同说话人的嵌入不同。一般来说，1 个 one-hot 编码的说话人身份嵌入被用于 many-to-many VC 中，但实际的 VC 编码器需要产生可泛化到训练过程中 unseen 的说话人嵌入。

本文使用全局风格标记(GST)[21]作为也被广泛使用的语音合成说话人编码器。GST 能够将变长声学特征序列压缩为固定维说话人风格嵌入并隐式学习说话人嵌入风格标签。说话人编码器结构如图 6 所示。GST 提取说话人风格嵌入需要 2 个步骤：

1)参考编码器采用连续堆叠的方式对声学特征序列进行压缩 Conv2d 层，然后将双向门控循环单元(Bi-GRU)层的最后 1 个状态作为输出参考编码器。将变长声学特征序列转化为固定长度的嵌入，称为参考嵌入。

2)将引用嵌入传递到样式标记层。在样式标记层中，有几个等长的样式标记(例如 A、B、C、D)被初始化以模拟不同组件语音(例如情感、节奏、音高、语速等)。多头自注意力模块学习参考嵌入和每个样式 Token。注意力模块输出 1 组组合权重，表示每个样式标记对引用嵌入的贡献。样式标记的加权和命名为风格嵌入。

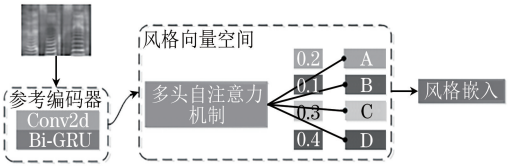

图 6 说话人编码器结构

Fig.6 Structure of speaker encoder

2.3 解码器

解码器将所有编码器的输出作为输入，重构表示，并将解码后的表示提供给重构的声码器。在实现细节上，实验中使用的网络架构如图 7 所示。解码器包含 3 个 Conv1d 层，1 个 BLSTM 层和 1 个线性层。每个 Conv1d 层的内核大小为 5，通道数为 512，步长为 1。BLSTM 层的单元数为 512。

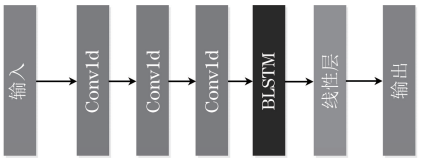

图 7 解码器结构

Fig.7 Decoder structure

2.4 多任务损失

使用 CTC 的 VQVC 与传统 VQVC 的相似之处在于两者都采用内容编码器和说话人编码器分别提取内容和说话人表示。区别在于内容编码器的输出是否受到 CTC 辅助网络 loss 的监督。因此，本文提出的模型成为 1 个多任务学习框架，模型训练需要减少自重建误差、VQ loss 和 CTC loss。总损失函数可以写成：

$$L = L_{\mathrm{MSE}}(x,\hat{x}) + \alpha L_{\mathrm{VQ}} + \beta L_{\mathrm{CTC}} \tag{3}$$

其中：x 和 $\hat{x}$ 分别代表输入的 SSPR 和重构的 SSPR；L_{MSE} 是 x 和 $\hat{x}$ 之间的均方误差(MSE)；α 和 β 分别是 L_{VQ} 和 L_{CTC} 的超参数，α 设置为 1，β 随着训练步数的增加线性增加，在 5×10^4 步数达到最大值 0.001 后保持固定。权重递增策略能够在初始训练阶段保持模型稳定。模型的训练步数设置为 3×10^5。CTC 辅助网络仅在训练阶段使用。因此，在转换过程中不需要对其进行额外的输入和计算。

最后，采用 MelGAN[22]作为声码器，以较快解码速度获得高保真波形。

3 实验结果与分析

本文实验的硬件环境：选用单个 GPU，其型号为 NVIDIA GeForce RTX 2080 Ti。深度学习网络模型搭建使用 PyTorch 框架，版本为 1.8。

