采用双重交换表示分离的任意说话人语音转换

章子旭,简志华*
(杭州电子科技大学通信工程学院,浙江杭州 310018)

摘 要: 在任意说话人语音转换中,训练阶段通常采用编码器对同一说话人语音进行解耦,然后用解码器进行自重构,而转换阶段的解码器是对源语音的内容信息与目标语音的个性特征进行耦合,因此解码器在转换阶段与训练阶段会存在性能失配现象,影响语音转换性能. 对此提出了一种采用双重交换表示分离的语音转换方法DERS-VC(Double Exchange Representation Separation Voice Conversion). 该方法在训练阶段的自重构过程中,用同一说话人的语音模拟不同说话人的语音进行自监督训练. 训练过程引入转换不变损失和周期循环一致损失,通过双重交换表示分离的循环过程使自重构语音与原始语音更加逼近. 实验结果表明,DERS-VC算法在梅尔倒谱距离(Mel-Cepstral Distortion,MCD)上比现有的AGAIN-VC(Activation Guidance and Adaptive Instance Normalization Voice Conversion)转换方法平均降低了4.03%,平均意见分(Mean Opinion Score,MOS)提升了3.62%,转换语音质量和相似度都有提高. 这说明,通过双重交换表示分离的方法可以更好地训练解码器,实现更好性能的任意说话人之间的语音转换.

关键词: 语音转换;任意说话人;双重交换;表示分离
基金项目: 国家自然科学基金(No.61201301, No.61772166)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)06-2141-10
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230246

Any-to-Any Voice Conversion Using Double Exchange Representation Separation

ZHANG Zi-xu, JIAN Zhi-hua*
(*School of Communication Engineering*, *Hangzhou Dianzi University*, *Hangzhou*, *Zhejiang* 310018, *China*)

Abstract: In any-to-any voice conversion, the encoder was usually utilized to disentangle the same speaker's speech and then the decoder was used for self-reconstruction in the training phase, but the decoder in the conversion phase coupled the content information of source speech and the personality characteristics of target speech. Therefore, there existed performance mismatch between the decoder in the conversion phase and the training phase, which deteriorated the performance of voice conversion. This paper proposed a voice conversion method named DERS-VC (Double Exchange Representation Separation Voice Conversion) using double exchange representation separation. In self-reconstruction process of the training phase, the proposed method applied the speech of the same speaker to simulate the voice of different target speakers for self-supervised training. Meanwhile, the conversion invariance loss and the cycle consistency loss were introduced, and the cycle process of separation was conducted by double exchange representation separation to make the self-reconstructed speech closer to the original speech. The experimental results demonstrated that DERS-VC had an average reduction of 4.03% in MCD (Mel-Cepstral Distortion), and had an increment of 3.62% in MOS (Mean Opinion Score), compared with the AGAIN-VC (Activation Guidance and Adaptive Instance Normalization Voice Conversion) method, and the quality and similarity of converted speech both had been improved. This shows that the method of double exchange representation separation can decrease the mismatch of the decoder and improve the performance of any-to-any voice conversion.

Key words: voice conversion; any-to-any; double exchange; representation separation
Foundation Item(s): National Natural Science Foundation of China (No.61201301, No.61772166)

收稿日期:2023-03-17;修回日期:2023-06-16;责任编辑:赵克
*通讯作者:简志华

1 引言

语音转换是在保留原始语音内容的前提下,将源语音中说话人特征转换成特定的目标说话人特征的一项技术[1]. 语音转换技术有着非常广泛的应用,在发音辅助、语音增强、信息安全等方面都起到了重要的作用.

早期的语音转换主要研究平行语料下的语音转换,即在语音内容相同但说话人个性特征不同的语音之间进行转换. 通过提取源和目标语音的特征,建立源和目标语音之间的特征映射. 早期语音转换技术包括矢量量化(Vector Quantization, VQ)[2]、高斯混合模型(Gaussian Mixed Model, GMM)[3]、动态核偏最小二乘回归(Dynamic Kernel Partial Least Squares regression, DKPLS)[4]、非负矩阵分解(Non-negative Matrix Factorization, NMF)[5]等方法. 但收集平行语料库并在源和目标语音之间进行时间对齐通常费时费力,不利于实际应用.

近年来,非平行语料之间的语音转换得到了广泛的研究,比如基于语音后验图(Phonetic Posterior Grams, PPG)[6]、深度神经网络(Deep Neural Networks, DNN)[7]等模型的转换方法. 其中生成对抗网络(Generative Adversarial Nets, GAN)[8]能够在不显示概率密度分布的情况下直接学习接近目标的生成分布,其派生出的循环生成对抗网络(Cycle Generative Adversarial Network, CycleGAN)[9]语音转换利用生成器和鉴别器进行对抗式学习,同时引入对抗损失、循环一致损失和身份映射损失学习正向映射和反向映射,实现了高相似度的一对一转换. 星形生成对抗网络(Star Generative Adversarial Network, StarGAN)[10]语音转换针对多说话人之间的转换场景实现了非平行语料下的多对多语音转换,但不论是一对一的语音转换,还是多对多的转换,都局限于训练集内,在面对训练集外的语音需要转换时,即任意说话人之间的语音转换,转换性能有所下降.

目前,对于任意说话人之间的语音转换广泛使用解耦的方法,基于解耦的模型将语音视为内容信息和说话人个性特征的组成,从而将两者分离并进行说话人个性特征变换,实现语音转换. 其中主要方法有AdaIN-VC(Adaptive Instance Normalization Voice Conversion)[11]、AutoVC[12]. AdaIN-VC通过自适应实例归一化技术分离说话人信息和内容信息并重新解码得到目标Mel谱图. AutoVC利用端到端损失(Generalized End-to-End Loss, GE2E Loss)对模型进行训练,并在内容编码器上设计一个信息瓶颈,经过仔细调整的瓶颈特征将说话人信息与内容分离. 这些方法都实现了任意说话人之间的语音转换,但也都有一定的不足之处. AdaIN-VC需要使用两个不同的编码器分别提取说话人嵌入和内容嵌入,即说话人编码器和内容编码器,而这两个编码器在分离说话人个性特征和内容信息的功能上有所重复,增加了系统的复杂性. AutoVC则需要用预先训练好的说话人编码器提取说话人嵌入,因此语音转换性能高度依赖说话人嵌入的准确性. 同时AutoVC中信息瓶颈的维度大小对说话人个性特征的残留或丢失非常敏感,一旦选择不当,严重影响转换性能. 针对这两个问题,AGAIN-VC[13]在仅使用一个编码器的情况下实现了任意说话人之间的语音转换,利用IN(Instance Normalization)层和激活引导层代替信息瓶颈分离了语音的内容信息和说话人个性特征. 但不管是AGAIN-VC还是AdaIN-VC和AutoVC,它们在训练阶段都使用自编码器结构,采用自重构过程进行训练,解码器都是由同一个说话人的语音训练得到,而在转换阶段,解码器需要面对源与目标两个不同的说话人语音作为输入,因此解码器在转换阶段与训练阶段会存在失配现象,影响语音转换性能.

