基于对抗学习和增强优化的深度转换语音还原方法

苏兆品[1,3,4],周晓琳[1],张国富[1,3,4*],廉晨思[2,4],王年松[2,4],岳 峰[1,3]

(1. 合肥工业大学计算机与信息学院,安徽合肥 230601;2. 安徽省公安厅物证鉴定管理处,安徽合肥 230000;
3. 智能互联系统安徽省实验室(合肥工业大学),安徽合肥 230009;4. 音视频智能防识联合实验室,安徽合肥 230000)

摘 要: 语音转换(Voice Conversion,VC)是一种采用深度学习将源说话人声音转换为目标说话人声音的人工智能技术,不仅被广泛应用于电影配音、个性化语音定制等,也被恶意分子应用于电信诈骗、身份伪造、政治社会操纵等,给个人隐私、社会稳定乃至国家安全带来严重危害. 相比较于深度转换语音的检测,如何由深度转换语音恢复出源说话声音,即深度转换语音还原,对追踪真实说话人,防止VC非法使用,具有更重要的研究意义和实用价值. 而目前相关的研究还较少. 为此,本文提出了一种基于对抗学习和增强优化的深度转换语音还原方法. 具体来说,首先分析了深度转换语音与源语音和目标语音的相似度,提出基于初步还原-增强优化的深度转换语音还原框架. 其次,基于动态卷积和注意力机制设计对抗还原网络,通过生成器、分类器和鉴别器的对抗学习,从转换语音中学习尽可能多的源说话人信息. 然后,设计包含音色提取器、内容提取器和声码器的增强优化网络,将初步还原语音中的音色信息和深度转换语音中的内容信息进行深度融合,生成优化后的还原语音. 最后,在Free-VC、TriAAN-VC、BNE-PPG-VC三种高性能语音转换模型的数据集上验证所提方法的有效性. 对比实验结果表明,本文方法针对三种语音转换模型的还原语音,在与真实语音的平均余弦相似度上分别提高了11.9、8.7和7.1个百分点,在说话人验证系统的平均等错率EER(Equal-Error-Rate)上分别降低了4.30、3.40和3.98个百分点,说明本文方法不仅可以有效恢复出源说话人语音,而且对未知深度转换语音也有一定的适用性.

关键词: 语音转换;深度转换语音;还原语音;对抗学习;增强优化;深度神经网络

基金项目: 教育部人文社会科学研究规划基金项目(No.24YJA870011);安徽省重点研究与开发计划项目(No.202104d07020001)

中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112(2025)06-1815-14

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20240819

第二十七届中国科协年会学术论文

Adversarial Learning and Enhanced Optimization Based Restoration Method for VC-Generated Speeches

SU Zhao-pin[1,3,4], ZHOU Xiao-lin[1], ZHANG Guo-fu[1,3,4*], LIAN Chen-si[2,4], WANG Nian-song[2,4], YUE Feng[1,3]

(1. *School of Computer and Information Technology*, *Hefei University of Technology*, *Hefei*, *Anhui* 230601, *China*;
2. *Department of Physical Evidence Identification*, *Anhui Public Security Department*, *Hefei*, *Anhui* 230000, *China*;
3. *Intelligent Interconnected Systems Laboratory of Anhui Province* (*Hefei University of Technology*), *Hefei*, *Anhui* 230009, *China*;
4. *Joint Laboratory of Intelligent Prevention and Recognition of Audio and Video*, *Hefei*, *Anhui* 230000, *China*)

Abstract: Voice conversion is an artificial intelligence technology that uses deep learning to convert the voice of a source speaker into the voice of a target speaker. It is widely used not only in movie dubbing, personalized voice customization, etc., but also used by malicious individuals in telecom fraud, identity forgery, political and social manipulation, etc., posing serious threats to personal privacy, social stability, and even national security. Compared with the detection of VC-generated speeches, how to restore the source speech from VC-generated speeches, that is, VC-generated speeches restoration, has more important research significance and practical value for tracking real speakers and preventing the illegal use of VC technologies. However, there are still few related studies. In this paper, a restoration method for VC-generated speeches is proposed based on adversarial learning and enhancement optimization. Specifically, the similarity of the VC-generated

收稿日期:2024-09-03;修回日期:2025-06-05;责任编辑:孙瑶
*通讯作者:张国富

speech with the source and target speech is first analyzed, and a restoration framework is present based on preliminary restoration-further optimization. Then, an adversarial restoration network is designed based on dynamic convolution and attention mechanisms, aiming to learn as much source speaker information as possible from VC-generated speech through adversarial learning of generator, classifier, and discriminator. After that, an enhanced optimization network, consisting of timbre extractor, content extractor, and sound encoder, is designed to generate optimized restored speech by deeply fusing timbre information in the preliminary restored speech and the content information in the deep converted speech. Finally, the effectiveness of the proposed method is validated on datasets of three high-performance speech conversion models: BNE-PPG-VC, TriAAN-VC, and Free VC. Comparative experimental results show that the restored speech for the three VC models improves the mean of cosine similarity with the source speech by 11.9, 8.7, and 7.1 percentage points respectively, and reduces the mean of equal-error-rate of speaker verification system by 4.30, 3.40, and 3.98 percentage points respectively, which indicates that the proposed method can not only effectively recover the source speaker speech, but also is also applicable to unknown VC-generated speech.

Key words: voice conversion; voice conversion generated speeches; restored speech; adversarial learning; enhanced optimization; deep neural network

Foundation Item(s): MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No.24YJA870011); Anhui Province Key Research and Development Program Project (No.202104d07020001)

1 引言

随着深度学习的迅速发展,语音转换技术(Voice Conversion,VC)迎来了前所未有的关注,不仅被广泛应用于说话人匿名[1,2]、电影配音[3]、个性化语音定制[4]、歌曲风格转换[5]等多样化场景,实现声音的即时操控与编辑,极大地丰富了声音处理的可能性与用户体验. 然而,当深度转换语音应用于电信诈骗、身份伪造冒充、政治社会操纵等场景时[6],便可能成为破坏社会稳定与国家安全的利器.

为了有效抵御深度转换技术带来的风险,国内外学者致力于深度伪造语音的检测,取得了丰富的研究成果[7~10]. 值得注意的是,相比较于深度转换语音的检测,如何由深度转换语音恢复出源说话声音,即深度转换语音还原,对追踪真实说话人,防止VC非法使用,具有更重要的研究意义和实用价值. 目前仅有文献[11]在2023年提出了一种用于语音转换的说话人主动溯源框架VoxTracer,利用信息隐藏技术,将原始说话人身份隐藏在深度转换语音中,从而实现在追踪时恢复出隐藏的身份信息,进一步恢复出源说话人的原始语音. 然而,实际应用场景无法事先将源说话人身份信息嵌入到深度转换语音中,只能依赖深度转换语音中的信息还原出源说话人语音,这是一个具有强挑战性的研究问题. 主要原因在于:

(1)现有语音转换技术的目的是将源说话人声音转换为目标说话人声音,使得深度转换语音与目标说话人具有很强的音色相似性,而与源说话人的音色相似度却很低,即源说话人音色信息在深度转换语音中以弱信息的形式存在.

(2)以BNE-PPG-VC[12]、TriAAN-VC[13]、Free-VC[14]为代表的语音转换模型理论和方法有所不同,使得深度转换语音的组成也不完全相同,这进一步加剧了转换语音还原的难度.

基于上述背景,本文在深入分析深度转换语音的基础上,提出一种基于对抗学习和增强优化的深度转换语音还原方法,高效利用有限的源说话人信息,实现从深度转换语音中还原出源说话人语音.

2 研究动机

2.1 深度转换语音分析

语音转换是利用深度学习技术,保留原始语音的内容,同时改变说话人的音色、音高、音长等声学特征,从而模仿另一个说话人的声音. 以BNE-PPG-VC[12]、TriAAN-VC[13]、Free-VC[14]为代表的VC方法均是先从源语音中解耦出语音的内容,然后与目标语音中提取的音色信息相结合,重构转换后的语音[15].

Free-VC采用VITS框架进行高质量波形重构,利用WavLM特征并对其施加信息瓶颈来解开内容信息的纠缠,使用预训练的说话人编码器提取出目标说话人的音色信息,将提取的源语音的内容信息和目标说话人的音色信息相结合,重构转换后的语音.

