采用表示分离自编码器的任意说话人语音转换

简志华，章子旭
（杭州电子科技大学通信工程学院，浙江 杭州 310018）

摘 要：针对非平行语料库下任意说话人之间的语音转换存在语言内容信息和说话人个性特征难以分离，从而导致语音转换的性能不佳的问题，提出了一种采用表示分离自编码器的语音转换方法 RSAE-VC。该方法将语音信号的说话人个性特征视为时不变，而将内容信息视为时变，利用编码器中的实例归一化和激活引导层将两者进行分离，再由解码器将源语音的内容信息与目标语音的个性特征进行合成，从而生成转换后的语音。实验结果表明，RSAE-VC 在梅尔倒谱距离上比现有的 AGAIN-VC 转换方法平均降低了 3.11%，在基音频率均方根误差上降低了 2.41%，MOS 分和 ABX 值分别提升了 5.22%和 8.45%。RSAE-VC 方法通过自内容损失进行约束使语音更好地保留内容信息，通过自说话人损失将说话人个性特征更好地从语音中分离，可以确保说话人个性特征尽少地遗留在内容信息中，从而提高语音转换性能。

关键词：语音转换；表示分离；自适应实例归一化；自内容损失；自说话人损失
中图分类号：TP391.42
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2024044

Any-to-any voice conversion using representation separation auto-encoder

JIAN Zhihua, ZHANG Zixu
School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: In view of the problem that it was difficult to separate speaker personality characteristics from semantic content information in any-to-any voice conversion under non-parallel corpus, which led to unsatisfied performance, a voice conversion method, called RSAE-VC (representation separation auto-encoder voice conversion) was proposed. The speaker's personality characteristics in the speech were regarded as time invariant and the content information as time variant, and the instance normalization and activation guidance layer were used in the encoder to separate them from each other. Then the content information of the source speech and the personality characteristics of the target one was utilized to synthesize the converted speech by the decoder. The experimental results demonstrate that RSAE-VC has an average reduction of 3.11% and 2.41% in Mel cepstral distance and root mean square error of pitch frequency respectively, and has an increasement of 5.22% in MOS and 8.45% in ABX, compared with the AGAIN-VC (activation guidance and adaptive instance normalization voice conversion) method. In RSAE-VC, self-content loss is applied to make the converted speech reserve more content information, and self-speaker loss is used to separate the speaker personality characteristics from the speech better, which ensure the speaker personality characteristics be left in the content information as little as possible, and the conversion performance is improved.

Keywords: voice conversion, representation separation, adaptive instance normalization, self-content loss, self-speaker loss

收稿日期：2023−10−12；**修回日期：**2024−01−17
基金项目：国家自然科学基金资助项目（No.61201301, No.61772166）
Foundation Items: The National Natural Science Foundation of China (No.61201301, No.61772166)

0　引言

语音是最重要的通信方式之一，目前语音通信的安全性已经引起人们的广泛关注，其中为了保障语音通信的隐私安全，语音转换技术已经得到学术界和产业界的持续研究和应用。语音转换是指在保留源语音内容的前提下，将源语音中说话人特征信息转换成特定的目标说话人特征信息的一项技术[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、话者隐秘通信、信息安全等方面都起到了非常重要的作用。

早期的语音转换主要研究平行语料库之间的转换，通过提取源语音和目标语音的特征，建立源语音和目标语音之间的特征映射。最早的语音转换采用矢量量化（VQ, vector quantization）模型进行特征映射[2]，通过码本进行频谱特征的映射，简单易实现，但由于将连续的语音信号进行离散分割，转换后的语音质量不佳。对于特征空间不连续的问题，文献[3]提出了高斯混合模型（GMM, Gaussian mixed model）方法，通过训练一个连续的转换函数模拟源和目标频谱特征之间的关系，提高了语音质量，但会引起过平滑问题，并且无法捕获时间和频谱的动态细节。为了考虑语音特征之间的动态相关性，文献[4]提出了动态核偏最小二乘（DKPLS, dynamic kernel partial least squares）回归方法。DKPLS 通过源语音特征核变换进行非线性建模，并串联相邻语音帧进行动态建模，但需要大量参数以及存在过平滑问题。为了解决这 2 个问题，文献[5]提出了非负矩阵分解（NMF, non-negative matrix factorization）转换方法。NMF 根据稀疏表示的原理将一个矩阵分解为 2 个非负矩阵进行特征转换，在有限训练数据的语音转换中十分有效，但 NMF 方法也局限于平行语音转换。平行语音转换在收集大型平行语料库时通常费时费力，不利于实际运用的推广。

近年来，非平行语音转换得到了广泛的研究，比如基于语音后验图（PPG , phonetic posterior gram）[6]、深度神经网络（DNN, deep neural network）[7]等转换方法，其中生成对抗网络（GAN, generative adversarial network）[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（CycleGAN, cycle generative adversarial network）[9]利用对抗损失和循环一致损失进行训练，达到了与平行语音转换相当的效果，但仅局限于一对一的语音转换。星形生成对抗网络（StarGAN, star generative adversarial network）[10]实现了非对称语料情况下的多对多语音转换，但网络在训练中仍存在模式崩溃问题，生成器无法持续学习。

一对一和多对多的转换都局限于训练集内的语音转换，在面对训练集外的语音转换时，往往性能不佳。因此，任意说话人之间的语音转换具有非常重要的现实意义。学术界也进行了很多探索，其中 AUTOVC[11]利用端到端损失进行预训练，设计了一个信息瓶颈，经过仔细调整的瓶颈特征将说话人个性特征与内容信息分离。AdaIN-VC（adaptive instance normalization voice conversion）[12]采用变分自动编码器，通过自适应实例归一化技术分离说话人信息和内容信息。这些方法都实现了任意说话人之间的语音转换，但也有一定的不足之处。AUTOVC 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。AdaIN-VC 使用 2 个独立的编码器分别提取说话人嵌入和内容嵌入，但 2 个编码器在功能上作用上有所重复且系统训练复杂。