3.1 数据集

本文采用 VCTK[23]语料库和 CMU Arctic[24]语料库对所提的方法进行评价。在 VCTK 数据库中有 109 个不同口音的英语使用者，经过预处理后得到了 43 398 条语句作为训练集。与 VCTK 语料库不同，CMU Arctic 语料库是 1 个平行英语语料库。男性说话人“bdl”和女性说话人“slt”被选为目标扬声器。男性说话人“rms”和女性说话人“clb”被选为源说话人。本文使用 VCTK 和 CMU Arctic 进行训练和测试，此外，选择了 CMU Arctic 语料库的一部分来评估该方法，其中每个演讲者包括 40 个

语句。首先,从语料库中选取波形数据并转换为 1.6×10^4 Hz,然后,WavLM 大型预训练模型提取 1 024 维表示并将其发送到内容编码器,同时提取 80 维梅尔语谱图并发送到说话人编码器中。

3.2 实验设置

本文选择 WavLM 学习的表征作为内容表示提取的输入。WavLM 预训练模型在 LibriSpeech 960h 语料库进行训练,学习 1 024 维的瓶颈特征,用于后续学习内容表示,SSPR 经过内容编码器最终得到 1 024 维表示,说话人表征通过 GST 最终产生 256 维的风格嵌入。VQ 嵌入字典中的嵌入维度设置为 64。为了用英语音素对应的 VQ 对音素进行建模,VQ 嵌入字典中加上了 CTC 的 1 个空白标签的嵌入共计 40 个。CTC 辅助网络仅在训练阶段使用。基于编码器-解码器结构将 2 种表示拼接送入解码器中学习重构的梅尔语谱表征,用于波形转换。本文选择 MelGAN 声码器,MelGAN 是一种条件波形合成模型,具有速度快、模型参数小、非自回归结构等优点。其他型号配置与文献[25]相同。

用于比较的 3 条基线模型如下:

1)VQVC。从文献[26]中的模型学习得到。该系统以非平行语料为基础,声学特征为梅尔语谱特征,模型结构包括内容编码器、嵌入词典、解码器和 GST 音色编码器 4 部分。本文使用 Adam 优化器以 0.000 3 的学习率训练 VQVC 基线模型。小批量大小设置为 8。将 VQ 嵌入的数量设置为 40,训练步数设置为 3×10^5。

2)CTC-VQVC。重现了文献[25]模型,该模型由 1 个带有 CTC 辅助损失的内容编码器、1 个矢量量化器、1 个说话人编码器和 1 个解码器组成。CTC 损失的权重在模型训练过程中线性增加,在 5×10^4 步内最大值为 0.001。

3)FragmentVC[15]。利用预训练模型获取源说话人语音的潜在语音结构,利用 80 维梅尔语谱图获取目标说话人的音色特征。通过加入注意力机制实现 2 个不同特征空间的对齐,使用 WavLM 预训练模型代替 wav2vec2.0 模型,其余配置不变。

3.3 实验评估

3.3.1 主观评估

为了评估整体语音合成质量,本文收集了所有神经模型输出的平均意见分数(MOS)以及真实记录。语音质量包括自然度和相似度。本文进行性别内和性别间的语音转换实验,并对每个模型测试 4 种转换:rms→slt(男性到女性),rms→bdl(男性到男性),clb→slt(女性到女性),clb→bdl(女性到男性)。所有 4 组说话人在训练期间都是隐形的,测试集中每个说话人随机选择了 40 句话。每次主观评估测试都招募了 20 名志愿者。对于来自 CMU 数据集的随机样本,参与者被要求以 1~5 分的标准对语音的自然度进行评分。分数越高代表语音自然度越高。

语音自然度和说话人相似度的评价结果分别如表 1 和表 2 所示,统计了 VQVC、CTC-VQVC、FragmentVC 和 SSPR-CTC-VQVC 4 个模型在性别内、性别间和平均 3 个方面对语音自然度和说话人相似度的主观评价结果。从表 1 和表 2 可以看出,SSPR-CTC-VQVC 在性别内和性别间语音转换上都明显优于其他方法,并取得了非常好的分数,语音自然度 MOS 平均值为 3.29 和说话人相似度 MOS 平均值为 3.22。CTC-VQVC 在语音自然度和说话人相似度方面优于 VQVC 模型,这是因为 CTC-VQVC 添加的 CTC 文本监督信息有利于内容建模。此外,与其他 3 种模型相比,FragmentVC 在语音自然度方面的表现较差,可能是因为该模型使用了 SSPR 取代 80 维梅尔语谱重新训练的声码器,解码器输出特征与声码器的输入特征不匹配导致的。