TriAAN-VC通过三重自适应注意归一化来实现源语音到目标语音的转换,利用内容编码器提取源语音内容特征,说话人编码器提取目标说话人音色信息,其次将二者的信息放入瓶颈层,然后经过解码器对信息进行处理,最后通过门控循环单元(Gated Recurrent Unit,GRU)层和PostNet层,重构转换后的语音.

BNE-PPG-VC将瓶颈特征提取器与sequence to sequence(seq2seq)合成模块相结合,利用瓶颈特征提取器提取出与说话人无关的源语音内容信息,使用预训练的说话人编码器提取出目标说话人的音色信息,将

提取的源语音的内容信息和目标说话人的音色信息相结合,重构转换后的语音.

为了说明研究的可行性,首先基于VCTK数据集[16],分别利用BNE-PPG-VC、TriAAN-VC、Free-VC三种方法,进行男性转换为男性(M2M)、男性转换为女性(M2F)、女性转换为女性(F2F)、女性转换为男性(F2M)四种情形的语音转换,分别获得100条转换语音.然后分别计算深度转换语音与源语音和目标语音音色的余弦相似度(Cosine Similarity,CS),其平均结果如图1所示,其中余弦相似度计算公式如式(1)所示,其中$\boldsymbol{x}$和$\boldsymbol{y}$代表两条语音的说话人嵌入向量,采用说话人编码器对语音进行编码.图1中CS1为深度转换语音与目标语音音色相似度的平均值,CS2为深度转换语音与源语音音色相似度的平均值.

$$\mathrm{CS}=\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\times\|\boldsymbol{y}\|} \tag{1}$$

由图1可以看出,由于语音转换的目标性,使得深度转换语音与目标说话人的音色相似度更高,平均音色相似度值大都在80%以上;而与源说话人的音色相似度较低,在50%左右.换言之,虽然深度转换语音与源说话人的音色相似度较低,但仍存在源说话人的音色信息.因此,本文将着重考虑如何对深度转换语音中包含的源说话人音色信息进行增强,实现深度转换语音的还原.

2.2　StarGAN-VC启发

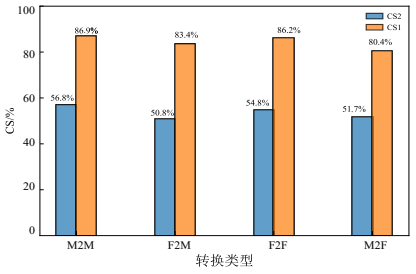
StarGAN-VC[17]利用StarGAN的图像到图像转换原理,实现了"多对多"的非平行数据集下的语音音色转换.其核心在于对语音的梅尔能量进行转换,通过生成器和鉴别器的相互作用,以及多任务分类器的辅助,使得模型能够学习到更广泛的语音特征,同时保持声音的质量和自然度.由于StarGAN-VC具有其独特的多域转换能力和灵活性,因此,可以利用StarGAN-VC模型的特性进行一个转换语音向多个源说话人进行相似性的学习,然后根据学习到的源说话人信息生成还原语音,实现转换语音还原任务.

由此启发,论文研究一种基于对抗还原和优化增强的深度转换语音还原方法(Restoration method for VC-Generated Speeches,Re-VCGS),以最大限度利用深度转换语音中包含的源说话人音色信息.

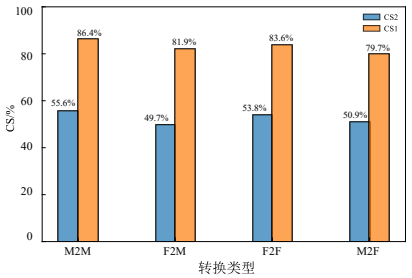
3　深度转换语音还原方法

Re-VCGS方法整体框架如图2所示,包含两个部分:

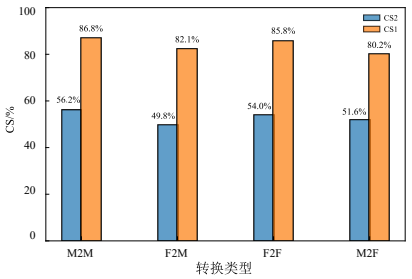
(1)对抗还原网络.通过生成器与分类器和鉴别器之间的对抗学习,得到初步还原语音,而生成器包括编码器、动态卷积和解码器.



(a) Free-VC



(b) BNE-PPG-VC



(c) TriAAN-VC

图1　三种语音转换数据集下音色相似度结果

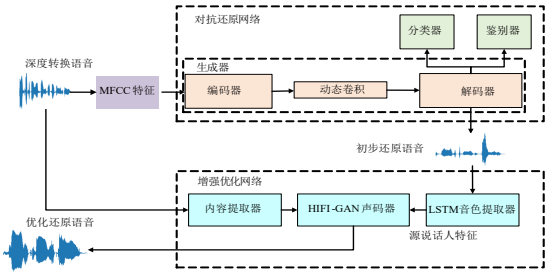


图2　Re-VCGS整体框架图

(2)增强优化网络.采用预训练模型,利用基于长短期记忆网络(Long Short-Term Memory,LSTM)的音色提取器提取初步还原语音的音色信息,利用内容提取器提取深度转换语音的内容信息,并利用声码器融合成优化后的还原语音.

3.1　对抗还原网络

对抗还原网络利用对抗生成思想,基于卷积神经网络(Convolutional Neural Networks,CNNs)[18]架构进行设计,包括生成器、鉴别器和分类器,如图3所示.生成

器的目的是学习并生成高质量的、具有源说话者特性的语音样本，同时保留转换语音的内容信息，实现转换语音到源语音的还原；鉴别器的目的是区分真实语音样本和生成器生成的虚假语音样本，并通过与生成器的对抗性训练来指导生成器不断提高其生成能力，从而生成高质量的、具有源说话者特性的语音样本；分类器的目的是准确预测输入语音特征的属性类别，实现分类器与生成器的交互，推动整个模型的优化和改进，促使生成器生成更加高质量的语音样本.

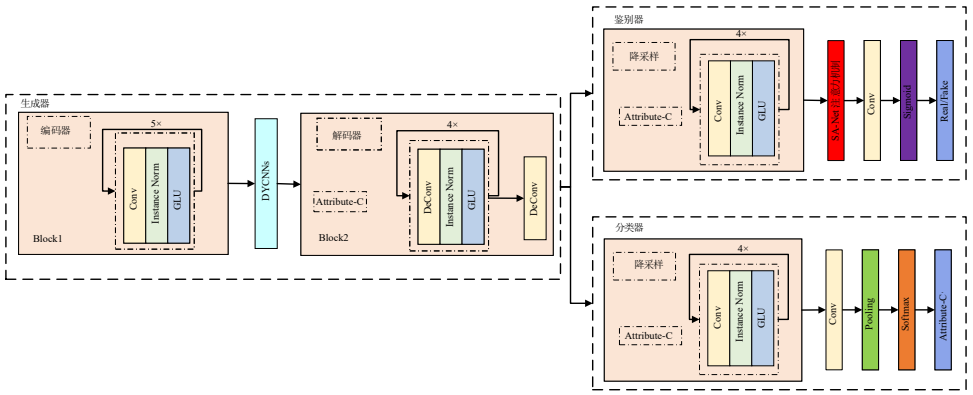


图3 对抗还原网络

3.1.1 生成器(Generator)

生成器主要包括编码器(Encoder)、动态卷积(Dynamic Convolutional Neural Networks，DYCNNs)和解码器(Decoder). 实际的编码器、动态卷积、解码器参数如表1所示，其中包括生成器、鉴别器、分类器. 括号中的数字分别表示输入通道(S=110)、输出通道、卷积核大小、步长、填充大小.

(1)编码器(Encoder). 包括五个降采样层(DownSample_1，DownSample_2，DownSample_3，DownSample_4，DownSample_5)，每个降采样层包括卷积层、归一化层和门控线性单元(Gated Linear Unit，GLU)层. 编码器接收预处理的特征，卷积层从输入特征中提取出高级别的语音表示. 通过卷积操作，可以捕捉到语音信号中的局部相关性. 为加速模型训练过程和提高模型的生成效果，采用了实例归一化层，独立地对每个样本的均值和方差进行归一化处理. GLU层有助于模型在复杂的语音信号中捕捉更关键的特征，并且能够在处理时间序列数据时更好地捕捉到这些长期依赖关系，从而提升还原的准确性和自然度.

