第 48 卷第 5 期
2023 年 9 月

声 学 学 报
ACTA ACUSTICA

Vol. 48, No. 5
Sept., 2023

融合梅尔谱增强与特征解耦的噪声鲁棒语音转换*

陈乐乐 张雄伟† 孙 蒙 张星昱

(陆军工程大学 指挥控制工程学院 南京 210007)

2022 年 9 月 27 日收到
2022 年 12 月 15 日定稿

摘要 提出了一种融合梅尔谱增强与特征解耦的噪声鲁棒语音转换模型, 即 MENR-VC 模型。该模型采用 3 个编码器提取语音内容、基频和说话人身份矢量特征, 并引入互信息作为相关性度量指标, 通过最小化互信息进行矢量特征解耦, 实现对说话人身份的转换。为了改善含噪语音的频谱质量, 模型使用深度复数循环卷积网络对含噪梅尔谱进行增强, 并将其作为说话人编码器的输入; 同时, 在训练过程中, 引入梅尔谱增强损失函数对模型整体损失函数进行了改进。仿真实验结果表明, 与同类最优的噪声鲁棒语音转换方法相比, 所提模型得到的转换语音在语音自然度和说话人相似度的平均意见得分方面，分别提高了 0.12 和 0.07。解决了语音转换模型在使用含噪语音进行训练时, 会导致深度神经网络训练过程难以收敛, 转换语音质量大幅下降的问题。

关键词 语音转换, 噪声鲁棒, 梅尔谱增强, 特征解耦

PACS 数 43.50, 43.60 **DOI：**10.12395/0371-0025.2022093

Noise robust voice conversion with the fusion of Mel-spectrum enhancement and feature disentanglement

CHEN Lele ZHANG Xiongwei† SUN Meng ZHANG Xingyu

(*College of Command and Control Engineering, Army Engineering University of PLA* Nanjing 210007)

Received Sept. 27, 2022
Revised Dec. 15, 2022

Abstract A novel noise-robust voice conversion model named MENR-VC which combines Mel-spectrum enhancement and feature decoupling is proposed in this paper. Speech content, fundamental frequency, and speaker identity vector features are extracted by three encoders in the model, and mutual information is introduced as a correlation metric to achieve speaker identity conversion via minimizing mutual information for feature decoupling. To overcome the limitations of noisy speech, a deep complex recurrent convolutional network is employed by the model to enhance the noisy Mel spectrum, which serves as input to the speaker encoder. Additionally, the Mel-spectrum enhancement loss function is introduced during the training process to improve the overall loss function of the model. The simulation results demonstrate that similar optimal noise-robust voice conversion methods are outperformed by the proposed model, with an enhancement of 0.12 and 0.07 in the average opinion scores of speech naturalness and speaker similarity of the converted speech respectively. The training of deep neural network can be easily converged when noisy speech is used as training data in the proposed voice conversion model, and the quality of the converted speech is also satisfatory.

Keywords Voice conversion, Noise robustness, Mel-spectrum enhancement, Feature decoupling

引言

语音转换 (VC) 是一种在不改变语音内容的前提下, 将源说话人的声音改变为目标说话人声音的技术[1-2]。

传统的语音转换方法包括高斯混合模型[3]、频率曲折[4-5] 以及基于模板[6] 的方法等, 这种基于参数统计的方法, 其最终目标是得到从源语音到目标语音的映射函数, 主要采用平行语料数据集进行训

* 国家自然科学基金项目 (62071484) 资助
† 通讯作者: 张雄伟, xwzhang9898@163.com

练。随着深度学习的蓬勃发展, 生成对抗网络[7-8]、编码器−解码器[9]、特征解耦[10-11] 等方法被相继提出, 以上基于数据驱动的方法不仅突破了平行语料数据集的约束, 还通过解耦语音内容和说话人身份, 灵活实现任意说话人之间的语音转换[12-13]。上述方法多采用在实验室或安静的声学场景下录制的干净语音数据集进行训练和测试, 但在实际的应用场景中, 背景噪声是个不可避免的环境因素, 无论是源说话人还是目标说话人, 语音中若掺杂着噪声因素, 转换后的语音质量和可懂度必将受到不可估计的影响。