为了解决任意说话人之间语音转换中解码器在转换阶段与训练阶段的失配问题,论文提出了一种采用双重交换表示分离的语音转换方法(DERS-VC, Double Exchange Representation Separation Voice Conversion). 该方法将原始语音与通过初次自重构得到的重构语音分别作为源与目标语音进行双重交换表示分离,通过自重构过程来构建不同说话人的语音并用于训练,整个训练过程采用转换不变损失和周期循环一致损失进行约束,解决解码器性能失配的问题. 同时双重交换过程两次更换说话人个性特征,第一次交换用于合成新语音,第二次交换用于重建语音,更好地实现说话人个性特征与内容信息的解耦,提升语音转换的性能.

2 DERS-VC语音转换

2.1 整体结构

传统的采用自编码器的语音转换在将同一个说话人原始语音Mel谱图$\boldsymbol{X}_1$生成重构语音Mel谱图$\boldsymbol{X}_2$的自重构过程中,训练的目标是将$\boldsymbol{X}_2$尽可能在某种度量准则下接近$\boldsymbol{X}_1$,从而通过训练获得网络中各个模块的参数,然后将训练好的网络用于转换语音. 但实际上,在转换阶段,源语音与目标语音分属不同说话人,解码器的耦合对象来自于两个不同的说话人,而训练解码器是使用同一个说话人的语音. 因此,解码器在训练阶段和转换阶段会存在失配现象. 为解决这一失配问题,论文采用双重交换表示分离的自编码器结构. 图1是DERS-VC语音转换系统的自编码器的训练过程整体框图,图2是该系统转换过程的流程图.

在训练阶段,源语音Mel谱图$\boldsymbol{X}_1$生成对应自重构Mel谱图$\boldsymbol{X}_2$的过程是使用AGAIN-VC的基本架构. 同

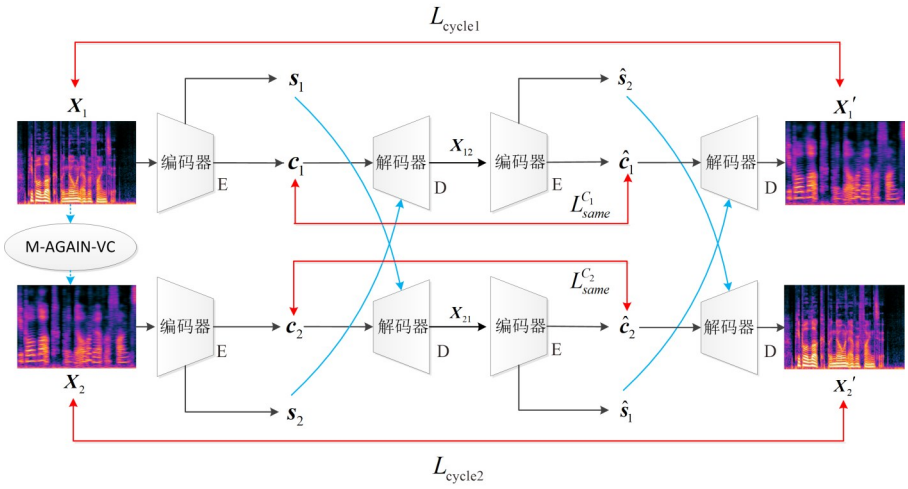


图1　DERS-VC系统的训练过程流程图

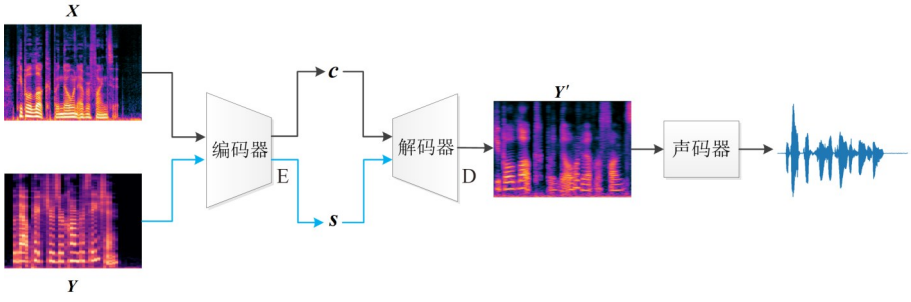


图2　DERS-VC转换阶段流程图

时为了提升自重构的性能,在这一基本架构的基础上引入表示损失,包括自内容损失和自说话人损失[14,15],使自重构过程中内容信息表示不会重叠或丢失.将改进后的AGAIN-VC记为M-AGAIN-VC(Modified AGAIN-VC),其结构如图3所示.

从图1和图3可以看出,DERS-VC模型由编码器E、解码器D以及内嵌在M-AGAIN-VC的相关编码器Er组成.编码器E将原始语音Mel谱图$\boldsymbol{X}_1$映射为内容信息,即(E:$\boldsymbol{X}_1 \rightarrow \boldsymbol{c}_1$),同时将重构语音Mel谱图$\boldsymbol{X}_2$经过编码器中IN(Instance Normalization)层获取的均值矢量$\boldsymbol{\mu}$和方差矩阵$\boldsymbol{\sigma}$作为说话人表示,即(E:$\boldsymbol{X}_2 \rightarrow \boldsymbol{s}_2$).编码器的内部结构如图4所示.