目前，任意说话人之间的语音转换普遍使用内容编码器和说话人编码器，内容编码器将语音内容信息从语音中分离并映射到潜空间中，说话人编码器将语音的说话人个性特征从语音中分离并映射到潜空间中，但在训练过程中需要同时训练 2 个编码器，且内容编码器和说话人编码器的作用类似，这增加了训练的难度和复杂度，需要消耗大量资源，完全可以只用一个编码器来分离内容和说话人嵌入。AGAIN-VC（activation guidance and adaptive instance normalization voice conversion）[13]仅通过单个编码器即可分离内容和说话人嵌入，但 Again-VC 仅通过 Mel 谱图之间的差异进行约束，只能使转换得到的 Mel 谱图之间差异最小，会使语音中的内容和说话人特征信息分离不够彻底，无法保证重构后语音中的内容和说话人嵌入与原始语音保持一致，影响了语音转换性能。

针对目前非平行语料库下任意说话人之间语音转换训练过程中仅使用单个约束，不能很好地实现内容信息和说话人特征的解耦，影响语音转换性能的问题，本文提出了一种表示分离自编码器的语音转换（RSAE-VC, representation separation auto-encoder voice conversion）方法。该方法仅通过单个编码器即可分离内容和说话人嵌入，

在训练阶段，通过编码器对语音进行特征提取，并使用自适应实例归一化（AdaIN, adaptive instance normalization）将语音中包含的说话人个性特征去除，从而只保留语音中的内容信息。同时引入自内容损失（SC-Loss, self-content loss）[14]和自说话人损失（SS-Loss, self-speaker loss）[15]，进一步保证说话人个性特征在转换过程中的一致性，进而获得更好的转换性能。在转换阶段，首先将分别提取到的源语音的内容信息与目标语音中的说话人信息输入解码器，然后将两者合成并解码得到转换后的语音。

1 RSAE-VC 语音转换

RSAE-VC 只使用一个编码器进行内容和说话人的分离，其中 IN（instance normalization）层在图像风格转换中常作为一种风格归一化技术，而在语音中则将不同说话人的相同内容语句统一风格，可以视为去除了不同说话人的个性特征，只保留相同的内容信息[16]。因此可以从编码器生成的潜变量中提取出语音的内容信息，将语音的内容信息和说话人的个性特征分离，从而可以独立于语音的内容信息只改变语音的个性特征，同时添加 Sigmoid 激活引导层，更好地学习内容信息。分离后的内容信息再与通过 AdaIN 层重新添加的目标说话人信息输入解码器生成目标的 Mel 谱图。训练过程中加入 SC-Loss 和 SS-Loss，保证合成语音的内容和说话人信息与原始语音的内容和说话人信息在潜空间中映射一致，提高转换语音的质量。

1.1 模型训练

1.1.1 模型结构

RSAE-VC 语音转换系统使用编/解码器结构，并在编码器中增加 IN 层与 Sigmoid 激活引导层。通过文献[14]的实验结果对比发现，在使用 Sigmoid 激活函数作为激活引导时，分离的内容信息中说话人个性特征的识别准确率大幅降低，同时说话人个性特征信息中的说话人个性特征的识别准确率高达 93.2%，相较于其他的激活函数 ReLU、ELU 和 tanh 有了显著的提升，因此 Sigmoid 激活引导层在引导学习内容信息方面比其他激活函数更合适。在解码器中增加 AdaIN 层，将目标语音中的说话人个性特征与源语音中的内容信息合成 Mel 谱图，最终将合成的 Mel 谱图经过声码器合成目标语音。RSAE-VC 语音转换系统的训练原理框架如图 1 所示，转换原理框架如图 2 所示。

在图 1 训练阶段中，原始语音 Mel 谱图 $\boldsymbol{X}$ 经过编码器 E_{c} 解耦得到内容信息 $\boldsymbol{c}$ 和说话人个性特征信息 $\boldsymbol{s}$。同时，原始语音 Mel 谱图 $\boldsymbol{X}$ 经过相关编码器 E_{d} 后的输出 $\boldsymbol{z}'$ 减去内容信息 $\boldsymbol{c}$ 就得到相关说话人个性特征信息 $\boldsymbol{s}'$。内容信息 $\boldsymbol{c}$ 和说话人个性特征信息 $\boldsymbol{s}$ 再经过解码器 D 得到重构 Mel 谱图 $\boldsymbol{X}'$，再将重构 Mel 谱图 $\boldsymbol{X}'$ 分别经过编码器 E_{c} 和相关编码器 E_{d} 得到重构后的内容信息 $\hat{\boldsymbol{c}}$ 和相关说话人个性特征 $\hat{\boldsymbol{s}}'$。最后分别利用 $\boldsymbol{c}$ 与 $\hat{\boldsymbol{c}}$ 构造自内容损失 SC-Loss，$\boldsymbol{s}'$ 与 $\hat{\boldsymbol{s}}'$ 构造自说话人损失 SS-Loss，从而采用这 2 个损失函数训练编码器 E_{c}、解码器 D 和相关编码器 E_{d}。在图 2 转换阶段中，源语音 Mel 谱图 $\boldsymbol{X}$ 和目标语音 Mel 谱图 $\boldsymbol{Y}$ 分别经过编码器 E_{c} 得到源语音的内容信息 $\boldsymbol{c}$ 和目标语音的说话人个性特征信息 $\boldsymbol{s}$，两者经过解码器 D 后就可以得到转换语音 Mel 谱图 $\boldsymbol{Y}'$。使用声码器可将 Mel 谱图转换为语音信号，此时将得到的 Mel 谱图 $\boldsymbol{Y}'$ 输入声码器即可得到相对应的转换后的语音信号。