针对噪声场景下的语音转换, 有学者通过引入域对抗训练来学习噪声不变的潜在特征[14-15], 试图解开声学特征中的噪声条件。文献 [16] 提出了一种基于矢量量化和对比预测编码的方法—NoiseVC, 通过噪声增强提高语音内容和说话人身份的解耦精度。2021 年亚马逊公司受去噪自动编码器框架的启发, 提出了针对嘈杂语音背景下的语音转换框架—Voicy[17], 该框架能够执行非并行零样本语音转换。Xue 等基于最新的 Glow-WaveGAN 模型[18], 提出了一种噪声可控的 WaveGAN[19], 通过编码器直接从波形中学习无噪声的声学表示, 该模型在目标说话人噪声语音转换中获得了不错的效果。但无论是上述的 Voicy 方法还是噪声可控的 WaveGAN 模型, 其转换效果还是受限于自动语音识别 (ASR) 模型的性能, 且均未考虑语音特征矢量解耦的充分性。

因此, 在前人研究的基础上, 结合 VQMIVC 模型[13]提出了一种融合梅尔谱增强及特征解耦的噪声鲁棒语音转换模型 (Mel-spectrum Enhanced Noise Robust Voice Conversion, MENR-VC)。MENR-VC 模型包括 3 个编码器: 语音内容编码器、音高编码器以及添加了梅尔谱增强模块的说话人编码器。使用 3 个编码器分别提取语音内容、基频和说话人身份 3 个特征矢量, 并在训练时引入互信息来度量 3 个语音特征矢量解耦的充分性,梅尔谱增强模块采用改进后的 DCCRN[20] 网络模型。首先, 输入含噪语音的梅尔谱, 使用复数卷积循环网络对幅度和相位部分进行训练, 将训练所得的增强梅尔谱直接作为说话人编码器的输入。同时, 在训练过程中引入梅尔谱增强损失函数对模型整体损失函数进行了改进。最后, 分别从 3 个编码器中提取对应特征矢量输入到解码器中重构语音梅尔谱, 再通过神经声码器 Parallel WaveGAN[21] 得到转换后的语音波形。实验结果表明, 在目标说话人语音含噪的情况下, 使用 MENR-VC 模型得到的转换语音在自然度和说话人相似度方面均有显著提高。引入梅尔谱增强模块对带噪语音进行增强, 改善了带噪语音梅尔谱的频谱质量, 缓解了现有语音转换模型面对目标说话人语音含噪的情况下, 转换语音质量下降的问题; 引入互信息作为相关性度量, 通过最小化互信息降低语音内容、音高和说话人身份特征矢量之间的相关性, 提高特征解耦的充分性; 将梅尔谱增强模块和特征解耦模块进行联合训练, 避免了因分开训练造成的语音失真问题, 提高模型整体的噪声鲁棒性。

1　语音转换中的特征解耦

语音信息大致可分为语音内容、音高、韵律和音色。目前, 基于音色信息和语义信息解耦的语音转换方法是学者们的重点研究方向。图 1 给出了基于音色信息和语义信息解耦的语音转换框架。

语义信息是动态的, 各帧之间的语义表示差异巨大, 因此内容编码器的作用主要是以语义的形式对内容进行编码。内容编码器通过压缩编码得到内容表示, 而在压缩过程中部分不必要的信息如说话人身份等会被丢失。

说话人身份是与时间无关的静态信息, 为了从频谱中分离出说话人表示, 说话人编码器首先使用

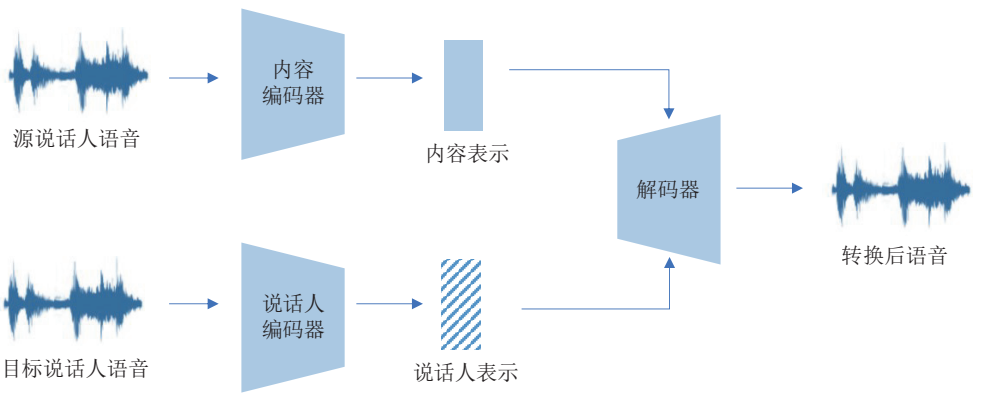

图 1　基于音色信息和语义信息解耦的语音转换框架

一维卷积层来获得帧级说话人特征, 然后采用平均池化层对帧级说话人信息进行聚合, 形成话语级说话人表示。

内容编码器以源说话人语音频谱作为输入, 提取出帧级的语音内容表示; 说话人编码器以目标说话人的语音频谱作为输入, 提取话语级的说话人表示; 最后, 解码器将内容和说话人表示作为输入, 重构转换后的语音频谱。