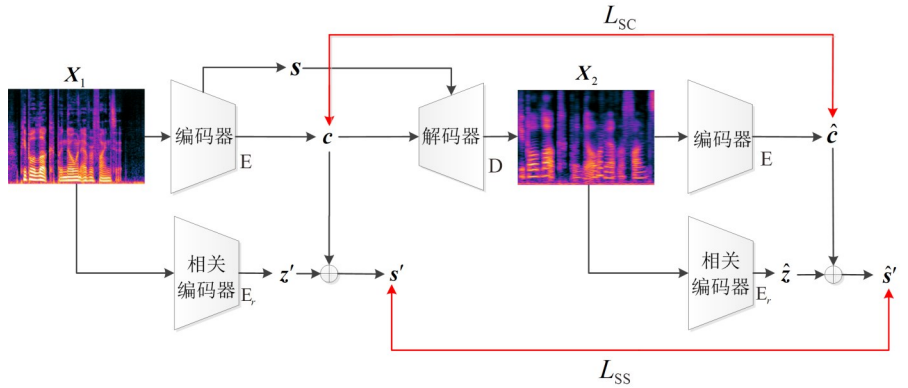


图3　M-AGAIN-VC自重构流程图

在编码器中,使用Conv1d层处理语音频率信息,在Conv1d层后使用6个ConvBlocks. ConvBlocks由Conv-Norm层、BatchNorm层、LeakyReLU层和IN层组成,其中IN层可以将语音中的说话人个性特征去除,将语音的内容信息保留[16]. Sigmoid激活函数作为信息瓶颈替代了矢量量化或降维,防止内容信息中有说话人个性特征的重叠或内容信息的丢失.

解码器作为语音转换中语音Mel谱图的合成模块,

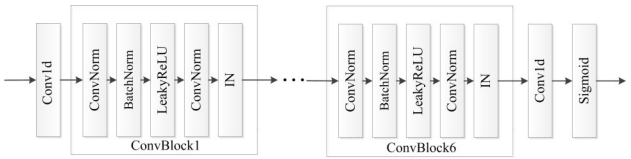


图4 编码器结构图

将内容信息表示和说话人个性特征表示合成出对应的语音Mel谱图，即($D:[\boldsymbol{c}_1,\boldsymbol{s}_2]\rightarrow\boldsymbol{X}_{12}$和$D:[\boldsymbol{c}_2,\boldsymbol{s}_1]\rightarrow\boldsymbol{X}_{21}$)，其中$\boldsymbol{X}_{12}$和$\boldsymbol{X}_{21}$都表示合成出的Mel谱图. 解码器结构如图5所示.

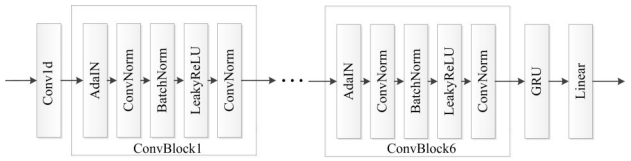


图5 解码器结构图

解码器的ConvBlocks模块中AdaIN层将说话人个性特征进行添加，最后经过GRU层和Linear层得到合成后的Mel谱图. AdaIN定义为

$$\mathrm{AdaIN}(\boldsymbol{H},\boldsymbol{\mu}(\boldsymbol{Z}),\boldsymbol{\sigma}(\boldsymbol{Z}))=\boldsymbol{\sigma}(\boldsymbol{Z})\mathrm{IN}(\boldsymbol{H})+\boldsymbol{\mu}(\boldsymbol{Z}) \tag{1}$$

其中，$\boldsymbol{H}$和$\boldsymbol{Z}$分别是源语音和目标语音的Mel谱图经过编码器卷积等计算生成的潜矢量序列. 由于在M-AGAIN-VC自重构过程中将语音与分离的内容信息映射在同一潜空间中，需要用到与编码器结构相似的相关编码器. 相关编码器使用与编码器一样的Conv1d层，之后也是使用6个ConvBlocks模块，但ConvBlocks中不需要IN层去除说话人个性特征表示. 同时相关编码器也无需Sigmoid层进行激活引导，详细结构如图6所示.

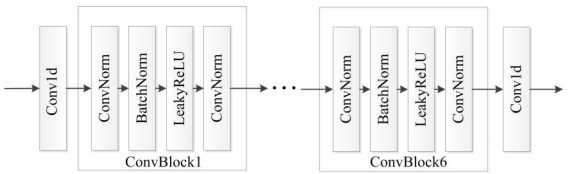


图6 相关编码器结构图

2.2 内容信息和说话人个性特征的解耦

由于在训练开始时，$\boldsymbol{X}_2$为$\boldsymbol{X}_1$的初始重构目标，因此$\boldsymbol{X}_1$与$\boldsymbol{X}_2$之间存在较大差距，为了在自重构过程中模拟不同目标之间转换的过程，将$\boldsymbol{X}_1$与$\boldsymbol{X}_2$视为两个不同说话人的语音. 假定$\boldsymbol{X}_1$是来自源说话人，$\boldsymbol{X}_2$是来自目标说话人. DERS-VC模型将Mel谱图$\boldsymbol{X}_1$与$\boldsymbol{X}_2$分别映射到内容潜空间C_1、C_2和说话人特征潜空间S_1、S_2，分别代表着输入语音经过编码器后分离为内容信息和说话人个性特征.

$$\{\boldsymbol{c}_1,\boldsymbol{s}_1\}=\{E(\boldsymbol{X}_1)\},\boldsymbol{c}_1\in C_1,\boldsymbol{s}_1\in S_1 \tag{2}$$

$$\{\boldsymbol{c}_2,\boldsymbol{s}_2\}=\{E(\boldsymbol{X}_2)\},\boldsymbol{c}_2\in C_2,\boldsymbol{s}_2\in S_2 \tag{3}$$

其中，$\boldsymbol{X}_1$与$\boldsymbol{X}_2$分别表示输入源语音的Mel谱图和目标语音的Mel谱图.

论文采用转换不变损失来实现内容信息表示与说话人个性特征表示的解耦并避免信息的重叠或丢失问题[17]，在交换不同说话人个性特征但保留内容信息不变的情况下进行转换，转换后语音的内容信息尽可能与源语音的内容信息一致，同时相同说话人个性特征表示之间也尽可能保持一致. 转换不变损失表示为

$$L_{\text{same}}^{C_1}=\left\|\boldsymbol{c}_1-\hat{\boldsymbol{c}}_1\right\|_1,L_{\text{same}}^{C_2}=\left\|\boldsymbol{c}_2-\hat{\boldsymbol{c}}_2\right\|_1 \tag{4}$$

$$L_{\text{same}}^{S_1}=\left\|\boldsymbol{s}_1-\hat{\boldsymbol{s}}_1\right\|_1,L_{\text{same}}^{S_2}=\left\|\boldsymbol{s}_2-\hat{\boldsymbol{s}}_2\right\|_1 \tag{5}$$

其中，$\|\cdot\|_1$表示1-范数，$\hat{\boldsymbol{c}}_1$和$\hat{\boldsymbol{s}}_1$分别表示源语音经编码器第二次分解后得到的内容信息和说话人个性特征，$\hat{\boldsymbol{c}}_2$和$\hat{\boldsymbol{s}}_2$分别表示目标语音经编码器第二次分解后得到的内容信息和说话人个性特征.