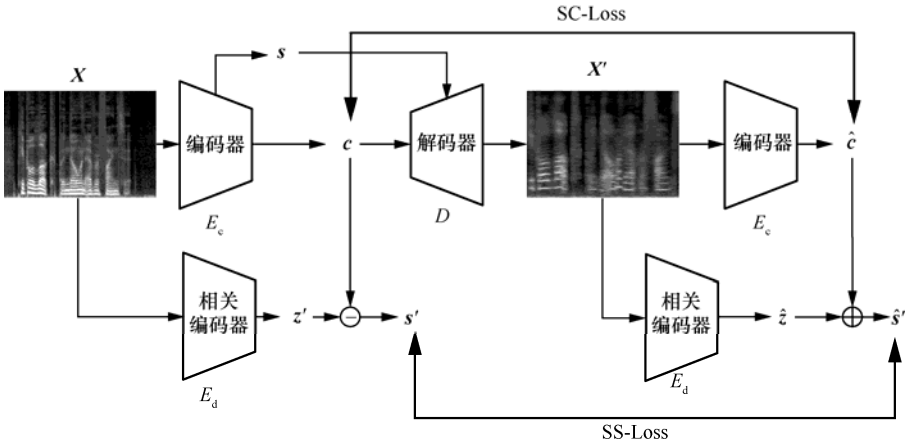


图 1 RSAE-VC 语音转换系统的训练原理框架

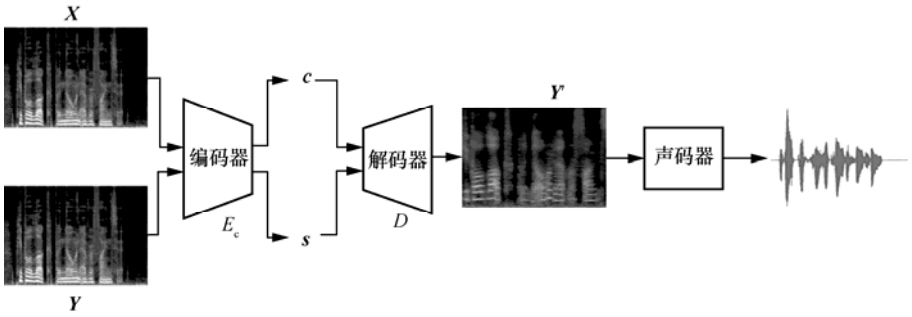


图 2　RSAE-VC 语音转换系统的转换原理框架

1.1.2　模型工作过程

RSAE-VC 使用编码器和解码器的结构，并且仅使用一个编码器提取语音的内容信息与说话人个性特征。编码器结构如图 3 所示。在编码器中，使用 Conv1d 层处理频率信息，在 Conv1d 层之后使用 6 个 ConvBlock，ConvBlock 由 ConvNorm 层、BatchNorm 层、LeakyReLU 层和 IN 层组成。内容信息通过 IN 层去除说话人个性特征得到，说话人个性特征经 IN 层得到均值矢量 $\boldsymbol{\mu}$ 和方差矩阵 $\boldsymbol{\sigma}$。

假设 $\boldsymbol{X}$ 为训练数据集中的任意一段语音的 Mel 谱图，E_c 表示编码器，D 表示解码器，则通过编码器卷积等计算生成的潜矢量序列 $\boldsymbol{z}$ 表示为

$$\boldsymbol{z} = E_c(\boldsymbol{X}) \tag{1}$$

得到 $\boldsymbol{z}$ 后，使用 IN 将潜矢量中包含的说话人个性特征归一化，得到语音的内容信息 $\boldsymbol{c}$

$$\boldsymbol{c} = \mathrm{IN}(\boldsymbol{z}) \tag{2}$$

同时，IN 方法可以表示为

$$\mathrm{IN}(\boldsymbol{z}) = \frac{\boldsymbol{z} - \boldsymbol{\mu}(\boldsymbol{z})}{\boldsymbol{\sigma}(\boldsymbol{z})} \tag{3}$$

其中，$\boldsymbol{\mu}(\boldsymbol{z})$ 和 $\boldsymbol{\sigma}(\boldsymbol{z})$ 表示通道的均值和方差。

由于在语音转换中将语音分为语音内容信息和语言风格信息两部分，而语言风格信息对应说话人个性特征，包含音高、响度等韵律信息，因此在语音转换中经过 IN 层将蕴含说话人个性特征的均值和方差去除后则留下语音的内容信息，不同说话人的相同内容语音经过归一化后，视为去除说话人的个性特征，只保留下相同的内容信息，以表征语音的内容信息，而均值和方差则表征语音的说话人个性特征。

经过 IN 层后，$p(\boldsymbol{c} \mid \boldsymbol{X})$ 则为具有零均值、单位方差的高斯分布[17]，即

$$p(\boldsymbol{c} \mid \boldsymbol{X}) = N(\boldsymbol{c}; 0, \boldsymbol{I}) \tag{4}$$

同时，潜矢量 $\boldsymbol{z}$ 经过 IN 层得到的均值矢量 $\boldsymbol{\mu}$ 和方差矩阵 $\boldsymbol{\sigma}$ 用作说话人嵌入并表示为 $\boldsymbol{s}$。