采取语音内容和说话人身份解耦的方式进行语音转换, 可以实现独立控制生成语音的说话人身份。除了语音内容和说话人身份, 音高信息同样对转换语音质量产生影响。现有研究通过添加多个编码器分别对语音信息进行矢量编码, 并通过瓶颈调优, 来约束自动编码器输入端语音成分的特征解耦。然而, 瓶颈调优是不健壮的, 各个语音特征分量之间解耦的充分性难以衡量。因此提出了 MENR-VC 模型, 引入相关性度量指标互信息, 在完成含噪语音转换的同时关注了语音信息各个特征矢量之间解耦充分性的问题。

2 MENR-VC 模型

图 2 给出了 MENR-VC 模型框架图, 该模型由梅尔谱增强模块和特征解耦模块构成。特征解耦模块中包含 3 个编码器: 说话人编码器、内容编码器和音高编码器, 通过最小化互信息降低 3 个编码器提取的特征矢量之间的依赖关系, 实现特征解耦并提高语音转换模型的性能。梅尔谱增强模块添加在说话人编码器之前, 可以获得噪声鲁棒的说话人矢量, 与特征解耦模块共同训练可提高模型的噪声鲁棒性。

2.1 梅尔谱增强模块

DCCRN 模型的网络结构如图 3(a) 所示。改进后的梅尔谱增强网络如图 3(b) 所示, 由于人类对于频率的感知在梅尔尺度上更加敏感, 因此将输入变为含噪语音的梅尔谱, 以及含噪语音经过短时傅里叶变换后得到的相位谱, 采用 stack 函数将梅尔谱和相位谱相结合输入到 DCCRN 网络中。经过 DCCRN 网络的训练后并未进行逆短时傅里叶变化, 而是经过一维卷积层处理后直接输出增强的梅尔谱。

图 3 的复数编码器模块包括复数二维卷积, 复数批归一化和实数 PReLU, 其结构如图 4 所示。输入的梅尔谱特征按照通道维度划分, 前 256 维作为实部, 后 256 维作为虚部, 分别输入到实部卷积和虚部卷积, 将卷积层的输出结果按复数运算步骤进行处理, 最后得出特征输出。

复数二维卷积模块中, 包含 4 个传统的二维卷积, 其作用是控制编码器中的复数信息流。其中, 复数卷积滤波器为 $\boldsymbol{U}=\boldsymbol{U}_r+\mathrm{j}\boldsymbol{U}_i$, 同时, 自定义复数矩阵为 $\boldsymbol{V}=\boldsymbol{V}_r+\mathrm{j}\boldsymbol{V}_i$, 实数矩阵 $\boldsymbol{U}_r$ 表示复数卷积核的实部, $\boldsymbol{U}_i$ 表示复数卷积核的虚部, 由此得到的复数卷积操作:

$$\boldsymbol{V}*\boldsymbol{U}=(\boldsymbol{V}_r*\boldsymbol{U}_r-\boldsymbol{V}_i*\boldsymbol{U}_i)+\mathrm{j}(\boldsymbol{V}_r*\boldsymbol{U}_i+\boldsymbol{V}_i*\boldsymbol{U}_r). \quad (1)$$