(2)动态卷积(DYCNNs). 为增强经过编码器输出特征的表示能力，提升生成器生成还原语音的质量，Re-VCGS采用动态卷积模块[19]，其中注意力机制包含平均池化层和两个全连接层，并在中间连接一个ReLU激活函数，最后使用一个Softmax函数，动态地调整每个卷积核的权重. 需要注意的是，注意力权值并不固定，随输入的变化而变化，使得动态卷积具有更强的特征表示能力. 在动态卷积之后，会接入批量归一化(BatchNorm)层和ReLU激活函数. 动态卷积层中卷积核数量K设置为4，在Softmax函数中，温度参数τ(τ=30)用于控制注意力分布的稀疏性. 使得注意力分布更加均匀，有利于训练初期多个卷积核的同时学习.

(3)解码器(Decoder). 包括四个上采样层(UpSample_1，UpSample_2，UpSample_3，UpSample_4)和一个反卷积(Deconvolution，DeConv)层. 每个上采样层包括反卷积层、归一化层和GLU层. 解码器根据编码器和动态卷积提供的潜在特征表示和说话人条件信息，将潜在特征表示逆变换映射回音频域，以确保能够准确地还原语音特征并保留关键信息.

3.1.2 鉴别器(Discriminator)

为了提高生成器生成还原语音的能力，本文基于SA-NET注意力机制[20]设计鉴别器模块，区分真实语音样本和生成器生成的虚假样本，具体网络结构如表1所示.

鉴别器结构主要包括四个采样层(DownSample_1，DownSample_2，DownSample_3，DownSample_4)、SA-NET注意力机制、卷积层(Conv)、Sigmoid激活函数，将条件信息注入到网络的不同层级，在每个降采样之前，将信息沿空间维度进行扩展，然后与当前下采样层的输入张量沿着特征通道维度进行拼接，使得每一层的特征表示都包含了这部分全局信息；拼接后的张量被送入降采样层和SA-NET模块，用于减少数据的空间维度，同时保留重要的特征信息；最后通过Sigmoid激活函数，将最后一个降采样层的输出归一化到(0，1)区间内，实现鉴别器在细粒度上(局部片段级别)对输入进行真实性和虚假性的判断，而不仅仅是整体判断.

鉴别器融合了空间注意力和通道注意力，增强了

表1　对抗还原网络结构

Model	Layer	Structure
Generator	DownSample_1	[1, 32, (3, 9), (1, 1), (1, 4)]
	DownSample_2	[32, 64, (4, 8), (2, 2), (1, 3)]
	DownSample_3	[64, 128, (4, 8), (2, 2), (1, 3)]
	DownSample_4	[128, 64, (3, 5),(1, 1), (1, 2)]
	DownSample_5	[64, 5, (9, 5), (9, 1), (4, 2)]
	DYCNNs	(5, 5, 3, 1, 1)
	UpSample_1	[5+*S*, 64, (9, 5), (9, 1), (0, 2)]
	UpSample_2	[64+*S*, 128, (3, 5), (1, 1), (1, 2)]
	UpSample_3	[128+*S*, 64, (4, 8), (2, 2), (1, 3)]
	UpSample_4	[64+*S*, 32, (4, 8), (2, 2), (1, 3)]
	Deconv	[32+*S*, 1, (3, 9), (1, 1), (1, 4)]
Discriminator	DownSample_1	[1+*S*, 32, (3, 9), (1, 1), (1, 4)]
	DownSample_2	[32+*S*, 32, (3, 8), (1, 2), (1, 3)]
	DownSample_3	[32+*S*, 32, (3, 8), (1, 2), (1, 3)]
	DownSample_4	[32+*S*, 32, (3, 6),(1, 2), (1, 2)]
	SA-NET	(32,32)
	Conv	[32+*S*, 1, (36, 5), (36, 1), (0, 2)]
Classifier	DownSample_1	[1, 8, (4, 4), (2, 2), (1, 1)]
	DownSample_2	[8, 16, (4, 4), (2, 2), (1, 1)]
	DownSample_3	[16, 32, (2, 4), (2, 2), (0, 1)]
	DownSample_4	[32, 16, (1, 4),(1, 2), (0, 1)]
	Conv	[16, S, (1, 4), (1, 2), (0, 1)]

特征表示的能力,提高了鉴别器的鉴别能力,可以更好地促进生成器对源语音的还原效果.

3.1.3　分类器(Classifier)

为了实现分类器与生成器的对抗学习,促使生成器生成更加高质量的语音样本,本文设计了分类器,具体网络结构如表1所示.

分类器主要包括降四个采样层(DownSample_1,DownSample_2,DownSample_3,DownSample_4)、卷积层(Conv)、平均池化层、Softmax激活函数. 首先通过连续的四个降采样层,逐步降低输入数据的空间维度,有助于提取输入数据的层次化特征;其次经过一个卷积层调整特征的输出维度,再经过平均池化操作,满足Softmax层对输入数据的要求;最后,将扁平化处理后的特征向量通过Softmax层输出每个类别的预测概率.

分类器通过一系列的降采样、卷积和平均池化操作,能够从输入数据中提取出有用的特征,促使生成器生成更加高质量的语音样本.

3.2　增强优化网络

为了进一步利用深度转换语音中有限的源说话人特征,获得更高质量的还原语音,本文基于预训练模型设计了增强优化网络,对初步还原语音进一步优化,如图2所示,主要包括音色提取器、内容提取器和声码器.

(1)内容提取器. 使用Free-VC中预训练的前置编码器作为内容提取器,主要包含WavLM模型、瓶颈提取器和归一化流. WavLM模型接收原始波形作为输入,并产生1 024维的SSL(Self-Supervised Learning)特征,同时包含内容信息和说话者信息;瓶颈提取器被用于去除说话者信息等与内容无关的特征,从而提取出更加纯净的语音内容表示;归一化流学习数据的潜在分布,并通过一系列的可逆变换函数将原始数据映射到一个已知的简单分布上.

(2)音色提取器. 使用预训练的LSTM说话人编码器[21]作为音色提取器,包含3层,每层有256个隐藏节点,后面是一个256个单元的投影层,最后一层的L2归一化隐藏状态被视为说话人嵌入向量. Re-VCGS使用LSTM说话人编码器提取出初步还原语音的音色信息,相当于是将生成器中学习到的音色信息放大,可以提高还原语音的语音质量.

(3)HIFI-GAN声码器. HIFI-GAN声码器[22]采用对抗学习思想将内容提取器提取的内容信息和音色提取器提取到的说话人信息进行融合,生成高质量的还原语音.

3.3　训练和损失函数

由于增强优化网络采用预训练模型,Re-VCGS只需训练对抗还原网络. 目标是通过对抗训练优化生成器G,使其可以根据给定的语音序列特征(深度转换语音)生成新的高质量语音特征序列(还原语音),同时鉴别器D和分类器C能够准确地区分真实和还原的语音特征,并预测它们的属性类别. 其中,生成器G的输入为转换语音的声学特征以及还原说话人的标签,输出为还原语音的声学特征. 鉴别器D的输入为生成器G的还原语音的声学特征以及还原说话人的标签,输出为此条语音的真假概率. 分类器C的输入为生成器G的还原语音的声学特征,输出为类别概率.

因此,对抗还原网络需要考虑三种损失:

(1)对抗性损失. 对抗性损失是训练过程中用于驱动生成器和鉴别器相互竞争的关键部分. 鉴别器D和生成器G的对抗性损失定义分别如式(2)和式(3)所示:

$$L_{D-\mathrm{adv}}(D)=-E_{c\sim p(c),\boldsymbol{y}\sim p(\boldsymbol{y}|c)}\left[\log D(\boldsymbol{y},c)\right]-E_{\boldsymbol{x}\sim p(\boldsymbol{x}),c\sim p(c)}\left[\log\left(1-D\left(G(\boldsymbol{x},c),c\right)\right)\right] \tag{2}$$

$$L_{G-\mathrm{adv}}(G)=-E_{\boldsymbol{x}\sim p(\boldsymbol{x}),c\sim p(c)}\left[\log D\left(G(\boldsymbol{x},c),c\right)\right] \tag{3}$$

其中,$D(\boldsymbol{y},c)$表示鉴别器对真实语音特征$\boldsymbol{y}$属于类别c的概率输出;$G(\boldsymbol{x},c)$表示生成器将输入特征$\boldsymbol{x}$还原为目标类别c的生成特征;$\boldsymbol{x}\sim p(\boldsymbol{x})$表示具有任意属性的语音特征序列;$c\sim p(c)$表示从属性标签分布$p(c)$中随机采样还原说话人属性标签$c$;$\boldsymbol{y}\sim p(\boldsymbol{y}|c)$表示一个带有属性$c$的真实语音声学特征序列的训练样本. $L_{D-\mathrm{adv}}(D)$和

$L_{G-\mathrm{adv}}(G)$分别为鉴别器和生成器的对抗性损失.