总体转换不变损失表示为

$$L_{\text{same}}=L_{\text{same}}^{C}+L_{\text{same}}^{S}=\sum_{n=1}^{2}L_{\text{same}}^{C_n}+\sum_{n=1}^{2}L_{\text{same}}^{S_n} \tag{6}$$

其中，$L_{\text{same}}^{C_n}$和$L_{\text{same}}^{S_n}$表示内容转换不变损失和说话人转换不变损失.

2.3 双重交换与循环过程

DERS-VC在自重构过程中模拟不同目标进行语音转换，结合源语音内容$\boldsymbol{c}_1$和目标语音个性特征$\boldsymbol{s}_2$进行语音转换. 我们使用一个循环过程来训练编解码器，在此过程中两次交换内容信息和说话人个性特征，使用周期循环一致性作为损失函数L_{cycle}.

在第一次交换中，由于将源语音与重构语音视为不同语音，其Mel谱图经过编码器可以得到一组不同的内容信息$\{\boldsymbol{c}_1,\boldsymbol{c}_2\}$和一组不同的说话人个性特征$\{\boldsymbol{s}_1,\boldsymbol{s}_2\}$. 然后第一次交换说话人个性特征$\{\boldsymbol{s}_1,\boldsymbol{s}_2\}$与内容信息$\{\boldsymbol{c}_1,\boldsymbol{c}_2\}$，并经过解码器生成相应的Mel谱图$\{\boldsymbol{X}_{12},\boldsymbol{X}_{21}\}$，其中生成的$\boldsymbol{X}_{12}$具有目标说话人的个性特征，即生成在目标域中，而$\boldsymbol{X}_{21}$具有源说话人的个性特征. 这个过程可以表示为

$$\boldsymbol{X}_{12}=D(\boldsymbol{c}_1,\boldsymbol{s}_2) \tag{7}$$

$$\boldsymbol{X}_{21}=D(\boldsymbol{c}_2,\boldsymbol{s}_1) \tag{8}$$

在第二次交换中，$\boldsymbol{X}_{12}$和$\boldsymbol{X}_{21}$再次经过编码器分别得到内容信息$\{\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{c}}_2\}$和说话人个性特征$\{\hat{\boldsymbol{s}}_1,\hat{\boldsymbol{s}}_2\}$，和前面一样进行交换并再次生成Mel谱图，即

$$\boldsymbol{X}_1'=D(\hat{\boldsymbol{c}}_1,\hat{\boldsymbol{s}}_1) \tag{9}$$

$$\boldsymbol{X}_2'=D(\hat{\boldsymbol{c}}_2,\hat{\boldsymbol{s}}_2) \tag{10}$$

经过两次交换后，输出得到的$\boldsymbol{X}_1'$和$\boldsymbol{X}_2'$应与对应输入的$\boldsymbol{X}_1$和$\boldsymbol{X}_2$尽可能保持一致，并使用周期循环一致损失L_{cycle}来进一步约束，表示为

$$L_{\text{cycle1}}=\left\|\boldsymbol{X}_1-\boldsymbol{X}_1'\right\|_1, L_{\text{cycle2}}=\left\|\boldsymbol{X}_2-\boldsymbol{X}_2'\right\|_1 \quad (11)$$

$$L_{\text{cycle}}=L_{\text{cycle1}}+L_{\text{cycle2}} \quad (12)$$

2.4 自重构过程

表示损失包括自内容损失和自说话人损失,将自重构语音的内容信息表示和说话人个性特征表示与原始语音映射在同一潜空间中,使两者的内容信息和说话人个性特征分别保持一致,同时确保分离的内容信息和说话人个性特征相互独立,从而提高转换语音的质量.

在DERS-VC自重构过程中,内容信息和说话人个性特征输入解码器得到重构语音Mel谱图,重构语音Mel谱图并再次经过编码器后得到重构语音的内容信息表示为$\boldsymbol{c}'$,$\boldsymbol{c}'$与原始语音内容信息$\boldsymbol{c}$之间构成自内容损失L_{SC},表示为

$$L_{\text{SC}}=\left\|\hat{\boldsymbol{c}}-\boldsymbol{c}\right\|_1 \quad (13)$$

同时在图3中,M-AGAIN-VC重构语音Mel谱图时,经过相关编码器得到重构语音的相关说话人个性特征信息$\widehat{\boldsymbol{s}}'$,与原始语音经过相关编码器得到的相关说话人个性特征$\boldsymbol{s}'$之间构成自说话人损失L_{SS},表示为

$$L_{\text{SS}}=\left\|\widehat{\boldsymbol{s}}'-\boldsymbol{s}'\right\|_1 \quad (14)$$

则表示损失L_{SR}为

$$L_{\text{SR}}=L_{\text{SC}}+L_{\text{SS}} \quad (15)$$

因此,DERS-VC的总体损失函数可以表示为

$$L_{\text{total}}=L_{\text{rec}}+\lambda_{\text{same}}L_{\text{same}}+\lambda_{\text{cycle}}L_{\text{cycle}}+\lambda_{\text{SR}}L_{\text{SR}} \quad (16)$$

其中,λ_{same}为转换不变损失的权重,λ_{cycle}为周期循环一致损失的权重,λ_{SR}为表示损失的权重.

3 实验及分析

3.1 实验数据及实验设置

实验使用VCTK数据集进行训练[18],该数据集收集了109个英语说话人的语音,每位说话人以各自不同的口音录制大约400个词句. 实验随机选取80个说话人,每个说话人随机选择200个语音作为训练集. 在测试方面,选取训练集内2个男声和2个女声,每个说话人取20个语音进行多对多语音转换性能测试. 为了更好地评价任意说话人之间语音转换的性能以及模型的泛化能力,实验使用VCC2018数据集进行任意说话人之间语音转换性能测试[19],同样随机选取2个男声和2个女声,每个说话人使用50个语音.

所有语音信号采样率为22 050 Hz,实验首先对语音信号进行端点检测,去除语音首尾静音部分. 在将生成的Mel谱图转换为语音时,考虑到语音生成的质量和速度,实验选用MelGAN[20]作为声码器. 根据MelGAN配置,对分帧后的语音信号进行1 024点的短时傅里叶变换,生成80 bin Mel谱图,并以连续的128帧作为系统的输入.