表 1 各语音转换方法的语音自然度 MOS 比较

Table 1 Comparison of speech naturalness MOS for each voice conversion methods

模型	同性别	跨性别	平均值
VQVC	2.08	2.18	2.13
CTC-VQVC	2.93	3.21	3.07
FragmentVC	1.54	1.54	1.54
SSPR-CTC-VQVC	3.36	3.22	3.29

表 2 各语音转换方法说话人相似度 MOS 比较

Table 2 Comparison of speaker similarity MOS for each voice conversion methods

模型	同性别	跨性别	平均值
VQVC	1.97	1.91	1.94
CTC-VQVC	2.88	2.99	2.94
FragmentVC	2.34	2.42	2.38
SSPR-CTC-VQVC	3.22	3.23	3.22

3.3.2 客观评价

本文采用梅尔倒谱失真(MCD)对图像的自然度进行客观评价,MCD 越低表示自然度越好。在每组 VC 系统中,本文使用 4 对转换(clb-slt,clb-bdl,

rms-slt,rms-bdl),为每对转换生成 40 条语句,共计 160 条语句。分别对同性别(clb-slt,rms-bdl)和跨性别(clb-bdl,rms-slt)各 80 句以及 VC 生成的 160 个语句进行 MCD 结果验证。

各语音转换方法的 MCD 比较如表 3 所示。从表 3 可以看出,同性别的 MCD 值整体低于跨性别的值,可能是因为同性别语音转换过程中说话人音色变化幅度小,语音质量稍高。总体来看,SSPR-CTC-VQVC 的 MCD 值低于 VQVC 和 CTC-VQVC。同时使用 VQ 和 CTC 辅助网络的性能优于单独使用 CTC 或仅引入嵌入词典的性能。VQ 和 CTC 辅助网络对内容编码器学习更多的内容信息起相互促进作用。FragmentVC 表现最差,可能是因为使用注意力机制的内容和音频之间的映射效果不好。

表 3 各语音转换方法的 MCD 比较

Table 3 Comparison of MCD of each voice conversion methods

单位:dB

模型	同性别	跨性别	平均值
VQVC	9.462	9.969	9.716
CTC-VQVC	8.683	9.293	8.988
FragmentVC	10.184	10.224	10.204
SSPR-CTC-VQVC	8.827	8.964	8.896

3.3.3 可视化分析

本文对 SSPR-CTC-VQVC 模型转换前后进行可视化分析,将真实声音和通过 SSPR-CTC-VQVC 模型得到的转换后声音进行梅尔谱图可视化。图 8 显示了 CMU 中的 slt 说话人的 Arctic_a0002.wav 真实语句和由 SSPR-CTC-VQVC 转换后语音的梅尔谱图对比(彩色效果见《计算机工程》官网 HTML 版)。

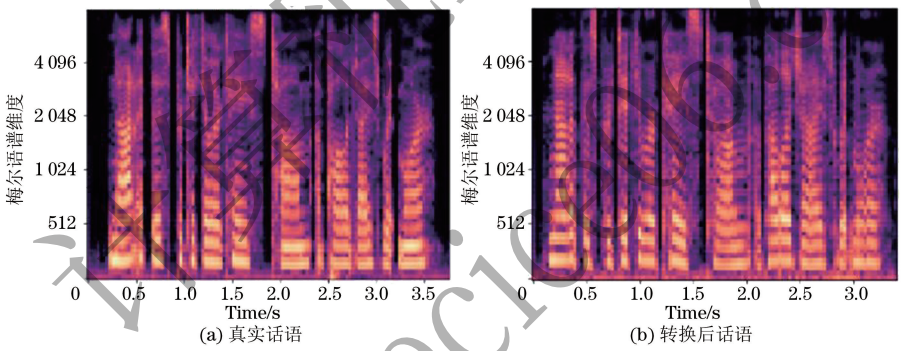

图 8 SSPR-CTC-VQVC 模型转换比较

Fig.8 Comparison of SSPR-CTC-VQVC model transformations

从图 8 可以看出,经过 SSPR-CTC-VQVC 转换后的语音与真实语音的轮廓一致,这表明转换后的语音内容得到保留。图 8 中越亮的点代表分贝越高,从同一频率下的对比可以看出,转换为目标说话人 slt 的音色、音高等身份信息也得到了保障。