图 4 为相关编码器结构，使用和编码器一样的

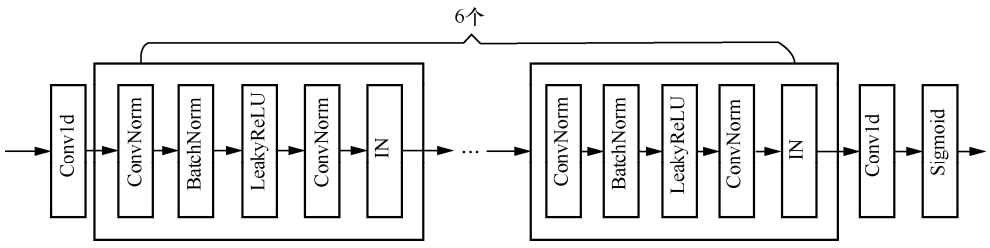


图 3　编码器结构

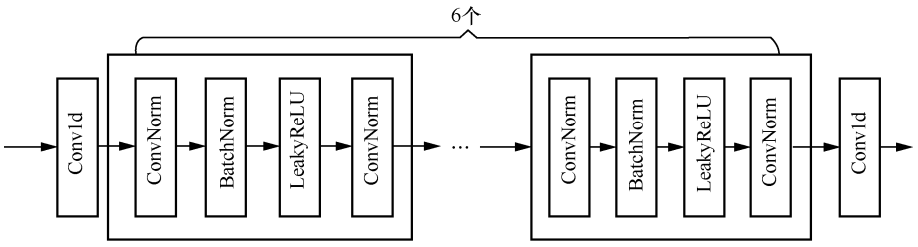


图 4　相关编码器结构

Conv1d 层，但不需要 IN 层去除说话人个性特征表示，也不需要 Sigmoid 层进行激活，Conv1d 层之后使用 6 个 ConvBlock。经过相关编码器得到相关语音特征表示，和经过编码器得到的内容表示映射在同一潜空间中。将语音内容信息 $\boldsymbol{c}$ 从相关语音特征表示潜矢量 $\boldsymbol{z}'$ 中减去，可得相关说话人的个性特征 $\boldsymbol{s}'$ [18]，表示为

$$\boldsymbol{s}' = \boldsymbol{z}' - \boldsymbol{c} \tag{5}$$

解码器结构如图 5 所示，在解码器网络中使用与编码器相同的 ConvBlock，并通过 AdaIN 层添加说话人个性特征，最后经过 GRU 层[19]和 Linear 层得到合成后的 Mel 谱图。

由于 $\boldsymbol{\mu}(\boldsymbol{z})$ 和 $\boldsymbol{\sigma}(\boldsymbol{z})$ 在转换过程中是时不变的，因此可以被视为说话人表示。在重构的解码阶段，分离的特征 $\boldsymbol{\mu}(\boldsymbol{z})$ 和 $\boldsymbol{\sigma}(\boldsymbol{z})$ 在 AdaIN 层中再次使用，其中 AdaIN 定义为

$$\mathrm{AdaIN}(\boldsymbol{H}, \boldsymbol{\mu}(\boldsymbol{z}), \boldsymbol{\sigma}(\boldsymbol{z})) = \boldsymbol{\sigma}(\boldsymbol{z})\,\mathrm{IN}(\boldsymbol{H}) + \boldsymbol{\mu}(\boldsymbol{z}) \tag{6}$$

其中，$\boldsymbol{H}$ 是源语音 Mel 谱图经过编码器去除说话人个性特征的内容信息表示，$\boldsymbol{z}$ 是目标语音 Mel 谱图经过编码器仅保留说话人个性特征的说话人表示。

在训练阶段，解码器对分离后的内容信息与说话人个性特征进行合成，得到重构后语音的 Mel 谱图 $\boldsymbol{X}'$

$$\boldsymbol{X}' = D(\boldsymbol{c}, \boldsymbol{s}) \tag{7}$$

RSAE-VC 采用 U-net 架构[20]，输入的 Mel 频谱图 $\boldsymbol{X}$ 通过多个 IN 层以去除说话人信息。其中的说话人嵌入 $\boldsymbol{\mu}(\boldsymbol{z})$ 和 $\boldsymbol{\sigma}(\boldsymbol{z})$ 跳过连接结构直接输入解码器中相应的多个 AdaIN 层中，以进行风格转换，最终生成 Mel 谱图 $\hat{\boldsymbol{X}}$ 并用于计算重构损失。

1.1.3 模型损失

重构后语音的 Mel 谱图和原始语音的 Mel 谱图之间构成重构损失，表示为

$$L_{\mathrm{rec}} = \left\| \boldsymbol{X}' - \boldsymbol{X} \right\|_1 \tag{8}$$

内容信息和说话人信息输入解码器得到重构语音 Mel 谱图，重构语音 Mel 谱图再次经过编码器后得到重构语音的内容信息表示 $\hat{\boldsymbol{c}}$，$\hat{\boldsymbol{c}}$ 与原始语音内容信息 $\boldsymbol{c}$ 之间构成 SC-Loss，表示为

$$L_{\mathrm{SC}} = \left\| \hat{\boldsymbol{c}} - \boldsymbol{c} \right\|_1 \tag{9}$$

同时，重构语音的 Mel 谱图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 $\hat{\boldsymbol{s}}'$，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 SS-Loss，表示为

$$L_{\mathrm{SS}} = \left\| \hat{\boldsymbol{s}}' - \boldsymbol{s}' \right\|_1 \tag{10}$$

使用结构相同的内容编码器从重构语音 Mel 谱图中获得重构后语音的内容信息，并与原始语音的内容特征形成 SC-Loss。在编码阶段添加相关编码器得到和内容信息在同一潜空间的语音相关特征信息。合成 Mel 谱图后也使用相同的相关编码器得到合成语音的相关说话人信息，使两者构成 SS-Loss，确保合成语音和原始语音之间的说话人信息保持不变，也确保说话人和内容之间是相互独立的。最终模型训练的总体失函数表示为

$$L_{\mathrm{total}} = L_{\mathrm{rec}} + \lambda_1 L_{\mathrm{SC}} + \lambda_2 L_{\mathrm{SS}} \tag{11}$$

其中，λ_1 是 SC-Loss 的权重系数，λ_2 是 SS-Loss 的权重系数。为了使总损失达到最小，通过反向传播计算损失函数梯度，不断优化参数进行更新找到最优解。

1.2 转换阶段

本节分别以 Mel 谱图的形式将待转换的源语音 $\boldsymbol{X}$ 和目标语音 $\boldsymbol{Y}$ 输入已经训练好的编码器中，编码器提取源语音的内容信息 $\boldsymbol{c}$ 和目标语音的个性特征 $\boldsymbol{s}$，再将源语音的内容信息 $\boldsymbol{c}$ 与目标语音的个性特征 $\boldsymbol{s}$ 通过解码器进行合成，则可以生成保留源语音的内容信息并具有目标说话人个性特征的语音 $\boldsymbol{Y}'$，实现语音转换。