复数相较于实数更加适合表示频谱信息, 因此与实数卷积相比, 复数卷积学习数据特征的能力更强, 经过增强的梅尔谱也学习到了更多幅度谱和相

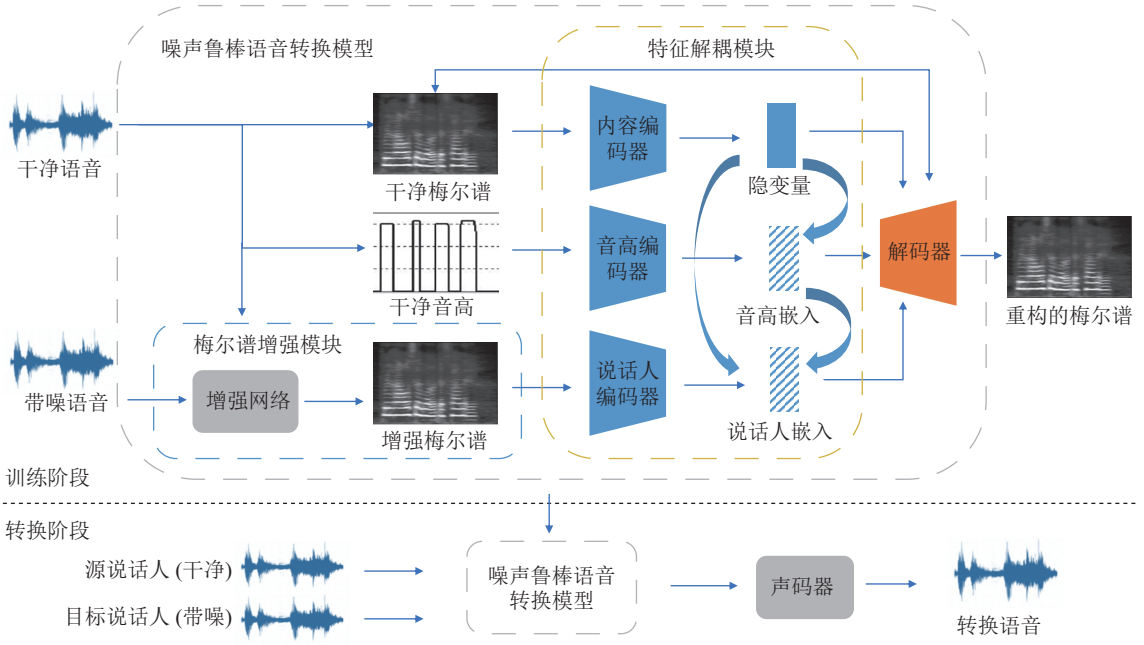

图 2 MENR-VC 模型框架图

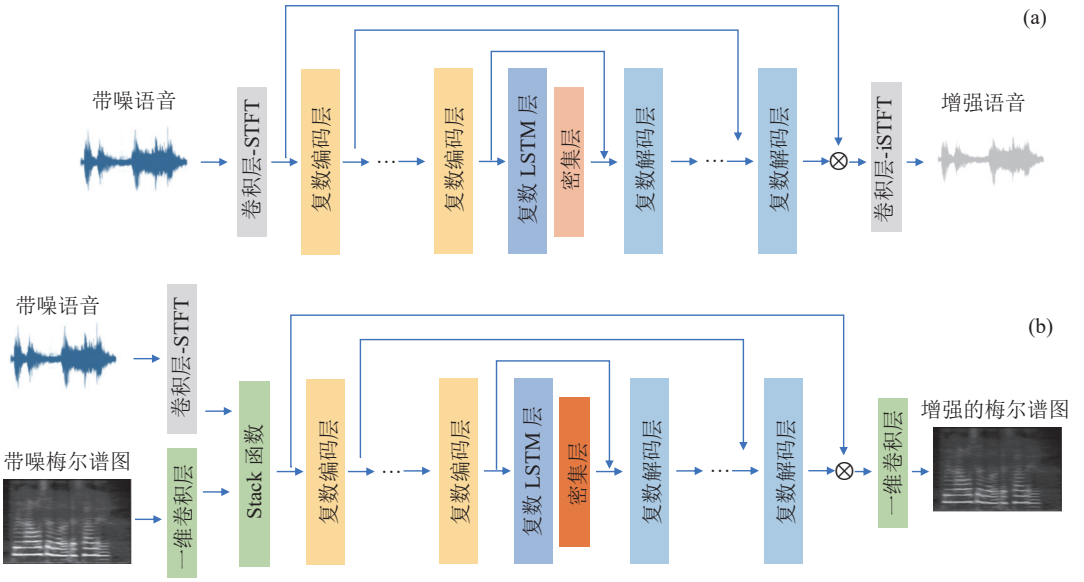

图 3　DCCRN 网络框架图

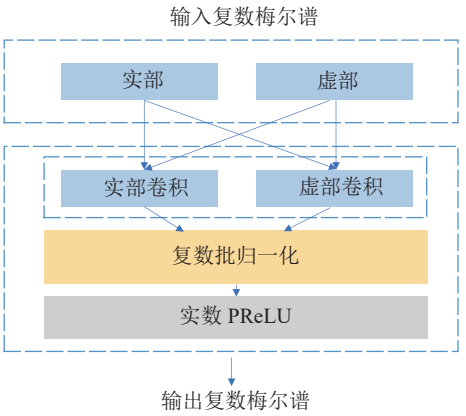

图 4　复数编码器结构

位谱之间的关联关系。

2.2　特征解耦模块

特征解耦模块包含 3 个编码器: 内容编码器、音高编码器和加入语音增强模块的说话人编码器, 使用 3 个编码器分别对语音的内容、音高和说话人身份进行特征编码, 同时引入了互信息, 通过最小化互信息降低各个特征矢量之间的依赖关系, 实现特征解耦, 提高语音转换性能。

内容编码器 E_c: 内容编码器采用矢量量化和对比预测编码 (VQ-CPC) 对输入语音提取内容信息。VQ-CPC 结构如图 5 所示。首先, 将输入语音参数化成为梅尔谱并输入到图 5 灰色虚线框表示的编码器中, 该编码器由一个跨步卷积层 (将输入下采样 2 倍) 和带有 ReLU 激活函数的 4 个线性层的堆栈组成, 并在每一层后应用层归一化。编码器的输出被投影到一系列连续的潜在向量 z 中, 接着使用 VQ 信息瓶颈将 z 向量离散化为 $\hat{z}$, VQ 通过信息瓶颈去除 $\hat{z}$ 中不重要的细节信息, 因此能够捕获更多底层语言信息。离散化的向量 $\hat{z}$ 被一个回归模型总结为上下文向量 $\boldsymbol{R}_t$, 接着, 使用 CPC[22] 预测未来语音样本, 可将跨多个时间步长的局部特征编码到内容表示中, 提高重建语音的准确度。