实验通过ADAM优化器进行训练,以初始学习率为0.000 5对DERS-VC模型进行训练,并将优化参数——矩估计的指数衰减率设置为β_1=0.9,β_2=0.999,Batch大小设置为32,训练步骤数为50 000. 实验部署在Python平台环境下,在4 GB GeForce RTX 3050 GPU上运行.

3.2 客观评价

实验将AdaIN-VC、AGAIN-VC作为基线模型进行对比,选用MCD来对比DERS-VC与基线系统的转换性能,MCD用于衡量转换后的语音与目标语音之间的频谱距离[21]. 在计算MCD的过程中首先使用动态时间规整(Dynamic Time Warping,DTW)对转换后语音与相应的目标语音进行逐帧对齐,然后再计算MCD、MCD的计算公式表示如下:

$$\text{MCD}=\frac{10\sqrt{2}}{\ln 10}\times\frac{1}{M}\sum_{m=1}^{M}\sqrt{\sum_{r=1}^{R}\left[y_m(r)-y_m'(r)\right]^2} \quad (17)$$

其中,$y_m(r)$和$y_m'(r)$分别为目标语音和转换后语音的第m帧梅尔倒谱特征矢量的第r维系数,R为梅尔倒谱系数的维数,M为总帧数.

为了进一步直观对比DERS-VC与基线系统的转换性能,实验同时采用基音频率均方根误差(F_0 Root Mean Square Error,F_0 RMSE)作为衡量语音转换性能的客观指标[22]. 在计算过程中根据先前DTW的对齐结果来计算转换后语音和对应的目标语音之间的F_0 RMSE. F_0 RMSE的计算公式表示如下:

$$F_0\,\text{RMSE}=\sqrt{\frac{1}{M}\sum_{i=1}^{M}(F_{0i}-F_{0i}')^2} \quad (18)$$

其中,F_{0i}和F_{0i}'分别表示目标说话人语音和对应的转换后语音的第i帧的基音频率,M表示语音总帧数.

实验分为三部分,第一部分为选择最佳的损失权重,包括转换不变损失权重、周期一致损失权重和表示损失权重. 第二部分测试任意说话人(训练集外)之间语音转换的性能,并与基线系统模型进行性能对照. 第三部分为DERS-VC多对多(训练集内)语音转换评测. 各个测试都分别包括女声转换到女声(F2F)、男声转换到男声(M2M)、女声转换到男声(F2M)和男声转换到女声(M2F)4种情况.

实验第一部分首先确定公式(16)总体损失函数中各个损失的权重,以便选定最优参数并进行模型训练. 实验随机选用VCC2018数据集中4组平行语音数据进行测试,包括两组男声和两组女声,每组包括50条语音. 在4种不同转换情形下进行平行语音转换并确定损失权重. 当仅考虑转换不变损失λ_{same}影响时,即$\lambda_{\text{cycle}}=0,\lambda_{\text{SR}}=0$,转换语音与目标语音的MCD在4种转

换情况下随λ_{same}变化如表1所示. 表中粗体数据表示当前取值下MCD最小.

表1 4种转换情形下λ_{same}的取值对MCD值的影响 单位:dB

λ_{same}	F2F	M2M	F2M	M2F
0.2	9.202	8.354	8.866	9.633
0.3	9.271	8.467	8.724	9.677
0.4	**9.074**	**8.191**	**8.635**	**9.427**
0.5	9.093	8.265	8.651	9.476
0.6	9.220	8.764	8.972	9.673
0.7	9.163	8.394	8.803	9.534
1	9.228	8.454	8.863	9.576

根据表1数据可以发现,在$\lambda_{same}=0.4$时,4种转换情况下的MCD均为最低,即转换性能最好,故选取$\lambda_{same}=0.4$. 在$\lambda_{same}=0.4$的前提下只考虑周期循环一致损失的影响时,即$\lambda_{SR}=0$,转换语音与目标语音的MCD在4种转换情况下随λ_{cycle}变化如表2所示. 表中粗体数据表示当前取值下MCD最小.

表2 4种转换情形下λ_{cycle}的取值对MCD值的影响 单位:dB

λ_{cycle}	F2F	M2M	F2M	M2F
0.2	9.147	8.383	8.783	9.587
0.3	9.276	8.391	8.728	9.651
0.4	9.217	8.294	8.676	9.569
0.5	**9.037**	**8.265**	**8.592**	**9.475**
0.6	9.184	8.403	8.728	9.562
1	9.161	8.551	8.992	9.794
1.5	9.217	8.796	9.172	9.819

表2数据显示,在$\lambda_{cycle}=0.5$时4种情况下MCD均比其他情况时低,且与表1最佳组对比时在F2F和M2F情况下转换性能有所提升,在M2M和F2M的情况下转换性能也保持在最佳值附近,故选取$\lambda_{cycle}=0.5$. 在当$\lambda_{same}=0.4,\lambda_{cycle}=0.5$时,转换语音与目标语音的MCD在4种转换情况下随$\lambda_{SR}$变化如表3所示. 表中粗体数据表示当前取值下MCD最小.

表3 4种转换情形下λ_{SR}的取值对MCD值的影响 单位:dB

λ_{SR}	F2F	M2M	F2M	M2F
1	9.240	8.299	8.751	9.774
1.4	9.151	8.471	8.821	9.664
1.5	9.198	8.309	8.694	9.564
1.6	9.136	8.273	8.735	9.560
1.7	**9.074**	**8.187**	**8.574**	**9.474**
2	9.259	8.383	8.797	9.602
2.5	9.283	8.412	8.654	9.648

根据表3数据可以发现,当$\lambda_{SR}=1.7$时在4种情况下MCD均为最低,同时综合表1、表2和表3的数据进行对比可以发现,当在设置$\lambda_{same}=0.4,\lambda_{cycle}=0.5,\lambda_{SR}=1.7$时,M2M、F2M情况下MCD均为最佳,且F2F和M2F情况下MCD均保持在最佳值附近,综合来看,选取$\lambda_{SR}=1.7$.

在设置$\lambda_{same}=0.4,\lambda_{cycle}=0.5,\lambda_{SR}=1.7$后,实验第二部分对DERS-VC和基线模型都使用相同的训练集进行训练,并使用相同的测试集进行任意说话人之间的转换性能测试. 实验随机选取VCC2018测试数据集中2个男声和2个女声,每个说话人包括50个语音,4种转换情形下转换语音的MCD对比如表4所示,4种转换情形下转换语音的F_0 RMSE对比如表5所示.