3.3.4 消融实验

关于解耦向量方面,本文对 VQ 和 CTC 辅助网络进行消融研究,分别进行了 MOS 评分(包括语音自然度和说话人相似度)和 MCD 评分的客观评价。表 4 显示了在 95%置信区间下,2 种网络模块对语音自然度 MOS 值和说话人相似度 MOS 值以及 MCD 客观的影响。从表 4 的 MOS 值可以看出,这 2 种辅助网络的模型在性别内和性别间转换的自然性和相似性方面总体优于仅使用 VQ 或 CTC 网络模型。从 MCD 值来看,同时使用 2 个网络的模型 MCD 值最小,优于仅使用 VQ 或 CTC 的网络模型。上述 2 组实验证明了 SSPR-CTC-VQVC 提出的 2 个辅助内容编码器在学习更纯粹的内容表示方面的有效性。

表 4 各语音转换方法的 MOS 和 MCD 比较

Table 4 Comparison of MOS and MCD of each voice conversion methods

模型	语音自然度 MOS	说话人相似度 MOS	MCD/dB
SSPR-CTC-VQVC	3.29	3.22	8.896
w/o VQ	3.27	3.19	9.328
w/o CTC	3.47	2.96	8.913

3.3.5 说话人分类

为了验证说话人编码器能够有效地提取说话人信息,本文对编码器提取的风格嵌入进行了说话人分类实验。风格嵌入是从选定的句子中生成的,它表示从句子中提取的关于说话人的信息。说话人分类器对风格嵌入进行分类的能力被用作说话人建模能力的度量。期望的结果是来自同一说话人的不同话语可以聚类在一起,不同说话人可以相互分离。较好的聚类表明说话人分类实验取得了较好的效果,从而证明了说话人编码器能够有效地提取说话人信息。该说话人分类器由 3 个 Conv1d 层、1 个

BLSTM 层和 1 个线性层组成。每个卷积层配对批量归一化和 ReLU 激活函数，BLSTM 节点数量为 128 个。线性层的激活函数是 Softmax，损失函数是交叉熵。

图 9 显示了 CTC-VQVC 和所提出的 SSPR-CTC-VQVC 与 t-SNE 在二维空间中的向量比较（彩色效果见《计算机工程》官网 HTML 版），其中图 9(a)和图 9(b)分别表示 CTC-VQVC 和 SSPR-CTC-VQVC。相同的颜色代表相同的说话人。在说话人分类实验中，本文从 VCTK 语料库中选取了 5 位女性说话人（p225、p228、p229、p230、p231）和 5 位男性说话人（p226、p227、p232、p237、p241），对每个人 10 条语音进行分类。选择的话语通过 GST 音色编码器生成其音色向量。从图 9(a)和图 9(b)中可以看出，同一说话人的 10 个话语都能聚类在一起，并且不同说话人之间也可以区分开，这表明 2 个系统的说话人编码器都学习了有意义的说话人嵌入。这表明 GST 提取的风格向量可以作为说话人身份。此外，与图 9(a)相比，图 9(b)中同一说话人 10 句话分布得更紧密。因此，与 CTC-VQVC 相比，SSPR-CTC-VQVC 在不同说话人身份方面具有更好的聚类性能。