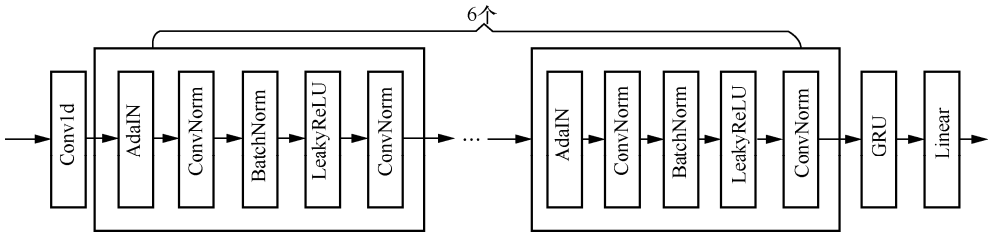


图 5 解码器结构

2　实验与结果

2.1　实验环境

实验使用的语音库为爱丁堡大学语音技术研究中心发布的 CSTR VCTK[21]，该数据集是以英语为母语并带有不同口音的英语说话人录制的语音数据集，其中包括 109 个说话人，每个说话人朗读大约 400 个语句，音频格式为 wav 文件。实验随机选取 80 个说话人的语音进行训练，并对每个说话人随机选择 200 个语音构成训练集。在测试方面，为了更好地评价任意说话人之间语音转换的效果以及模型的泛化能力，实验使用 VCC2018 数据集[22]进行性能测试，将本文提出的 RSAE-VC 方法与基线系统 AGAIN-VC 在非平行语料库上进行比较。对每个说话人，分别使用 50 个语句进行测试。

实验首先对语音进行预处理，语音数据采样率为 22.05 kHz，再对语音信号进行端点检测，去除静音部分。由于输出得到的是 Mel 谱图，同时考虑波形生成的质量和推理计算速度，实验使用 MelGAN[23]作为声码器。根据 MelGAN 的配置，对分帧后的语音信号进行 1 024 个点的短时傅里叶变换，生成 80 维 Mel 谱图，并以连续的 128 帧作为系统的输入特征参数。实验通过 ADAM 优化器进行训练，以初始学习率为 0.000 5 对 RSIN-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。

2.2　客观评价

2.2.1　性能指标

实验选用梅尔倒谱距离（MCD, Mel-cepstral distortion）[24]来衡量转换后的语音与目标语音的频谱距离，直观对比 RSAE-VC 方法与 AGAIN-VC 方法的转换性能，MCD 的计算式为

$$\mathrm{MCD}=\frac{10\sqrt{2}}{\ln 10}\times\frac{1}{M}\sum_{m=1}^{M}\sqrt{\sum_{r=1}^{R}(y_m(r)-y_m'(r))^2} \tag{12}$$

其中，$y_m(r)$ 和 $y_m'(r)$ 分别是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 是梅尔倒谱系数的维数，M 是总帧数。

为进一步对比 RSAE-VC 方法与 AGAIN-VC 方法的转换性能，实验同时采用基音频率 F_0 均方根误差（RMSE, root mean square error）[25]作为转换的客观指标之一。F_0 RMSE 的计算式为

$$F_0\ \mathrm{RMSE}=\sqrt{\frac{1}{M}\sum_{i=1}^{M}(F_{0i}-F_{0i}')^2} \tag{13}$$

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

2.2.2　参数设置

实验首先确定式(11)所示的模型总体损失函数中参数 λ_1 和 λ_2 的数值。在考虑语音仅包含内容信息和说话人个性特征，且内容信息和说话人个性特征相互独立的情况下，SC-Loss 和 SS-Loss 对 RSAE-VC 的性能影响相互独立。

当只考虑 SC-Loss 对 RSAE-VC 性能影响时，选用不同的 λ_1 系数对转换的性能进行对比。由于同一说话人的不同目标语音所提供的说话人信息无法做到完全一致，本文实验随机选用 VCC2018 数据集的 4 组平行语音数据进行测试，其中包括 2 组男声和 2 组女声，每组包括 50 条语音。在 4 种不同转换情形下进行实验，分别为女声转换到女声（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）。通过设置不同的 λ_1 值来调节 SC-Loss 的权重，对比合成语音的内容信息与原始语音的内容信息之间差异，并设置 $\lambda_2=0$。转换语音与目标语音在 4 种转换情形下 λ_1 的取值对 MCD 的影响如表 1 所示。

表 1　4 种转换情形下 λ_1 的取值对 MCD 的影响（$\lambda_2=0$）

λ_1	F2F	M2M	F2M	M2F
2.0	9.111	8.284	8.831	9.427
2.5	9.119	8.401	8.897	9.420
3.0	9.174	8.421	8.862	9.672
3.5	9.075	8.358	8.641	9.466
4.0	9.318	8.388	8.835	9.760

从表 1 中可以看出，当 λ_1=3.5 时，在 F2F 和 F2M 转换情形下其 MCD 值最小，且在其他转换情形下转换性能也较好，说明模型在只考虑 SC-Loss 时整体的性能达到最优。因此选取 $\lambda_1=3.5$。

在 λ_1=3.5 的前提下，考虑 SS-Loss 对转换性能的影响，选用不同的 λ_2 系数对 RSAE-VC 进行测试。表 2 是 4 种转换情形下 λ_2 的取值对 MCD 的影响。

表 2　4 种转换情形下 λ_2 的取值对 MCD 的影响($\lambda_1 = 3.5$)

λ_2	F2F	M2M	F2M	M2F
0.5	9.105	8.194	8.664	9.576
0.6	9.143	8.125	8.517	9.545
0.7	9.142	8.301	8.679	9.590
1.0	9.233	8.437	8.580	9.536
1.5	9.252	8.467	8.835	9.484