表4 任意说话人之间语音转换时DERS-VC与基线系统的MCD对比 单位:dB

对比系统	F2F	M2M	F2M	M2F	Avg.
AdaIN-VC	10.250	9.174	9.901	10.423	9.937
AGAIN-VC	9.706	8.786	9.426	10.223	9.535
DERS-VC	9.365	8.611	8.882	9.745	9.150

表5 任意说话人之间语音转换时DERS-VC与基线系统的F_0 RMSE对比 单位:Hz

对比系统	F2F	M2M	F2M	M2F	Avg.
AdaIN-VC	50.504	38.578	37.829	47.851	43.690
AGAIN-VC	49.375	38.041	37.289	48.256	43.240
DERS-VC	48.378	37.786	37.178	48.244	42.896

表4数据显示,DERS-VC的MCD平均值相较于AdaIN-VC和AGAIN-VC分别提高了0.787和0.385(即7.91%和4.03%). 从表5数据来看,F_0 RMSE平均值则相较于AdaIN-VC和AGAIN-VC分别提高了0.794和0.344. 这说明DERS-VC方法在任意说话人之间语音转换方面取得了更好的性能. 在跨性别语音转换中,DERS-VC的性能对比AdaIN-VC和AGAIN-VC则更有优势,而在同性别语音转换中,DERS-VC的性能也都相应有所提升.

实验另外采用PESQ指标比较DERS-VC和基线系统的语音合成质量作为客观指标之一[23]. PESQ综合音频清晰度、音量、背景噪音音频中的可变延迟或滞后、丢失、音频干扰5个方面进行打分,PESQ评分值介于–0.5到4.5的分数,分数越高代表语音质量越好. 同时实验中将训练模型总时间除以模型训练步数记为时间复杂度,通过时间复杂度比较DERS-VC与基线系统之间的算法复杂情况. PESQ和时间复杂度的实验数值对比如表6所示.

表6 DERS-VC与基线系统的PESQ和时间复杂度对比

对比系统	PESQ	时间复杂度/s
AdaIN-VC	2.013	0.304
AGAIN-VC	2.004	0.177
DERS-VC	2.185	0.465

从表6可以看出,DERS-VC的PESQ数值要优于基线系统AdaIN-VC和AGAIN-VC,但总体上而言,三者的PESQ数值都不高. DERS-VC与AdaIN-VC、AGAIN-VC一样,都使用实例归一化对语音信号的内容信息和个性特征进行解耦,分离出的内容信息中仍有少量泄漏的说话人个性特征. 其次,由于这三者都使用两阶段重构语音信号的方法,即第一阶段将源语音的声学特征转换为目标语音的声学特征,第二阶段使用声码器将转换后的声学特征转换为语音信号. 由于两阶段分开训练,会造成转换模型预测的声学特征与声码器在训练期间使用的来自真实语音的声学特征具有不同分布. 这些都会导致转换语音的质量下降,PESQ数值偏低. 综合表4、表5和表6数据,DERS-VC虽然在时间复杂度上有所提高,但转换后的语音质量和相似度都有更好的性能.

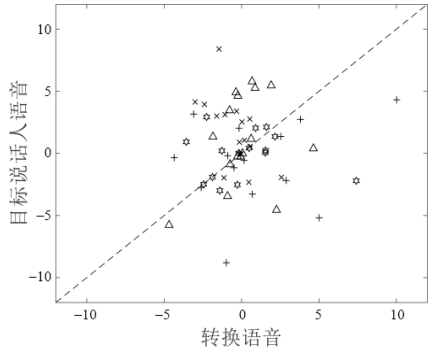
为了更加直观地对DERS-VC和基线系统的性能进行客观分析,实验采用梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient,MFCC)匹配度作为另一项评价指标[24]. 实验以相同源、目标语音作为转换对象,分别从经过AdaIN-VC、AGAIN-VC和DERS-VC的转换语音中提取相同帧的MFCC特征参数. 以转换后语音的MFCC参数作为横坐标,对应的目标语音MFCC参数作为纵坐标进行绘图. 若转换语音与目标语音的MFCC参数有较高匹配度,则两者之间的MFCC参数分布将趋向45°线. DERS-VC和基线系统的MFCC参数匹配情况如图7所示.

从图7可以看出,DERS-VC模型的转换语音与目标语音的MFCC参数更靠近于45°线,而基线模型的分布相对分散,这一结果表明本文提出的DERS-VC模型转换的语音与目标语音更加匹配,MFCC相似度更高,在语音特征相似度方面明显优于其他各个基准模型.

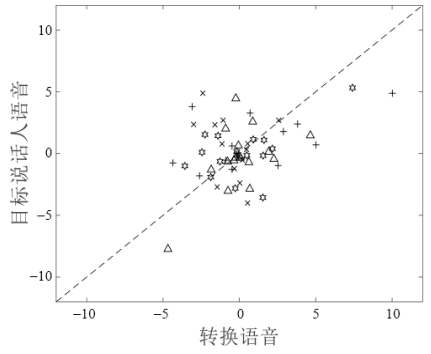
在任意说话人之间语音转换实验中,通过去除DERS-VC中使用的损失函数进行消融实验,并对比MCD和F_0 RMSE结果,实验数值对比如表7所示.

通过表7数据可以得出,当使用完整的损失函数时MCD和F_0 RMSE都比删除其中任何一个损失函数好得多,且当删除转换不变损失或周期一致损失时,转换性能下降明显,这也说明在自重构过程中额外使用双重交换过程模拟不同说话人之间语音转换的过程对于语音表示的解耦能力有所提高,从而提升了语音转换的性能,极大地改善了自重构过程中的解码器在训练阶段与转换阶段的失配问题.

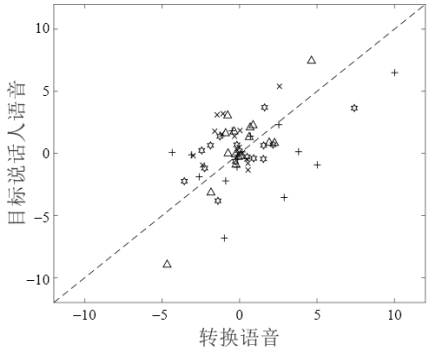
实验第三部分使用训练集中的语音进行多对多语音转换测试,实验随机选取VCTK训练数据集中2个男声和2个女声,每个说话人包括20个语音. DERS-VC与基线模型在多对多语音转换的MCD、F_0 RMSE和PESQ对比如表8所示.