(2)分类损失. 分类损失旨在确保分类器能够准确地根据样本(真实样本、生成器生成的样本)预测其对应的条件,同时生成器能够生成与给定条件一致的样本. 分类器C和生成器G的分类损失定义分别如式(4)和式(5)所示:

$$L_{C-\mathrm{cls}}(C)=-E_{c\sim p(c),\boldsymbol{y}\sim p(\boldsymbol{y}|c)}\left[\log p_C(c|\boldsymbol{y})\right] \tag{4}$$

$$L_{G-\mathrm{cls}}(G)=-E_{\boldsymbol{x}\sim p(\boldsymbol{x}),c\sim p(c)}\left[\log p_C\left(c|G(\boldsymbol{x},c)\right)\right] \tag{5}$$

其中,$p_C(c|\boldsymbol{y})$表示分类器C对输入特征$\boldsymbol{y}$属于类别c的概率预测;$L_{C-\mathrm{cls}}(C)$和$L_{G-\mathrm{cls}}(G)$分别为分类器和生成器的分类损失.

(3)循环一致性损失和恒等映射损失. 为了确保生成器G生成的样本既符合还原的条件c,又能保留原始输入$\boldsymbol{x}$的重要特征,生成器G使用循环一致性损失以及恒等映射损失,分别如式(6)和式(7)所示:

$$L_{G-\mathrm{cyc}}(G)=E_{c'\sim p(c),\boldsymbol{x}\sim p(\boldsymbol{x}|c'),c\sim p(c)}\left[\left\|G(G(x,c),c')-\boldsymbol{x}\right\|_\rho\right] \tag{6}$$

$$L_{G-\mathrm{id}}(G)=E_{c'\sim p(c),\boldsymbol{x}\sim p(\boldsymbol{x}|c')}\left[\left\|G(\boldsymbol{x},c')-\boldsymbol{x}\right\|_\rho\right] \tag{7}$$

其中,$G\left(G(\boldsymbol{x},c),c'\right)$表示将生成特征再次转回原类别$c'$的结果;$c'\sim p(c)$表示从属性标签分布$p(c)$中采样一个原始转换语音属性标签$c'$;$\boldsymbol{x}\sim p\left(\boldsymbol{x}|c'\right)$表示具有属性$c'$的真实语音声学特征序列的训练样本;$L_{G-\mathrm{cyc}}(G)$和$L_{G-\mathrm{id}}(G)$分别为生成器的循环一致性损失和恒等映射损失.

综上,生成器G、鉴别器D、分类器C的损失分别如式(8)~式(10)所示:

$$L_D(D)=L_{D-\mathrm{adv}}(D) \tag{8}$$

$$L_C(C)=L_{C-\mathrm{cls}}(C) \tag{9}$$

$$\begin{aligned}L_G(G)=&L_{G-\mathrm{adv}}(G)+\lambda_{\mathrm{cls}}L_{G-\mathrm{cls}}(G)\\&+\lambda_{\mathrm{cyc}}L_{G-\mathrm{cyc}}(G)+\lambda_{\mathrm{id}}L_{G-\mathrm{id}}(G)\end{aligned} \tag{10}$$

其中,λ_{cls}、λ_{cyc}和λ_{id}表示不同损失的权重,用于平衡不同损失的重要性.

基于上述损失函数,对抗还原网络的训练过程可以描述如下:

(1)首先提取转换语音以及源语音的MFCC特征和F0基频特征,为后续的生成对抗网络训练提供帮助.

(2)其次,固定生成器,训练鉴别器、分类器. 将真实的源语音特征输入鉴别器和分类器中,期望其输出为真实,将转换语音特征输入生成器,生成还原语音的特征,然后将这些特征输入鉴别器和分类器中,期望输出为假.

(3)然后,固定鉴别器、分类器,训练生成器. 在此阶段,生成器的目标是欺骗鉴别器和分类器,使其将生成的语音特征误认为是真实的. 生成器接收转换语音特征和源语音标签作为输入,生成还原后的语音特征. 将这些生成的语音特征输入鉴别器和分类器,并计算对抗性损失.

(4)最后,重复上述两个步骤,交替训练生成器、鉴别器和分类器,直到达到收敛状态.

3.4 推理

对于待还原的深度转换语音,首先将其MFCC特征放入生成器G中,生成初步还原语音;然后将初步还原语音放入预训练的音色编码器中,提取初步还原语音的音色信息,再利用预训练的内容编码器提取深度转换语音的内容信息;最后将内容信息与音色信息放进HIFI-GAN声码器中,融合生成所需要的还原语音.

4 实验结果与分析

为了验证Re-VCGS的有效性,本文与StarGAN-VC模型进行对比分析. 在使用StarGAN-VC进行转换语音还原任务时,在训练过程中,采用StarGAN-VC模型的训练方法. 在推理时首先通过待处理的转换语音对学习到的所有源说话人进行还原,然后通过相似度计算,选取出相似度最高的那条还原语音,即初步还原语音,最后通过增强优化网络得到转换语音的还原语音.

在本节中,我们首先对实验数据集和参数设置进行说明,然后通过消融实验、对比实验以及泛化性实验进行性能验证.

4.1 数据集和实验设置

面向深度转换语音的还原研究,目前为止没有可利用的公开数据集. 因此本文首先基于VCTK数据集构造转换语音数据集CS-VCTK,对110个说话人进行语音转换. 对于每个说话人,分别采用Free-VC、TriAAN-VC、BNE-PPG-VC三种算法生成400条转换语音,其中300条作为训练,100条作为测试,共计约132 000条,如表2所示.

表2 CS-VCTK数据集

	Free-VC	TriAAN-VC	BNE-PPG-VC	合计
训练集	300	300	300	99 000
测试集	100	100	100	33 000

对于特征参数,从每个语音片段中提取频谱包络、对数基频(log F0)和非周期性成分(Aps),帧移为5 ms. 随后,从每个频谱包络中提取了36个Mel倒谱系数(Mel Frequency Cepstral Coefficents,MFCC).

使用Adam优化器来训练网络,对于生成器、鉴别器、分类器,其学习率都设置为0.000 1,在训练过程中我们根据训练步数对学习率进行调整,目的是在训练初期使用较大的学习率以快速收敛,然后在训练后期逐渐减小学习率以避免过拟合,提高模型的稳定性和

泛化能力. 学习率在训练过程中按固定步数间隔进行调整,具体条件如下:当前训练步数超过预设的衰减起始步数(1×10^5步),并且当前训练步数是学习率调整步数(1×10^4步)的整数倍时,才对学习率进行一定的调整. λ_{cyc}的值设置为3,λ_{cls}和λ_{id}的值设置为2. 深度转换语音的还原任务分为男性转换为男性(M2M)、男性转换为女性(M2F)、女性转换为女性(F2F)、女性转换为男性(F2M)四种情形.

为了评估还原语音质量,采用余弦相似度(CS)和等错误率(Equal Error Rate,EER)进行客观评估,采用平均主观意见分(Mean Opinion Score,MOS)进行主观评估. 其中,余弦相似度计算过程如式(1)所示;EER是错误接收率(False Acceptance Rate,FAR)与错误拒绝率(False Rejection Rate,FRR)相等的点,定义在接收者操作特征(Receiver Operating Characteristic curve,ROC)曲线上,使用基于ECAPA-TDNN[23]模型训练的说话人验证系统进行EER计算;MOS在1~5之间,1~5分别代表"差""较差""一般""好""优秀".

4.2　消融实验

为了验证各模块的性能,在数据集CS-VCTK上对各模块分别进行实验,包括动态卷积模块DYCNNs、Shuffle Attention注意力机制(SA-NET)和增强优化网络(enhancement Optimization Networks,OptN). 由表3~表5所示,首先仅采用Encoder-Decoder模块,接着在此基础上加上DYCNNs模块,然后在鉴别器中引入注意力机制SA-NET,最后通过优化模块的引入得到提出的完整的还原网络结构.