从表 2 中可以看出，当 λ_2=0.6 时，在 M2M 和 F2M 转换情形下其 MCD 值最小，且在其他转换情形下的转换性能也较好，说明在包含 SC-Loss 下考虑 SS-Loss 对模型的性能影响时，模型的性能达到最优。因此在之后实验的性能对比中，以 λ_1=3.5 和 λ_2=0.6 与基准模型进行对比。

2.2.3 消融实验

为了对比不同激活函数对语音转换性能的影响，实验选用 tanh、ReLU 和不同参量的 Sigmoid 作为实验对象，4 种情形下转换语音的 MCD 及重构损失 L_{rec} 的对比如表 3 所示。

表 3　不同激活函数在 4 种转换情形下转换语音的 MCD 及重构损失的对比

激活函数	MCD				L_{rec}
	F2F	M2M	F2M	M2F	
None	11.304	9.120	11.112	11.372	0.152
tanh	11.840	11.064	11.029	11.976	0.203
ReLU	12.163	11.917	11.919	12.237	0.174
ELU	11.874	11.827	11.714	12.010	0.150
Sigmoid1	9.730	8.792	9.216	9.576	0.151
Sigmoid2	9.105	8.194	8.664	9.576	0.147
Sigmoid3	9.736	8.792	9.639	10.189	0.149

表 3 中，Sigmoid1、Sigmoid2、Sigmoid3 分别表示参量 α 取值为 0.01、0.1 和 1。通过表 3 使用不同激活函数在 4 种情形下进行语音转换的对比可以发现，当不使用激活函数或使用其他激活函数（ReLU、ELU、tanh）进行激活引导时，转换得到的语音在 MCD 性能上都比使用 Sigmoid 激活函数要差，且当 Sigmoid 中 α 取值为 0.1 时，无论是 MCD 还是重构损失 L_{rec} 均为性能最佳，即使用 Sigmoid 作为激活引导层相较其他激活函数（ReLU、ELU 和 tanh）有了显著的提升。因此在进行任意说话人之间的语音转换过程中使用 Sigmoid 函数构成激活引导层。

为测试 RSAE-VC 在非平行语料库情况下任意说话人之间语音转换的性能，实验随机选用 VCC2018 数据集的 4 组非平行语音数据，包括其他的 2 组男性和 2 组女性，源语音每组包括 50 条语音，目标语音每组包括 10 条语音。源语音与转换语音均为训练集外数据，且目标语音和源语音在内容上不相同。实验将 RSAE-VC 与 AGAIN-VC、仅使用 SC-Loss 的 RSAE-VC（RSAE(SC-L)）、仅使用 SS-Loss 的 RSAE-VC（RSAE(SS-L)）以及借鉴 AdaIN-VC 的双编码器 RSAE-VC（RSAE(2Enc)）在非平行语料库下任意说话人之间的语音转换性能进行对比，4 种转换情形下转换语音的 MCD 和 F_0 RMSE 的对比分别如表 4 和表 5 所示。

表 4　4 种转换情形下转换语音的 MCD 对比

方法	F2F	M2M	F2M	M2F
AGAIN-VC	9.706	8.786	9.330	10.223
RSAE(SC-L)	10.317	9.019	9.426	10.569
RSAE(SS-L)	9.588	8.557	9.190	10.069
RSAE(2Enc)	10.250	9.174	10.163	10.826
RSAE-VC	9.486	8.477	9.011	9.894

表 5　4 种转换情形下转换语音的 F_0 RMSE 对比

方法	F2F/Hz	M2M/Hz	F2M/Hz	M2F/Hz
AGAIN-VC	97.371	69.956	72.332	96.497
RSAE(SC-L)	97.889	72.759	69.878	97.418
RSAE(SS-L)	99.337	71.319	68.936	97.810
RSAE(2Enc)	98.101	72.779	69.377	96.677
RSAE-VC	94.483	67.662	69.358	96.579

从表 4 可以看出，4 种情形下 RSAE-VC 的 MCD 值与 AGAIN-VC 相比均有所下降，且在 M2M 和 F2M 转换时下降幅度最大，说明

RSAE-VC 在低频语音转换有更明显的效果，在高频处语音转换也较 AGAIN-VC 有所提升。当仅使用自内容损失在任意说话人之间语音转换效果比 AGAIN-VC 效果稍差，由于在训练过程中只将训练集中语音的内容信息损失降到最低，但在训练集外的语音进行转换时无法保证内容信息没有丢失，因此在仅使用自内容损失的情况下转换效果没有显著提升。而仅使用自说话人损失时，由于可以学习到内容信息和说话人个性特征分离的情况，在转换过程中可以更好地分离，转换后 4 种情形下 MCD 值均有所下降。

从表 5 可以看出，RSAE-VC 与 AGAIN-VC 相比，除了 M2F 的 F_0 RMSE 有所上升外，其余 3 组 F_0 RMSE 均有所下降，且在 M2M 和 F2M 时下降幅度较大，而 M2F 时的 F_0 RMSE 对比 AGAIN-VC 也在合理范围之内。当单独使用自内容损失或自说话人损失不能保证有些韵律信息体现在说话人个性特征中，因此在仅使用单个损失函数的情况下转换语音的 F_0 RMSE 性能无明显提升。

2.2.4　性能评价

为了从更多方面对 RSAE-VC 和 AGAIN-VC 的性能进行客观分析，实验采用梅尔频率倒谱系数（MFCC, Mel frequency cepstrum coefficient）[26]作为另一项评价指标。实验分别用 RSAE-VC 模型与 AGAIN-VC 模型生成转换语音，并提取 MFCC 特征参数。以转换语音的 MFCC 为横坐标、对应的目标语音 MFCC 参数为纵坐标进行绘图，结果如图 6 所示，其中，△、×、✡、+分别表示随机选取的 4 帧转换语音与相应的目标语音。