(*a*) AdaIN-VC的MFCC参数匹配图



(*b*) AGAIN-VC的MFCC参数匹配图



(*c*) DERS-VC的MFCC参数匹配图

图7　3种转换方法的MFCC参数匹配对比图

表7　DERS-VC消融实验中的MCD和F_0 RMSE对比

对比不同损失	MCD/dB	F_0 RMSE/Hz
DERS-VC *w/o* same Loss	9.553	43.697
DERS-VC *w/o* Cycle Loss	9.496	43.628
DERS-VC *w/o* SR Loss	9.342	43.574
DERS-VC	9.150	42.896

通过表8的数据可以发现,DERS-VC相比于基线模型在多对多语音转换情况下也都有更好的性能,MCD数值有大幅的下降,语音转换的质量也有提高.

表8 多对多语音转换DERS-VC与基线系统的性能对比

对比系统	MCD/dB	F_0 RMSE/Hz	PESQ
AdaIN-VC	9.829	45.679	1.786
AGAIN-VC	9.120	45.774	1.967
DERS-VC	9.062	44.391	2.032

总体来说，DERS-VC模型不论是在任意说话人之间还是多对多语音转换下，相比于基线模型，在MCD、F_0 RMSE和PESQ的客观指标下性能都有提高，改善了语音转换效果.

3.3 主观评价

实验选用反映语音质量的平均意见分MOS对转换语音进行主观评价和测试. MOS意见分将语音质量分为5个等级[25]，从低到高依次表示语音质量的提高，其中1表示语音有较大失真，语音质量很差，2表示语音有部分失真且难以听懂，3表示语音有少部分失真可以部分听懂，4表示较少失真且语音大部分内容可以听懂，5表示语音完全不失真，语音内容完全可以听懂. 测试者在实验环境下对转换语音进行打分，判断转换后语音的质量. 本文在F2F、F2M、M2M、M2F 4种转换情形下进行MOS分评测，每种情况随机挑选30条转换后的语音，20名受试者对转换语音进行打分，MOS分值越高，代表语音质量越好，任意说话人之间的语音转换实验结果如图8所示，多对多语音转换实验结果如图9所示.

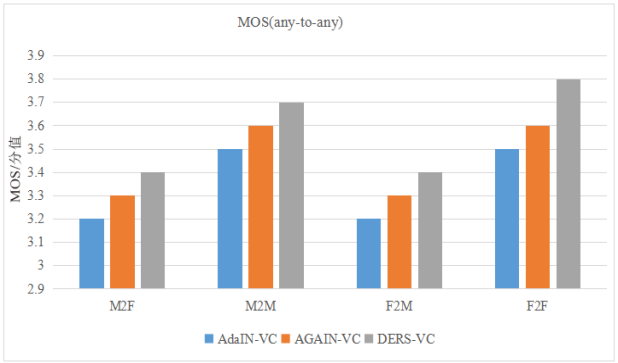


图8 4种情形下任意说话人之间转换语音的MOS比较

通过图8和图9可以发现，不论是任意说话人之间还是多对多语音转换实验中，DERS-VC在4种情况下的MOS都有所提升，可以反映出语音质量的提高. 而且在同性间语音转换时，DERS-VC取得了更高的MOS，说明DERS-VC在同性间语音转换有着更好的性能，转换语音有着更高的语音质量. 在跨性别语音转换中，相比于基线系统DERS-VC也有显著的提高.

另外实验选用ABX来评测转换语音与目标语音的相似度[25]. 测试者对源语音、目标语音和转换语音进行试听，当转换语音与源语音接近时积0分，当与目标语音接近时积1分，累计得出所有待测语音总分后除以测试语音个数得到ABX分. 实验分别对DERS-VC与基线模型在四种不同情况下转换的语音进行ABX测试，每种情况下随机挑选30条目标语音和相应的30条转换后语音，20名受试者对转换语音打分，ABX分值越高，代表语音相似度越高. 任意说话人之间的语音转换实验结果如图10所示，多对多语音转换实验结果如图11所示.

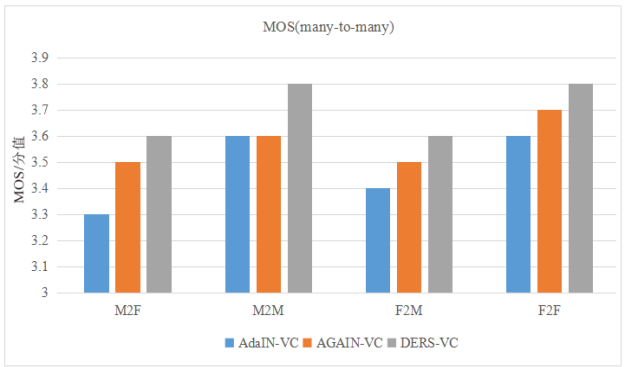


图9 4种情形下多对多转换语音的MOS比较

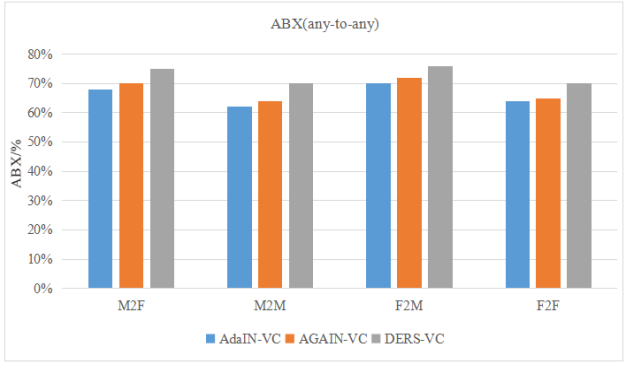


图10 四种情形下任意说话人之间转换语音的ABX比较

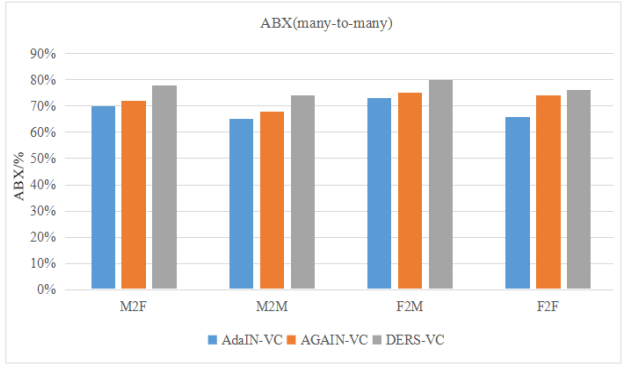


图11 四种情形下多对多转换语音的ABX比较

通过图10和图11可以发现，在任意说话人之间以及多对多语音转换中，DERS-VC在ABX测试中都取得了更高的评分. 由于异性之间的说话人个性特征差异

明显,因此在跨性别转换时ABX分值更高,表示转换得更加明显,DERS-VC对比基线系统能取得更佳的转换相似度. 而在同性之间的语音转换相似度上,DERS-VC也有较好的性能,说明该方法在低频到低频、高频到高频的语音转换上表现更好.