说话人编码器 E_s: 图 6 展示了说话人编码器的结构。该结构中包括一个 ConvBank 层, 两个卷积块、一个 Dense 块和平均池化层。ConvBank 层源自于 2017 年提出的端到端语音合成模型 Tactron 中的 CBHG 模块[23], CBHG 模块由卷积层组 (CB)、Highway 网络 (H) 以及双向门控递归单元 (G) 组成。其中的 ConvBank 层能够从序列中提取表示, 具体而言, 输入序列首先与 N 个一维卷积滤波器集卷积, 其中第 n 个集合包含宽度为 n 的滤波器 C_n ($N=1,2,3,\cdots,n$)。这些滤波器显式地建模局部和上下文信息 (类似于建模 ungram、biggram, 直到 N-gram)。使用 ConvBank 层能够扩大感受野, 更好地捕获长时信息, 通过应用平均池化层能够强制说话人编码器只学习全局信息。说话人编码器接收梅尔谱增强网络输出的增强语谱作为输入, 生成噪声鲁棒的说话人矢量表示, 通过对全局语音特征的捕获来控制说话人的身份信息。

音高编码器 E_{f0}: $f0$ 中不仅包含音高信息, 同时还有部分节奏信息和说话人身份信息, 所以无法直接使用 $f0$ 编码音高。与内容编码器和说话人编码器不同, 音高编码器直接从波形中提取 $f0$, 并且对每个语音进行对数归一化处理, 提取与说话人无关的音高矢量。