由实验结果可以得出以下结论:

(1)在Free-VC、TriAAN-VC、BEN-PPG-VC三种数据集上的结果表明,动态卷积模块、Shuffle Attention注意力模块、增强优化模块对整体方法性能均有提升作用. 引入动态卷积网络模块,相较于Encoder+Decoder模块,在三种数据集上转换语音还原平均相似度(MCS)分别提高了4.9、3.6和2.9个百分点,在EER上分别降低了1.30、1.26和1.20个百分点. 在引入动态卷积模块后,在鉴别器中加入Shuffle Attention注意力机制模块,相较于Encoder+Decoder模块,在三种数据集上转换语音还原平均相似度(MCS)分别提高了6.7、4.9和4.0个百分点,在EER上分别降低了1.78、1.62和1.96个百分点. 在此基础上引入增强优化网络,可以看出,相较于Encoder+Decoder模块,在三种数据集上转换语音还原平均相似度(MCS)分别提高了11.9、8.7和7.1个百分点,在EER上分别降低了4.30、3.40和3.98个百分点.

(2)动态卷积模块的引入使得还原语音的相似度以及EER性能都有所提高,但在不同的语音转换模型上会有不同差异. 其中,由于Free-VC数据集中的残留源说话人信息较多,因此在加入动态卷积模块后增强了模型对源说话人特征的学习能力,因此,在Free-VC数据集上的相似性提高了4.9个百分点,EER降低了

表3　Free-VC数据集下消融实验结果　　单位:%

模块	类型	CS	EER
Encoder+Decoder	M2M	67.8	24.62
	F2M	60.0	
	F2F	67.2	
	M2F	63.8	
	平均值	64.7	
Encoder+Decoder+DYCNNs	M2M	71.3	23.32
	F2M	67.4	
	F2F	70.2	
	M2F	69.7	
	平均值	69.6	
Encoder+DYCNNs+Decoder+SA	M2M	72.5	22.84
	F2M	70.2	
	F2F	71.8	
	M2F	71.4	
	平均值	71.4	
Encoder+DYCNNs+Decoder+SA+OptN	M2M	77.1	20.32
	F2M	73.5	
	F2F	79.4	
	M2F	76.7	
	平均值	76.6	

表4　TriAAN-VC数据集下消融实验结果　　单位:%

模块	类型	CS	EER
Encoder+Decoder	M2M	69.1	24.74
	F2M	61.8	
	F2F	67.6	
	M2F	62.4	
	平均值	65.2	
Encoder+Decoder+DYCNNs	M2M	70.4	23.48
	F2M	67.2	
	F2F	69.6	
	M2F	68.3	
	平均值	68.8	
Encoder+DYCNNs+Decoder+SA	M2M	71.2	23.12
	F2M	69.3	
	F2F	70.8	
	M2F	69.2	
	平均值	70.1	
Encoder+DYCNNs+Decoder+SA+OptN	M2M	77.3	21.34
	F2M	70.6	
	F2F	74.6	
	M2F	73.2	
	平均值	73.9	

表5 BNE-PPG-VC数据集下消融实验结果 单位:%

模块	类型	CS	EER
Encoder+Decoder	M2M	70.6	25.42
	F2M	65.1	
	F2F	66.5	
	M2F	61.1	
	平均值	65.8	
Encoder+Decoder+DYCNNs	M2M	71.2	24.22
	F2M	68.2	
	F2F	68.2	
	M2F	67.5	
	平均值	68.7	
Encoder+DYCNNs+Decoder+SA	M2M	71.8	23.46
	F2M	69.1	
	F2F	69.4	
	M2F	69.2	
	平均值	69.8	
Encoder+DYCNNs+Decoder+SA+OptN	M2M	73.6	21.44
	F2M	72.9	
	F2F	74.1	
	M2F	71.2	
	平均值	72.9	

1.3个百分点,由于TriAAN-VC和BNE-PPG-VC两种语音转换模型对源说话人信息的抹除相对较多,动态卷积模块对源说话人信息的学习能力有限,其相似性分别提高了3.6个百分点和2.9个百分点,EER分别降低了1.26个百分点和1.20个百分点.

(3)在鉴别器中加入Shuffle Attention注意力机制模块进一步提升了Re-VCGS性能.鉴别器的目的是通过对抗学习,判定生成器的还原语音和源说话人语音之间的差异,来提高生成器的生成还原语音质量.但语音转换模型的机理区别,导致转换语音中残留的源说话人信息也有所差异.对于Free-VC,其含有的源说话人信息较多,可以更充分地学习到源说话人特征,相似性提高了6.7个百分点,EER降低了1.78个百分点,而对于TriAAN-VC和BNE-PPG-VC相似性分别提高了4.9个百分点和4.0个百分点,EER分别降低了1.62个百分点和1.96个百分点.

(4)增强优化网络的设计进一步优化了Re-VCGS方法的表现能力.这是由于优化模块中的音色提取器可以对初步还原语音的音色进行进一步提取,对初步还原语音的音色信息进行增强,从而提高还原语音的语音质量,达到对初步还原语音优化的目的.对于Free-VC,其转换语音与源语音的平均相似度为53.5%,经增强优化网络后提升到76.6%,EER降低了4.30个百分点;引入增强优化模块后,TriAAN-VC和BNE-PPG-VC转换语音与源语音平均相似度从52.9%、52.5%提升到73.9%、72.9%,EER分别降低了3.40个百分点和3.98个百分点.

4.3 主观质量评价

在语音转换还原任务方面,MOS能够有效衡量还原后的语音在保留源语音内容的程度,可以对还原语音的自然度和源说话人相似度进行主观评价.

在自然度评估中,评估者会听到来自源说话人的原始录音样本,这些样本会与通过两种不同还原方法生成的样本随机混合在一起,然后呈现给评估者.这样做的目的是减少评估者对原始样本和还原样本之间的潜在偏见,从而客观地评价还原语音的自然度(Naturalness).

在相似度(Similarity)的评估中,还原的语音样本会直接与源说话人的原始录音样本进行比较,这种比较更直接地反映了还原系统在还原方面的能力.同样的,对于每种还原对(如F2F、M2M、M2F、F2M),都会从数据集中选取10个句子进行评估.

实验共邀请了20名评估者(10男,10女,年龄20~27岁),均具备语音信号处理相关领域的学术背景(硕士及以上学历).所有评估者均通过听力筛查测试,并接受过语音质量主观评估培训.评估者未参与本研究的任何设计或数据集采集环节,以规避利益冲突.

在培训阶段,为了使评估者明确MOS评分标准,评估者首先听取5组示例音频(包含高、中、低质量样本),该样本与实验数据集并不相关.在正式评估阶段,每组样本以随机顺序播放,评估者需在独立房间中对音质自然度和说话人相似度分别进行评分,每项评分间隔至少10 s以避免疲劳效应.

他们在安静的房间内使用耳机进行测评,以确保评估环境的一致性和减少外界干扰.评估者可以重复播放每个样本多次,并在提交结果前随时更改他们对任何样本的评分.这种灵活性有助于评估者更准确地表达他们对还原语音质量的看法,实验结果如表6所示.