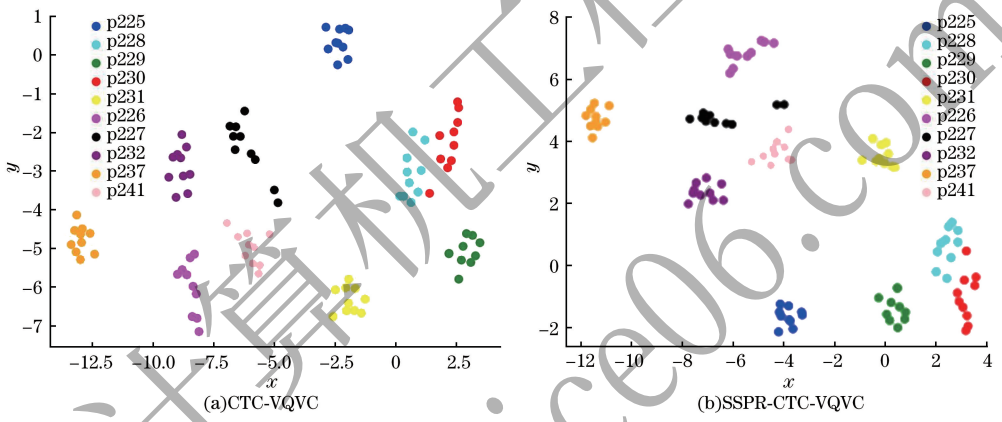

图 9 CTC-VQVC 和 SSPR-CTC-VQVC 可视化风格嵌入比较

Fig. 9 Visual style embedding comparison between CTC-VQVC and SSPR-CTC-VQVC

4 结束语

本文提出一种基于自监督预训练表示同时混合 VQ 和 CTC 的单次 VC 方法 SSPR-CTC-VQVC。使用 GST 获取语音中说话人信息，使用基于 WavLM 自监督预训练模型学习的表征送入内容编码器中学习最初内容信息的嵌入。与传统使用梅尔语谱作为声学特征学习内容嵌入相比，使用基于预训练模型学习的 SSPR 减少了嵌入向量中内容信息的损失，提高了语音质量。采用矢量量化和基于联结时序分类的监督模型，从内容表示中去除其中冗余的说话人信息。这 2 个辅助网络模块对净化的内容向量在性能上是互补的。实验结果表明，本文提出的方法可以很好地从 SSPR 中学习纯的内容信息，极大地提高了转换后的语音质量。但 SSPR-CTC-VQVC 学习的内容表示仍包含少量的说话人信息，为了更好地进行语音信息解耦，下一步将设计更好提取纯净内容的方法。比如在 VQ 层后加入互信息进行内容表征提纯，或者基于梯度反转达到类似于生成对抗网络的作用，进一步将对抗的思想引入到表征学习中，使得说话人表征和内容表征有效分离。

参考文献

[1] 张凯，朱立新，赵义正. 基于重训练高斯混合模型的语音转换方法[J]. 声学技术，2010，29(1)：52-55.
ZHANG K，ZHU L X，ZHAO Y Z. A voice conversion method based on retraining GMM[J]. Technical Acoustics，2010，29(1)：52-55. (in Chinese)

[2] CHEN L H，LING Z H，LIU L J，et al. Voice conversion using deep neural networks with layer-wise generative training[J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2014，22(12)：1859-1872.

[3] SUN L F，KANG S Y，LI K，et al. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks[C]// Proceedings of IEEE International Conference on Acoustics，Speech and Signal Processing. Washington D. C.，USA：IEEE Press，2015：4869-4873.

[4] SUN L F，LI K，WANG H，et al. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training[C]// Proceedings of the IEEE International Conference on Multimedia and Expo. Washington D. C.，USA：IEEE Press，2016：1-6.

[5] 高俊峰，陈俊国. 基于 Style-CycleGAN-VC 的非平行语料下的语音转换[J]. 计算机应用与软件，2021，38(9)：133-139，159.
GAO J F，CHEN J G. Voice conversion with non-parallel corpus based on Style-CycleGAN-VC[J]. Computer Applications and Software，2021，38(9)：133-139，159. (in Chinese)

[6] KAMEOKA H，KANEKO T，TANAKA K，et al. ACVAE-VC：non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder[EB/OL]. [2023-