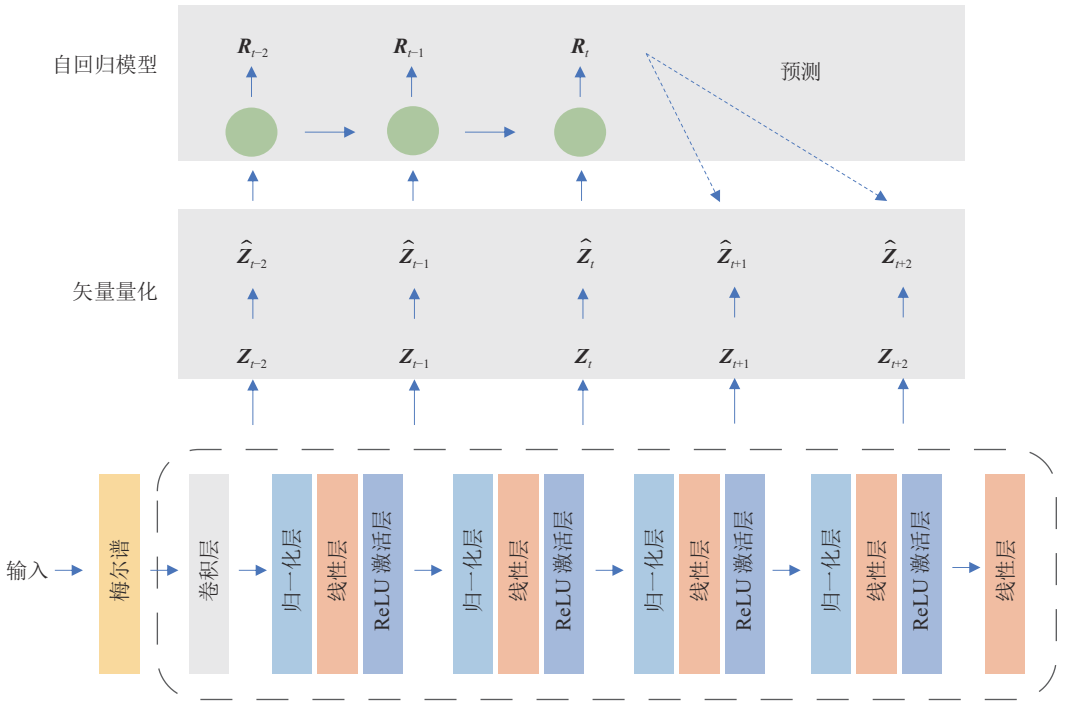

图 5 VQ-CPC 结构

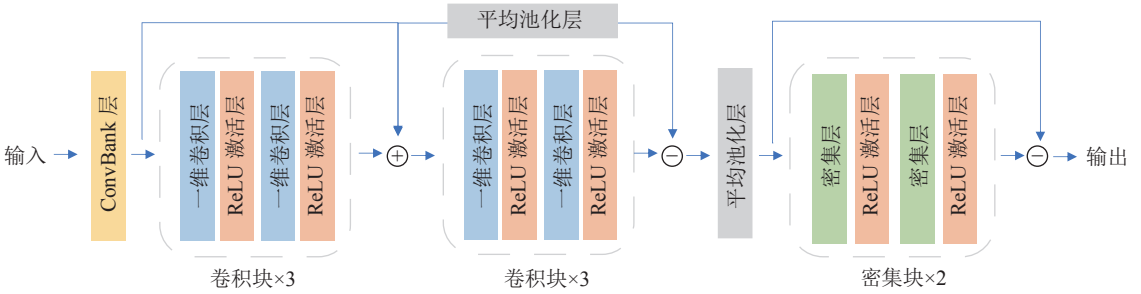

图 6 说话人编码器结构

互信息: 互信息 (MI) 是衡量两个随机变量之间依赖程度的一个量。对于两个连续随机变量 x 和 y, MI 定义如下:

$$I(x;y)=\int p(x,y)\log_{10}\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y=\mathrm{E}_{p(x,y)}\left[\log_{10}\frac{p(x,y)}{p(x)p(y)}\right], \tag{2}$$

其中, $p(x,y)$为联合分布, $p(x)$和 $p(y)$为边际分布。文中选用对比对数比上界 (CLUB)[24] 以对比学习的方式, 利用正、负样本对之间的条件概率之差, 提出了 MI 上界估计量。由于 MENR-VC 模型的目标是对语音内容、说话人身份和音高等特征矢量进行特征解耦, 因此利用 CLUB 来减少以上 3 个特征矢量两两之间的相互依赖。互信息的具体损失函数在 2.3 节中介绍。

解码器 *D*: 解码器将以上 3 个编码器的输出表示映射为转换语音的梅尔谱。具体地, 分别对内容和说话人矢量进行上采样和重复线性插值操作, 再结合音高矢量, 最终生成转换语音的梅尔谱。

2.3 改进的损失函数

MENR-VC 框架的整体损失函数包括增强模块损失、特征解耦模块损失 (内容编码器损失和互信息损失) 以及重构模块损失:

$$L_{\mathrm{VC}}=\beta L_{\mathrm{Ehan}}+L_{\mathrm{VQ}}+L_{\mathrm{CPC}}+\gamma L_{\mathrm{MI}}+\lambda L_{\mathrm{REC}}, \tag{3}$$

其中, $\beta\geqslant 0$, $\gamma\geqslant 0$, $\lambda\geqslant 0$, 用来控制客观项的权重。实验中, 这些超参数分别设置为 $\beta=10$, $\gamma=1\times10^{-2}$, $\lambda=10$。其中, γ 的取值与文献 [13] 实验所得的最优值保持一致。