表6 主观评价结果

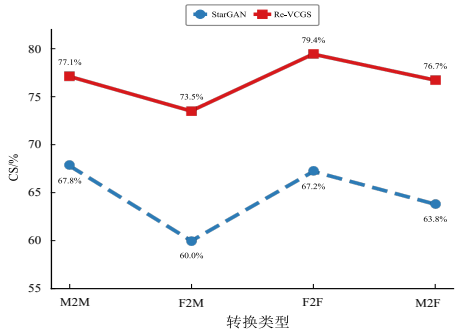
	StarGAN方法		Re-VCGS方法	
	Naturalness	Similarity	Naturalness	Similarity
F2M	2.78±0.04	2.88±0.06	3.22±0.05	3.42±0.05
F2F	3.24±0.08	3.26±0.07	3.42±0.06	3.68±0.09
M2M	3.16±0.09	3.20±0.09	3.66±0.07	3.84±0.09
M2F	2.56±0.07	2.62±0.08	3.34±0.06	3.62±0.08
Average	2.94±0.07	2.99±0.07	3.41±0.06	3.64±0.08

从表6中数据可以看出,本文方法在保持语音的自然性和源说话人的特征方面具有较好的性能.在F2F、M2M、F2M、M2F四种还原类型中,本文方法在多数场景下表现优异,表明其在同性别和异性别之间的还原都有较好的适应性.但整体上同性别之间的还原语音自然度和相似度都优于异性别,可能因为方法中使用的

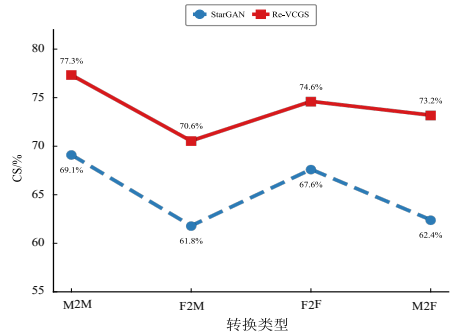
简单对数尺度下的线性变换不足以充分模拟异性别间的音高转换. 表6中,MOS评分标准差均不超过0.1,这是因为评估者均为语音信号处理领域相关人员,对评分标准理解一致. 评估前进行标准化培训,明确评分维度和示例样本,减少了主观理解差异.

4.4　客观质量评价

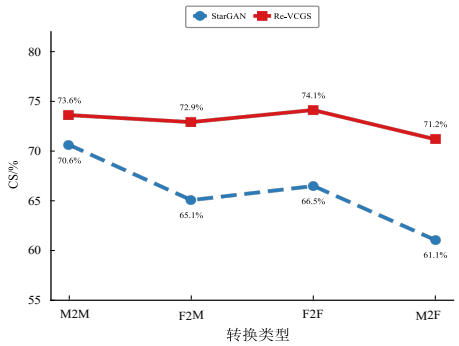
图4和图5给出了在Free-VC、TriAAN-VC、BNE-PPG-VC模型下的测试结果.



(a) Free-VC



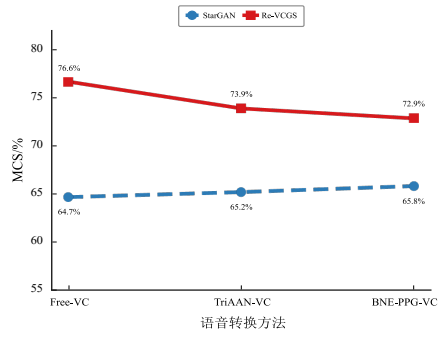
(b) TriAAN-VC



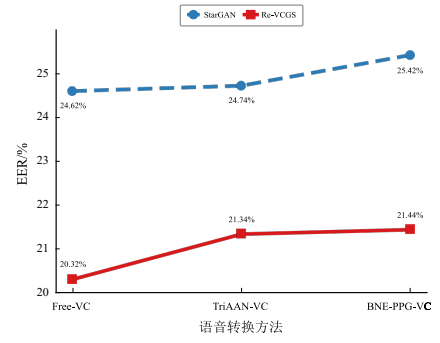
(c) BNE-PPG-VC

图4　三种语音转换模型下还原语音的CS指标

由图4可知,在Free-VC模型下构建的数据集上,本文方法在M2M、F2M、F2F、M2F的转换语音还原相似度(CS)上分别提高了9.3、13.5、12.2和12.9个百分点. 在TriAAN-VC模型下构建的数据集上,本文方法在M2M、F2M、F2F、M2F的转换语音还原相似度(CS)上分别提高了8.2、8.8、7.0和10.8个百分点. 在BNE-PPG-VC模型下构建的数据集上,本文方法在M2M、F2M、F2F、M2F的转换语音还原相似度(CS)上分别提高了3.0、7.8、7.6和10.1个百分点.



(a) MCS



(b) EER

图5　三种语音转换模型下还原语音的MCS指标和EER指标

由图5可知,在Free-VC模型下构建的数据集上,本文方法在MCS上提高了11.9个百分点,在EER上降低了4.3个百分点. 在TriAAN-VC模型下构建的数据集上,本文方法在MCS上提高了8.7个百分点,在EER上降低了3.4个百分点. 在BNE-PPG-VC模型下构建的数据集上,本文方法在MCS上提高了7.1个百分点,在EER上降低了3.98个百分点.

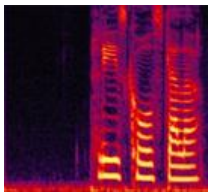
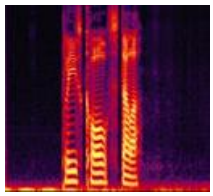
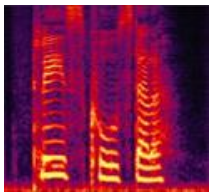
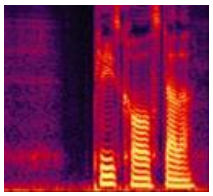
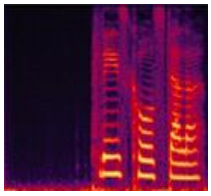
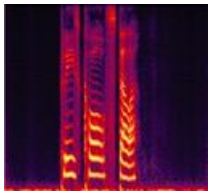
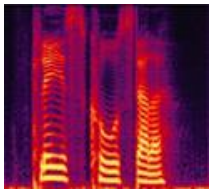
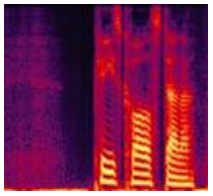
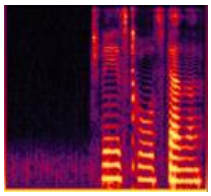
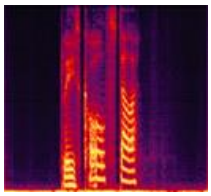
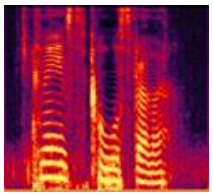
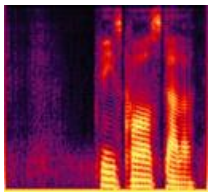
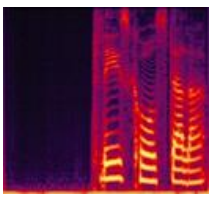
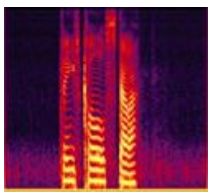
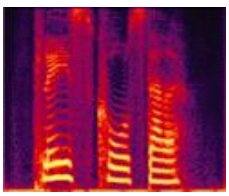
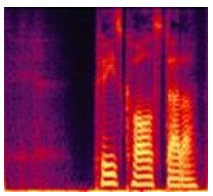
综上所述,本文提出的基于动态卷积和Shuffle Attention注意力的语音转换追踪算法在处理转换语音的说话人还原任务时具有一定的优势,由于动态卷积可以增强残留说话人信息特征的学习,所引入的SA-NET注意力机制可以提高生成器生成语音的质量,因此该网络能够追踪到更具有表现力的残留说话人语音.

为了进一步比较语音质量,随机选择一条源语音,分别利用三种转换模型得到深度转换语音,利用StarGAN-VC和Re-VCGS进行还原. 将源语音、转换语音、StarGAN-VC还原语音、Re-VCGS还原语音的梅尔谱图,以及源语音分别与转换语音、StarGAN-VC还原语音和Re-VCGS方法还原语音的相似度(CS)进行对比分

析,结果如表7~表9所示.

由表7~表9的梅尔谱图可以看出,振幅模式的不规则性在转换语音中尤为明显,表现为不自然的峰值和谷值,这与自然语音平滑连续的振幅变化形成鲜明对比. 而且,在转换中,共振峰显得不清晰或出现双重峰值,这可能是源语音和目标语音的特征混合在一起的结果. 转换语音噪声较高,在转换过程中未能够有效融合或消除源语音的某些特定频率成分,而这些在源语音中通常不会出现. Re-VCGS还原语音的梅尔谱图与源语音较为相似,但在不同数据集下的还原语音效果各有偏差. 造成还原语音效果不同的原因可能是因为各个语音转换模型的效果各有偏差,由于这些偏差,导致转换语音中残留的源说话人信息也会各有不同,因此还原语音结果也会有所偏差.