02-23]. https://arxiv.org/pdf/1808.05092.pdf.
[7] 谭智元.基于自编码器的零样本语音转换系统研究[D].天津:天津大学,2020.
TAN Z Y. A study of zero-shot voice conversion system based on auto-encoder[D]. Tianjin: Tianjin University, 2020. (in Chinese)
[8] SAITO Y, NAKAMURA T, IJIMA Y, et al. Non-parallel and many-to-many voice conversion using variational autoencoders integrating speech recognition and speaker verification[J]. Acoustical Science and Technology, 2021, 42(1): 1-11.
[9] 李燕萍,曹盼,左宇涛,等. 基于 i 向量和变分自编码相对生成对抗网络的语音转换[J]. 自动化学报,2022,48(7):1824-1833.
LI Y P, CAO P, ZUO Y T, et al. Voice conversion based on i-vector with variational autoencoding relativistic standard generative adversarial network[J]. Acta Automatica Sinica, 2022, 48(7): 1824-1833. (in Chinese)
[10] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionisttemporal classification: labelling unsegmentedsequence data with recurrent neural networks [EB/OL]. [2023-02-23]. https://mediatum.ub.tum.de/doc/1292048/12285.pdf.
[11] WANG C Y, WU Y, QIAN Y, et al. UniSpeech: unified speech representation learning with labeled and unlabeled data [EB/OL]. [2023-02-23]. https://arxiv.org/abs/2101.07597.
[12] VAN DEN OORD A, LI Y Z, VINYALS O. Representation learning with contrastive predictive coding[EB/OL]. [2023-02-23]. https://arxiv.org/pdf/1807.03748.pdf.
[13] CHUNG Y A, GLASS J. Generative pre-training for speech with autoregressive predictive coding[C]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Washington D. C., USA: IEEE Press, 2020: 3497-3501.
[14] LIU A T, YANG S W, CHI P H, et al. Mockingjay: unsupervised speech representation learning with deep bidirectional transformer encoders[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2020: 6419-6423.
[15] 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]//Proceedings of the International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2021: 5939-5943.
[16] RAO K, SENIOR A, SAK H. Flat start training of CD-CTC-SMBR LSTM RNN acoustic models[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2016: 5405-5409.
[17] ZHAO Z P, BAO Z T, ZHANG Z X, et al. Attention enhanced connectionist temporal classification for discrete speech emotion recognition[C]// Proceedings of the International Symposium on Computer Architecture. Phoenix, Arizona, USA:[s.n.], 2019: 1-10.
[18] 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.
[19] 邓宏贵,郭晟伟,李志坚. 基于哈夫曼编码的矢量量化图像压缩算法[J]. 计算机工程,2010,36(4):218-219,222.
DENG H G, GUO S W, LI Z J. VQ image compression algorithm based on Huffman coding[J]. Computer Engineering, 2010, 36(4): 218-219, 222. (in Chinese)
[20] CHOU J C, YEH C C, LEE H Y. One-shot voice conversion by separating speaker and content representations with instance normalization[EB/OL]. [2023-02-23]. https://arxiv.org/pdf/1904.05742.pdf.
[21] WANG Y X, STANTON D, ZHANG Y, et al. Style Tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis[EB/OL]. [2023-02-23]. http://arxiv.org/abs/1803.09017.
[22] KUMAR K, KUMAR R, DE BOISSIERE T, et al. MelGAN: generative adversarial networks for conditional waveform synthesis[EB/OL]. [2023-02-23]. http://arxiv.org/abs/1910.06711.
[23] VEAUX C, YAMAGISHI J, MACDONALD K. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning Toolkit[EB/OL]. [2023-02-23]. https://datashare.ed.ac.uk/handle/10283/3443.
[24] KOMINEK J, BLACK A W. The CMU Arctic speech databases[C]//Proceedings of the 5th ISCA Speech Synthesis Workshop. Washington D. C., USA: IEEE Press, 2004: 1-10.
[25] KANG X, HUANG H, HU Y, et al. Connectionist temporal classification loss for vector quantized variational autoencoder in zero-shot voice conversion[J]. Digital Signal Processing, 2021, 116(6):103110.
[26] HSU C C, HWANG H T, WU Y C, et al. Voice conversion from non-parallel corpora using variational auto-encoder[EB/OL]. [2023-02-23]. https://arxiv.org/abs/1610.04019v1.

编辑 薛晋栋