L_{Ehan} 表示增强梅尔谱模块的损失, 将其定义为增强后的梅尔谱与干净语音梅尔谱的均方误差 (MSE):

$$L_{\mathrm{Ehan}}=\frac{1}{T}\sum_{i=1}^{T}(M_{i_c}-N_E(M_{i_n}))^2, \tag{4}$$

其中, M_{i_c}表示干净语音梅尔谱, M_{i_n}表示含噪语音梅尔谱, N_E表示梅尔谱增强网络。

特征解耦模块的损失由矢量量化损失L_{VQ}、对比预测编码损失L_{CPC}和互信息损失L_{MI}三部分组成。通过最小化L_{VQ}和L_{CPC}优化内容编码器, VQ 损失函数:

$$L_{\mathrm{VQ}}=\frac{1}{T}\sum_{i=1}^{T}\left\|\mathrm{sg}(\boldsymbol{z}_i)-\widehat{\boldsymbol{z}}_i\right\|^2+\alpha\left\|\boldsymbol{z}_i-\mathrm{sg}(\widehat{\boldsymbol{z}}_i)\right\|^2, \tag{5}$$

其中, $\boldsymbol{z}_i$表示连续特征向量, $\widehat{\boldsymbol{z}}_i$表示使用最近邻搜索得到的离散向量, $\mathrm{sg}(\cdot)$表示梯度停止算子, 在反向传播时不计算此项梯度。另外, 为了保证重构效果, 引入系数α ($\alpha<1$)控制两部分比例。

采用对比预测编码, 能够鼓励离散化的$\widehat{\boldsymbol{z}}_i$向量捕获更多局部结构。给定$m$步的预测范围、可训练的预测矩阵$\boldsymbol{W}_m$ ($m=1,2,\cdots,M$)、正样本$\widehat{\boldsymbol{z}}_{t+m}$和负样本集合$\boldsymbol{N}_{t,m}$, 通过训练最小化 InfoNCE 损失[22]来区分m步后的正负样本。InfoNCE 损失如下所示:

$$L_{\mathrm{CPC}}=-\frac{1}{M}\sum_{m=1}^{m}\log_{10}\left[\frac{\exp\left(\widehat{\boldsymbol{z}}_{t+m}^{\mathrm{T}}\boldsymbol{W}_m\boldsymbol{R}_t\right)}{\sum\limits_{Z\in N_{t,m}}\exp\left(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{W}_m\boldsymbol{R}_t\right)}\right], \tag{6}$$

其中, $\boldsymbol{R}_t$表示由回归模型总结得出的上下文向量。

互信息损失与 VQMIVC 中相类似, 采用 CLUB[24]计算互信息的上边界, 可用$I(x,y)$来表示:

$$I(x,y)=E_{P(x,y)}\left[\log_{10}Q_{\theta_{x,y}}(x|y)\right]-E_{P(x)}E_{P(y)}\left[\log_{10}Q_{\theta_{x,y}}(x|y)\right]. \tag{7}$$

定义Z表示语音内容, s表示增强后的说话人身份, p表示基频, 且$x,y\in\{Z,s,p\}$, $Q_{\theta_{x,y}}(x|y)$为变分逼近网络。为了最大程度解耦内容、说话人和音高 3 个特征矢量, 需通过式 (7) 最小化互信息降低各个矢量两两之间的相关性, 因此总体的 MI 损失为

$$L_{\mathrm{MI}}=I(Z,s)+I(Z,p)+I(s,p). \tag{8}$$

在重构语音训练时, 重构的梅尔谱为$\widehat{\boldsymbol{x}}=\{\widehat{\boldsymbol{x}}_1,\widehat{\boldsymbol{x}}_2,\cdots,\widehat{\boldsymbol{x}}_t\}$, 加入干净语音的梅尔谱$\boldsymbol{x}_t$计算重构损失$L_{\mathrm{REC}}$:

$$L_{\mathrm{REC}}=\frac{1}{T}\sum_{t=1}^{T}\left[\left\|\widehat{\boldsymbol{x}}_t-\boldsymbol{x}_t\right\|_1^1+\left\|\widehat{\boldsymbol{x}}_t-\boldsymbol{x}_t\right\|_2^2\right]. \tag{9}$$