表7 Free-VC数据集下梅尔谱图对比

类型	F2F	F2M	M2F	M2M
源语音				
转换语音				
CS/%	54.8	50.8	51.7	56.8
StarGAN-VC还原				
CS/%	67.2	60.0	63.8	67.8
Re-VCGS还原				
CS/%	79.4	73.5	76.7	77.1

具体来说,在Free-VC模型下构建的数据集上,本文方法在F2F、F2M、M2F、M2M的转换语音还原相似度(CS)上分别为79.4%、73.5%、76.7%、77.1%. 在TriAAN-VC模型下构建的数据集上,本文方法在F2F、F2M、M2F、M2M的转换语音还原相似度(CS)上分别为74.6%、70.6%、73.2%、77.3%. 在BNE-PPG-VC模型下构建的数据集上,本文方法在F2F、F2M、M2F、M2M的转换语音还原相似度(CS)上分别为74.1%、72.9%、71.2%、73.6%.

综上,Re-VCGS还原语音相似度(CS)均可以达到70%以上,在Free-VC模型的实验结果整体上优于TriAAN-VC和BNE-PPG-VC,转换语音还原相似度的结果与梅尔谱图可视化结果一致. 此外,Re-VCGS方法整体上同性别之间的还原语音自然度和相似度都优于异性别,可能是因为方法使用的线性变换不足以充分模拟异性别间的音高转换.

4.5 泛化性

训练集是基于Free-VC、TriAAN-VC和BNE-PPG-VC三种方法获得,其中Free-VC是基于VITS框架,采用自监督学习和信息瓶颈提取内容特征,依赖隐空间解耦获取音色特征;TriAAN-VC是基于编码器-解码器结构,结合三重自适应注意力归一块,重构转换后的语音;BNE-PPG-VC结合瓶颈特征提取与序列到序列模型,使用混合逻辑分布(Mixture of Logistic,MoL)注意力

表 8　TriAAN-VC 数据集下梅尔谱图对比

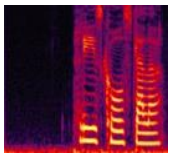
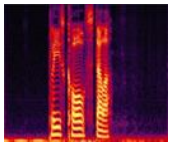
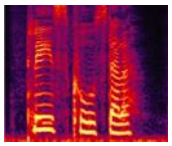
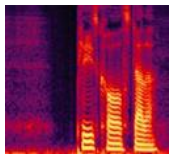
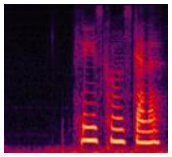
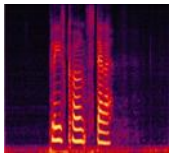
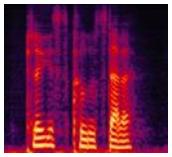
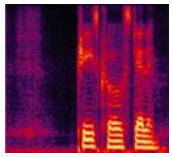
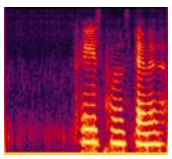
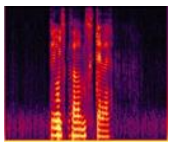
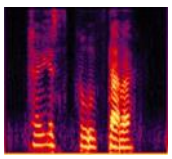
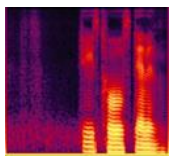
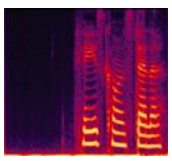
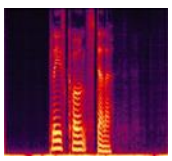
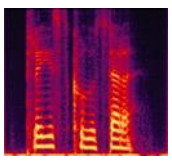
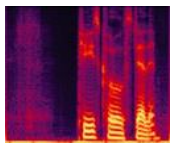
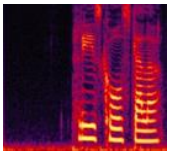
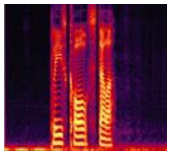
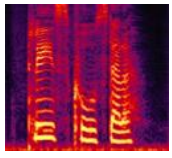
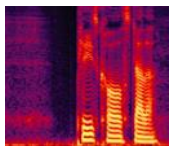
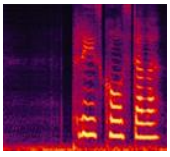
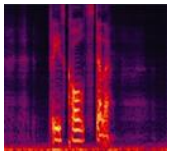
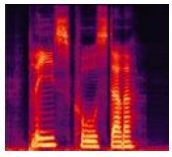
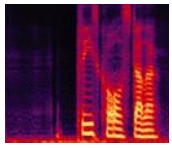
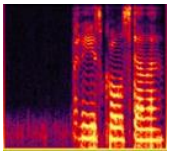
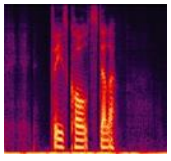
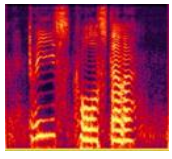
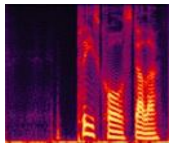
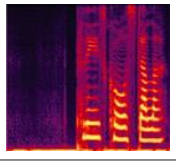
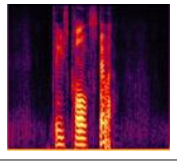
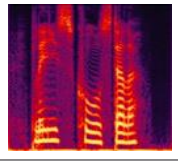
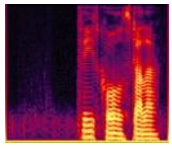
类型	F2F	F2M	M2F	M2M
源语音				
转换语音				
CS/%	54.0	49.8	51.6	56.2
StarGAN-VC 还原				
CS/%	67.6	61.8	62.4	69.1
Re-VCGS 还原				
CS/%	74.6	70.6	73.2	77.3

表 9　BNE-PPG-VC 数据集下梅尔谱图对比

类型	F2F	F2M	M2F	M2M
源语音				
转换语音				
CS/%	53.8	49.7	50.9	55.6
StarGAN-VC 还原				
CS/%	66.5	65.1	61.1	70.6
Re-VCGS 还原				
CS/%	74.1	72.9	71.2	73.6

处理长序列对音色进行建模. 而基于扩散模型的语音转换方法采用渐进式生成过程和概率建模,使转换的语音与目标说话人具有更高的相似性和保真性,而与源说话人具有更低的相似性,如图6和图7所示.

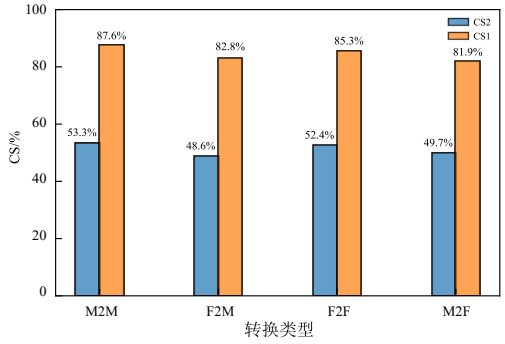


图6 Dddm-VC泛化性数据集下音色相似度结果

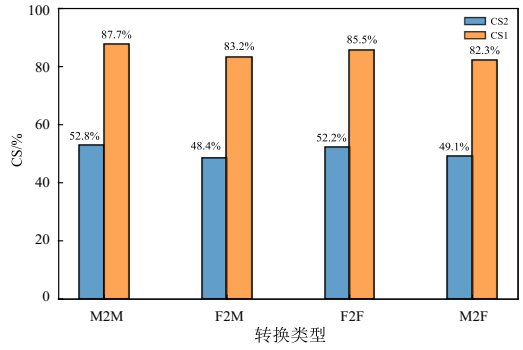


图7 Diff-HierVC泛化性数据集下音色相似度结果

基于此,本文选用Dddm-VC[24]和Diff-HierVC[25]两种基于扩散的语音转换模型构建的数据集进行泛化性能测试. 首先基于VCTK数据集按照4.1节的方法构建测试数据集Dddm-VCTK和Diff-VCTK,泛化性实验结果如表10所示. 可以得出以下结论:

表10 泛化性实验结果 单位:%

数据集	Type	CS	EER
Dddm-VCTK	M2M	72.1	22.12
	F2M	71.3	
	F2F	73.2	
	M2F	70.8	
	平均值	71.8	
Diff-VCTK	M2M	71.2	22.27
	F2M	70.4	
	F2F	74.3	
	M2F	70.6	
	平均值	71.6	

(1)本文所提出Re-VCGS方法对其他语音转换具有一定的泛化能力,在构建的Dddm-VCTK和Diff-VCTK数据集下测试结果可以达到70%以上的音色相似度.