当 MFCC 系数分布趋向于 45°线（用虚线表示）时，表明转换后语音与目标语音有较高的匹配度。从图 6 可以看出，RSAE-VC 模型的转换语音与目标语音的 MFCC 参数更聚拢于 45° 线，而 AGAIN-VC 模型的分布相对分散，这一结果表明本文提出的 RSAE-VC 模型的转换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基准模型 AGAIN-VC。

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，表 6 为不同方法的时间复杂度对比。从表 6 可以发现，当仅使用自内容损失时，由于重构语音需再次经过编码器得到重构后的内容信息与输入语音进行自内容损失，因此 RSAE(SC-L)的时间复杂度比 AGAIN-VC 提升 23%。当仅使用自说话人损失时，由于需要经过相关编码器得到相关说话人个性特征，因此在时间复杂度上比仅使用自内容损失提升 16%，相较于 AGAIN-VC 提升 44%。当同时使用自内容损失和自说话人损失时，时间复杂度与 AGAIN-VC 相比提升 45%。当使用内容编码器和说话人编码器 2 个编码器进行语音转换时，相较于仅使用一个编码器进行分离的情况提升 18%，并且使用 2 个编码器在任意说话人之间语音转换效果比使用一个编码器效果下降 9.6%，因此使用一个编码器情况下在复杂度和转换效果上都比使用 2 个编码器性能好。

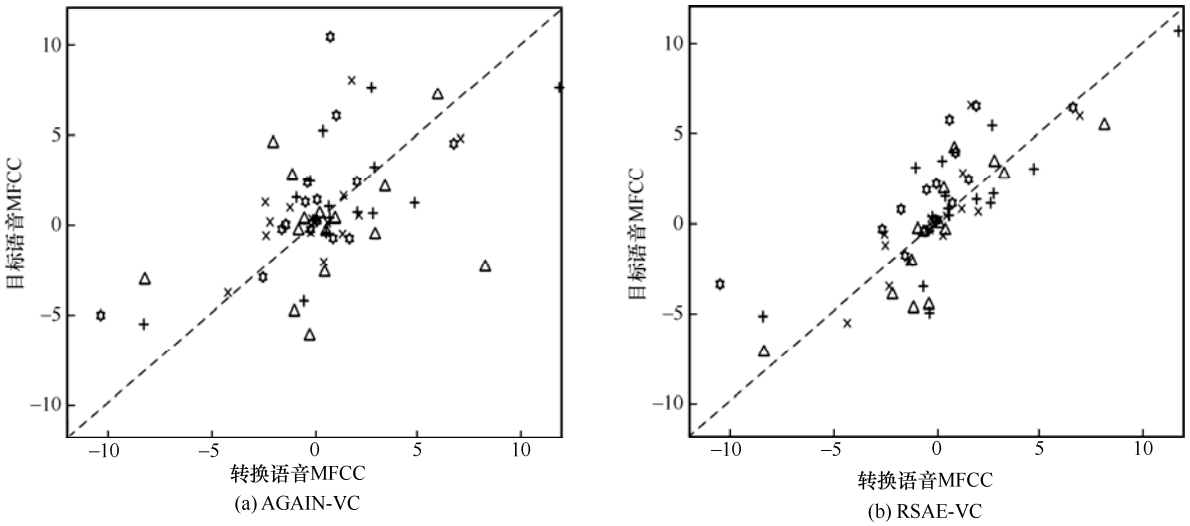


图 6　转换语音与目标语音的 MFCC 参数匹配

表 6 不同方法的时间复杂度对比

方法	时间/s
AGAIN-VC	0.177
RSAE(SC-L)	0.218
RSAE(SS-L)	0.255
RSAE(2Enc)	0.304
RSAE-VC	0.256

综合 MCD 值和 F_0 RMSE 的对比可以发现，本文所提 RSAE-VC 方法比 AGAIN-VC 方法均有不同程度的降低，同时 MFCC 参数匹配图也表明本文方法的匹配度优于基准模型 AGAIN-VC。

2.3 主观评价

对于主观评价，实验分别选用反映语音质量的平均意见分（MOS, mean opinion score）[27]和反映个性相似度的 ABX 值对目标语音和转换后语音进行测试。

MOS 意见分将语音质量分为 5 个等级，分别为 1（很差）、2（差）、3（一般）、4（好）、5（很好），然后让测试者在实验环境下对转换语音进行打分，从而判断转换后语音的质量。本文在 F2F、F2M、M2M、M2F 这 4 种情形下进行 MOS 评测，每种情况随机挑选 30 条目标语音和相应的 30 条转换后语音，20 名受试者对这些语音进行打分，MOS 值越高，代表语音质量越好，结果如图 7 所示。

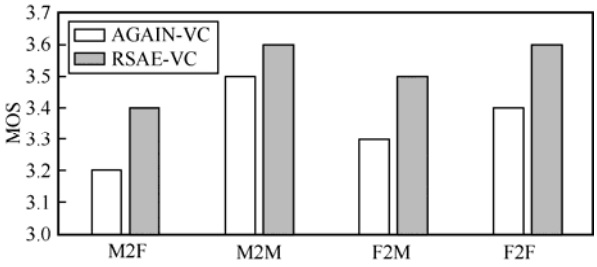


图 7 4 种情形下转换语音的 MOS 值比较

从图 7 中 2 种算法的 MOS 值对比来看，在 4 种不同情形的语音转换下，RSAE-VC 的转换语音相较于 AGAIN-VC 的转换语音在流畅度、自然度等方面都要好。由于同性语音之间的特征参数具有更小的差异，AGAIN-VC 与 RSAE-VC 在进行同性别之间语音转换时，语音质量都要优于跨性别之间语音转换。