4 结论

论文提出了一种双重交换表示分离的语音转换方法DERS-VC,该方法无需收集平行语料,可以实现任意说话人之间的语音转换. DERS-VC的自重构训练采用语音表示的双重交换和循环过程,引入转换不变损失函数和周期一致损失函数来限制内容信息和说话人个性特征不重叠,不仅解决了解码器失配导致的转换性能下降的问题,也改善了解耦过程中内容信息和说话人个性特征重叠或丢失的问题. 同时在自重构过程中引入表示损失函数,使转换语音更好地保留源语音中内容信息和目标语音中的说话人特征,提高了转换语音的质量. 主观和客观实验结果都表明DERS-VC方法在任意说话人之间和多对多语音转换性能上都有显著的提升,说明DERS-VC方法有效地提升了转换的整体性能,提高了转换语音的相似度和自然度.

参考文献

[1] SISMAN B, YAMAGISHI J, KING S, et al. An overview of voice conversion and its challenges: From statistical modeling to deep learning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 132-157.

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

[3] 徐宁, 杨震, 张玲华. 基于状态空间模型的子频带语音转换算法[J]. 电子学报, 2010, 38(3): 646-653.
XU N, YANG Z, ZHANG L H. Sub-band voice morphing algorithm based on state-space model[J]. Acta Electronica Sinica, 2010, 38(3): 646-653. (in Chinese)

[4] HELANDER E, SILEN H, VIRTANEN T, et al. Voice conversion using dynamic kernel partial least squares regression[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(3): 806-817.

[5] WU Z Z, VIRTANEN T, CHNG E S, et al. Exemplar-based sparse representation with residual compensation for voice conversion[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10): 1506-1521.

[6] SUN L F, LI K, WANG H, et al. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training[C]//2016 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2016: 1-6.

[7] HASHIMOTO T, SAITO D, MINEMATSU N. Many-to-many and completely parallel-data-free voice conversion based on eigenspace DNN[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(2): 332-341.

[8] ALAA Y, ALFONSE M, AREF M M. A survey on generative adversarial networks based models for many-to-many non-parallel voice conversion[C]//2022 5th International Conference on Computing and Informatics (ICCI). Piscataway: IEEE, 2022: 221-226.

[9] KANEKO T, KAMEOKA H, TANAKA K, et al. Cyclegan-VC2: Improved cyclegan-based non-parallel voice conversion[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2019: 6820-6824.

[10] 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 (SLT). Piscataway: IEEE, 2018: 266-273.

[11] CHOU J C, LEE H Y. One-shot voice conversion by separating speaker and content representations with instance normalization[C]//20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019). Piscataway: IEEE, 2019: 664-668.

[12] QIAN K Z, ZHANG Y, CHANG S Y, et al. Autovc: Zero-shot voice style transfer with only autoencoder loss[C]//36th International Conference on Machine Learning (ICML). Piscataway: IEEE, 2019: 5210-5219.

[13] CHEN Y H, WU D Y, WU T H, et al. Again-VC: A one-shot voice conversion using activation guidance and adaptive instance normalization[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2021: 5954-5958.

[14] WANG Q Q, ZHANG X L, WANG J Z, et al. DRVC: A framework of any-to-any voice conversion with self-supervised learning[C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 3184-3188.

[15] DANG T, TRAN D, CHIN P, et al. Training robust zero-shot voice conversion models with self-supervised fea-

tures[C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 6557-6561.

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

[17] WANG Y Y, SU J Q, FINKELSTEIN A, et al. Controllable speech representation learning via voice conversion and AIC loss[C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 6682-6686.

[18] WANG Z C, XIE Q C, LI T, et al. One-shot voice conversion for style transfer based on speaker adaptation[C]// ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 6792-6796.

[19] KANEKO T, KAMEOKA H, TANAKA K, et al. Maskcyclegan-VC: Learning non-parallel voice conversion with filling in frames[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2021: 5919-5923.

[20] SONG K, CONG J, WANG X S, et al. Robust MelGAN: A robust universal neural vocoder for high-fidelity TTS[C]//2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). Piscataway: IEEE, 2022: 71-75.

[21] ZHAO X T, LIU F, SONG C H, et al. Disentangling content and fine-grained prosody information via hybrid ASR bottleneck features for voice conversion[C]// ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2022: 7022-7026.

[22] LEE S H, NOH H R, NAM W J, et al. Duration controllable voice conversion via phoneme-based information bottleneck[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 1173-1183.

[23] 王师琦, 曾庆宁, 龙超, 等. 语音增强与检测的多任务学习方法研究[J]. 计算机工程与应用, 2021, 57(20): 197-202.
WANG S Q, ZENG Q N, LONG C, et al. Multi-task learning for speech enhancement and detection[J]. Computer Engineering and Applications, 2021, 57(20): 197-202. (in Chinese)

[24] 仲伟峰, 方祥, 范存航, 等. 深浅层特征及模型融合的说话人识别[J]. 声学学报, 2018, 43(2): 263-272.
ZHONG W F, FANG X, FAN C H, et al. Fusion of deep shallow features and models for speaker recognition[J]. Acta Acustica, 2018, 43(2): 263-272. (in Chinese)

[25] 车滢霞, 俞一彪. 约束条件下的结构化高斯混合模型及非平行语料语音转换[J]. 电子学报, 2016, 44(9): 2282-2288.
CHE Y X, YU Y B. Non-parallel corpora voice conversion based on structured Gaussian mixture model under constraint conditions[J]. Acta Electronica Sinica, 2016, 44(9): 2282-2288. (in Chinese)

作者简介

章子旭 男,1999年3月出生于浙江省杭州市. 现为杭州电子科技大学通信工程学院硕士研究生. 主要研究方向为语音转换.
E-mail: 18757179533@163.com

简志华 男,1978年出生于江西省新余市. 博士,现为杭州电子科技大学通信工程学院副教授、硕士生导师. 主要研究方向为智能语音处理、语音转换、语音鉴伪以及语音中的隐私保护.
E-mail: jianzh@hdu.edu.cn