(2)同性别间的还原效果相较于异性别间的还原效果更好,可能是因为Re-VCGS方法中使用解码器生成语音时音高的变换只采用线性变换,不足以充分模拟异性别间的音高转换.

针对泛化性实验,我们也邀请了20个人参与了主观性测评,实验结果如表11所示. 由表11中数据可以看出,本文方法在Dddm-VCTK和Diff-VCTK测试集上也能保持还原语音的自然性以及与源说话人的相似性. 在F2F、M2M、F2M、M2F这四种泛化性还原方案中,本文方法在多数还原场景下表现优异,表明其在同性别和不同性别之间的还原都有较好的泛化性能.

表11 泛化性主观评价结果

数据集	Type	Naturalness	Similarity
Dddm-VCTK	F2M	3.12±0.06	3.38±0.09
	F2F	3.38±0.05	3.58±0.05
	M2M	3.32±0.06	3.54±0.08
	M2F	3.08±0.07	3.32±0.09
	Average	3.22±0.06	3.45±0.08
Diff-VCTK	F2M	3.02±0.08	3.26±0.09
	F2F	3.28±0.04	3.38±0.05
	M2M	3.26±0.07	3.42±0.08
	M2F	2.94±0.09	3.24±0.09
	Average	3.12±0.07	3.32±0.07

5 结论

针对深度转换语音还原问题,本文提出了一种基于对抗学习和增强优化的深度转换语音还原方法,利用对抗学习生成初步还原语音,利用增强优化网络提升还原语音质量,提高与源说话人语音的相似度. 实验分析表明本文所提出Re-VCGS方法可以实现深度转换语音的还原,对追踪真实说话人,防止语音转换技术的非法使用具有重要的实用价值.

需要指出的是,随着深度学习技术的发展,语音转换技术呈现出多样化的发展,Re-VCGS仅针对代表性的Free-VC、TriAAN-VC、BNE-PPG-VC三种声音转换算法进行了尝试,验证了深度转换语音还原的可能性. 未来工作将着重关注高泛化性、高相似度的深度转换语音还原方法.

致谢 感谢合肥工业大学高性能平台提供的硬件支持.

参考文献

[1] QIAN J W, DU H H, HOU J H, et al. Speech sanitizer: Speech content desensitization and voice anonymization[J].

IEEE Transactions on Dependable and Secure Computing, 2021, 18(6): 2631-2642.

[2] LAL SRIVASTAVA B M, VAUQUIER N, SAHIDULLAH M, et al. Evaluating voice conversion-based privacy protection against informed attackers[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2020: 2802-2806.

[3] MUKHNERI F M, WIJAYANTO I, HADIYOSO S. Voice conversion for dubbing using linear predictive coding and hidden Markov model[J]. Journal of Southwest Jiaotong University, 2020, 55(4): 33.

[4] KANAGAWA H, NOSE T, KOBAYASHI T. Speaker-independent style conversion for HMM-based expressive speech synthesis[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2013: 7864-7868.

[5] LUO Y J, HSU C C, AGRES K, et al. Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2020: 3277-3281.

[6] 许裕雄, 李斌, 谭舜泉, 等. 语音深度伪造及其检测技术研究进展[J]. 中国图象图形学报, 2024, 29(8): 2236-2268.
XU Y X, LI B, TAN S Q, et al. Research progress on speech deepfake and its detection techniques[J]. Journal of Image and Graphics, 2024, 29(8): 2236-2268. (in Chinese)

[7] TAK H, TODISCO M, WANG X, et al. Automatic speaker verification spoofing and deepfake detection using Wav2vec 2.0 and data augmentation[C]//The Speaker and Language Recognition Workshop. ISCA, 2022: 112-119.

[8] WANG L, YEOH B, NG J W. Synthetic voice detection and audio splicing detection using SE-Res2Net-conformer architecture[C]//2022 13th International Symposium on Chinese Spoken Language Processing. Piscataway: IEEE, 2022: 115-119.

[9] YUE F, CHEN J L, SU Z P, et al. Audio Spoofing Detection Using Constant-Q Spectral Sketches and Parallel-Attention SE-ResNet[M]//Computer Security-ESORICS 2022. Cham: Springer Nature Switzerland, 2022: 756-762.

[10] XUE J X, ZHOU H, SONG H W, et al. Cross-modal information fusion for voice spoofing detection[J]. Speech Communication, 2023, 147: 41-50.

[11] REN Y Z, ZHU H C, ZHAI L M, et al. Who is speaking actually? Robust and versatile speaker traceability for voice conversion[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM, 2023: 8674-8685.

[12] LIU S X, CAO Y W, WANG D S, et al. Any-to-many voice conversion with location-relative sequence-to-sequence modeling[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1717-1728.

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

[14] LI J Y, TU W P, XIAO L. FreeVC: Towards high-quality text-free one-shot voice conversion[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2023: 1-5.

[15] 章子旭, 简志华. 采用双重交换表示分离的任意说话人语音转换[J]. 电子学报, 2024, 52(6): 2141-2150.
ZHANG Z X, JIAN Z H. Any-to-any voice conversion using double exchange representation separation[J]. Acta Electronica Sinica, 2024, 52(6): 2141-2150. (in Chinese)

[16] YAMAGISHI J, VEAUX C, MACDONALD K, et al. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)[J]. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019: 271-350.

[17] KAMEOKA H, KANEKO T, TANAKA K, et al. StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks[C]//2018 IEEE Spoken Language Technology Workshop. Piscataway: IEEE, 2018: 266-273.

[18] 黄赟, 张帆, 郭威, 等. 一种基于数据标准差的卷积神经网络量化方法[J]. 电子学报, 2023, 51(3): 639-647.
HUANG Y, ZHANG F, GUO W, et al. A quantification method of convolutional neural network based on data standard deviation[J]. Acta Electronica Sinica, 2023, 51(3): 639-647. (in Chinese)

[19] CHEN Y P, DAI X Y, LIU M C, et al. Dynamic convolution: Attention over convolution kernels[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 11027-11036.

[20] ZHANG Q L, YANG Y B. SA-net: Shuffle attention for deep convolutional neural networks[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2021: 2235-2239.

[21] 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. Piscataway: IEEE, 2018: 4879-4883.

[22] 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.

[23] DESPLANQUES B, THIENPONDT J, DEMUYNCK K. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification[C]//Interspeech 2020. ISCA, 2020: 3830-3834.

[24] CHOI H Y, LEE S H, LEE S W. DDDM-VC: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16): 17862-17870.

[25] CHOI H Y, LEE S H, LEE S W. Diff-HierVC: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation[C]//INTERSPEECH 2023. ISCA, 2023: 2283-2287.

作者简介

苏兆品 女，1983 年 8 月生，山东菏泽人. 副教授，硕士生导师，CCF 会员. 2004 年和 2008 年在合肥工业大学分别获得学士和博士学位. 主要研究方向为音频信息隐藏、深度学习和进化计算. 中国电子学会会员编号：E190027825M.
E-mail: szp@hfut.edu.cn

周晓琳 男，1999 年 10 月生，安徽蚌埠人. 硕士研究生. 2022 年在淮北师范大学获得学士学位. 主要研究方向为面向转换语音的溯源关键技术.
E-mail: 2022171228@mail.hfut.edu.cn

张国富 男，1979 年 3 月生，安徽合肥人. 教授，硕士生导师，CCF、CAA 会员. 2002 年和 2008 年在合肥工业大学分别获得学士和博士学位. 现为工业安全与应急技术安徽省重点实验室副主任. 主要研究方向为基于搜索的软件工程、音频安全和进化计算等.
E-mail: zgf@hfut.edu.cn

廉晨思 女，硕士，高级工程师. 主要研究方向为声纹鉴定.
E-mail: lchsi324@163.com

王年松 男，2002 年毕业于中国刑事警察学院公共安全图像专业，正高级. 主要研究方向为多媒体取证.
E-mail: 28640145@qq.com

岳 峰 1981 年 2 月生，安徽合肥人. 副研究员，硕士生导师. 2004 年、2009 年和 2015 年在合肥工业大学分别获得学士、硕士和博士学位. 主要研究方向为软件工程、音频信息隐藏和进化计算.
E-mail: yuefeng@huft.edu.cn