另外，实验选用 ABX 评测[28]方法对转换语音与目标语音的相似度进行测试。测试者对听到的语音做出判断，若接近目标语音则积 1 分，若接近源语音则积 0 分，最后将积分除以总的测试语音数目计算出 ABX 测试分值。本节分别对 AGAIN-VC 与 RSAE-VC 在 4 种不同情况下转换的语音做了 ABX 测试，随机挑选每种情况下 30 条目标语音和相应的 30 条转换后语音，让 20 名受试者打分，ABX 分值越高，代表语音相似度越高，结果如图 8 所示。

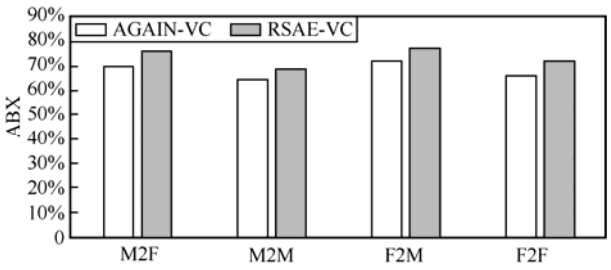


图 8 4 种情形下 ABX 评分比较

从图 8 的 ABX 相似度测试结果可以看出，RSAE-VC 转换语音与目标语音的相似度比 AGAIN-VC 好。由于跨性别之间的语音的特征参数差异更大，转换的程度更高，因此在跨性别转换时更加明显。而在同性之间的语音转换相似度上，RSAE-VC 也有较好的性能，说明 RSAE-VC 在低频到低频、高频到高频的语音转换上效果明显，即同频转换时表现更加优秀。

3 结束语

本文提出了 RSAE-VC 方法，更好地实现了语音中内容信息和说话人特征的解耦，有效地提升了非对称语音条件下任意说话人之间的语音转换性能。该方法不需要收集平行语料库，避免了平行语音收集的复杂度以及每进行一对说话人语音转换就需要重新训练模型的训练成本，同时也以简单高效的方式将任意源语音转换为任意目标语音，且提高了转换语音的质量。主观和客观实验结果都表明，RSAE-VC 方法显著优于现有的 AGAIN-VC 方法，有效地提升了转换语音的质量和说话人个性相似度。

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] MOUCHTARIS A, AGIOMYRGIANNAKIS Y, STYLIANOU Y. Conditional vector quantization for voice conversion[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2007: 505-508.

[3] AIHARA R, TAKASHIMA R, TAKIGUCHI T, et al. GMM-based emotional voice conversion using spectrum and prosody features[J]. American Journal of Signal Processing, 2012, 2(5): 134-138.

[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 posterior grams for many-to-one voice conversion without parallel data training[C]//Proceedings of IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE Press, 2016: 1-6.

[7] MURAKAMI H, HARA S, ABE M. DNN-based voice conversion with auxiliary phonemic information to improve intelligibility of glossectomy patients' speech[C]//Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Piscataway: IEEE Press, 2019: 138-142.

[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]//Proceedings of 5th International Conference on Computing and Informatics (ICCI). Piscataway: IEEE Press, 2022: 221-226.

[9] KANEKO T, KAMEOKA H, TANAKA K, et al. CycleGAN-VC2: improved cyclegan-based non-parallel voice conversion[C]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 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]//Proceedings of IEEE Spoken Language Technology Workshop (SLT). Piscataway: IEEE Press, 2018: 266-273.

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

[12] DENG C H, CHEN Y, DENG H F. One-shot voice conversion algorithm based on representations separation[J]. IEEE Access, 2020, 8: 196578-196586.

[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]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 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]// Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 3184-3188.

[15] DANG T, TRAN D, CHIN P, et al. Training robust zero-shot voice conversion models with self-supervised features[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6557-6561.

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

[17] WANG X, TAKAKI S, YAMAGISHI J, et al. A vector quantized variational autoencoder (VQ-VAE) autoregressive neural F_0 model for statistical parametric speech synthesis[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 28: 157-170.

[18] WU D Y, LEE H Y. One-shot voice conversion by vector quantization[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2020: 7734-7738.

[19] YANG S D, YU X Y, ZHOU Y. LSTM and GRU neural network performance comparison study: taking yelp review dataset as an example[C]//Proceedings of International Workshop on Electronic Communication and Artificial Intelligence (IWECAI). Piscataway: IEEE Press, 2020: 98-101.

[20] PRASAD S, MANU A, KAPOOR A, et al. Non-parallel denoised voice conversion using vector quantisation[C]//Proceedings of 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST). Piscataway: IEEE Press, 2022: 78-83.

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

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

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

[24] 周健, 刘荣敏, 窦云峰, 等. 采用 $L_{1/2}$ 稀疏约束的梅尔倒谱系数语音重建方法[J]. 声学学报, 2018, 43(6): 991-999.
ZHOU J, LIU R M, DOU Y F, et al. Speech reconstruction from Mel-frequency cepstral coefficients via $L_{1/2}$ sparse constraint[J]. Acta Acustica, 2018, 43(6): 991-999.

[25] 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.

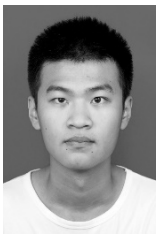
[26] 林云, 徐怀韬, 王森, 等. 基于特征融合的通信语音干扰效果客观评估[J]. 通信学报, 2023, 44(3): 105-116.
LIN Y, XU H T, WANG S, et al. Objective assessment of communication speech interference effect based on feature fusion[J]. Journal on Communications, 2023, 44(3): 105-116.

[27] PRIHASTO B, LIN Y X, LE P T, et al. CNEG-VC: contrastive learning using hard negative example in non-parallel voice conversion[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.

[28] SHAH N, SINGH M, TAKAHASHI N, et al. Nonparallel emotional voice conversion for unseen speaker-emotion pairs using dual domain adversarial network & virtual domain pairing[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.

[作者简介]

简志华（1978− ），男，江西新余人，博士，杭州电子科技大学副教授，主要研究方向为智能语音处理、语音转换、伪造语音检测、语音隐私保护等。



章子旭（1999− ），男，浙江杭州人，杭州电子科技大学硕士生，主要研究方向为语音转换。