2.4　MENR-VC 模型工作流程

MENR-VC 模型的工作流程包括训练阶段和转换阶段。

训练阶段: 首先, 对输入的干净语音及含噪语音进行预处理, 提取语音的梅尔谱和$f0$; 接着, 将干净语音梅尔谱M_c和噪声语音梅尔谱M_n输入到梅尔谱增强模块N_E中, 得到增强后的梅尔谱M_{Ehan}; 然后, 将M_{Ehan}, M_c, $f0$输入到特征解耦模块对应的编码器中, 通过迭代计算损失函数不断优化模型; 最后, 训练得到具有噪声鲁棒的说话人编码器、内容编码器、音高编码器以及解码器。MENR-VC 模型的训练算法如下所示。

MENR-VC模型的训练算法

输入: 干净语音的梅尔谱M_c、音高$f0$, 含噪语音的梅尔谱M_n, 学习率γ和β, 训练迭代次数N
输出: 训练得到的N_E, E_s, E_c, E_{f0}, D, 开始训练:

```
i = 1 %设置迭代次数%
for i <= N do
    M_Ehan ← f(M_c, M_n; N_E), s ← f(M_Ehan; E_s)
    Z ← f(M_c; E_c), p ← f(f0; E_f0)
    %每次更新时, 计算对数似然函数 L_{x,y} = log10 Q_{θ_{x,y}}(x|y)%
    θ_{x,y} ← θ_{x,y} + γ∇_{θ_{x,y}} L_{x,y}, x,y ∈ {Z,s,p}
    %每次更新时, 计算模型的整体损失 L_VC%
    θ ← θ − β∇_θ L_VC, θ ∈ {N_E, E_s, E_c, E_f0, D}
End for
```

返回: N_E, E_s, E_c, E_{f0}, D

转换阶段: 将源说话人和目标说话人语音同时输入到 MENR-VC 模型中, 得到由编码器重构的转换语音梅尔谱, 再将重构的梅尔谱输入到预先训练好的神经声码器中, 最后得到转换语音波形。

3　实验及结果分析

本节共设置了 3 组对比实验: (1) 噪声鲁棒语音转换模型对比实验; (2) 不同信噪比及不可见噪声环境下语音转换结果对比实验; (3) 使用不同语音增强方法的语音转换方法对比实验。采用主观评价指标、客观评价指标以及语谱图可视化等对实验结果进行评测和分析。

3.1　实验设置

实验选取的是 CSTR-VCTK[25]数据集, 该数据集包含 109 个说话人近 43600 句干净语音片段, 每个语音片段持续 4~10 s。按照 10∶1 的比例将数据集分为训练集和测试集, 再将训练集中每个说话人的十分之一语音数据提取出来作为验证集。实验中使用验证集对网络进行交叉验证, 训练集与测试集没有交集, 测试集中的说话人被认为是未出现过的说话人, 可用于执行 one-shot 语音转换。噪声数据来自 CHiME4[26]挑战赛的噪声数据集, 该噪声语料库包含了大约 8.5 h 的背景噪声, 记录在 4 个不同位置 (公共汽车、咖啡馆、步行区和街道) 的噪声数据。对 CSTR-VCTK 数据集内说话人的干净语音分别加入

不同的噪声, 使含噪语音的信噪比分别为−5 dB, 0 dB, 5 dB, 10 dB, 20 dB, 并使用干净语音和含噪语音对模型进行训练。声学特征提取阶段, 所有语音被下采样至 16 kHz, 使用 Librosa 从干净和含噪的语音中提取 80 维梅尔频谱图和 $f0$, 帧长为 25 ms, 帧移为 10 ms。

图 2 中 DCCRN 模型的通道数参考文献 [20] 中的设置: {32, 64, 128, 256, 256, 256}。卷积核的大小为 2, 步长为 1, DCCRN 使用了复数 LSTM, 实部和虚部大小分别是 128。最后一个 LSTM 层后连接了一个 1024 维的 dense 层。VQMIVC 模型中内容编码器的量化器由一个带有 512 个 64 维的可学习向量码本构成, 自回归模型是一个 256 维的单向 RNN 网络。对比预测模块中的预测步长 m 设置为 6, 负样本集合大小为 10。说话人编码器将 DCCRN 模块得到的 80 维梅尔谱作为输入, 经过 8 个 ConvBank 层抓取长时信息, 12 个卷积层、1 个平均池化层, 4 个线性层, 最后推导出 256 维的说话人表示。采用神经声码器 Parallel WaveGAN 对转换后的语音进行合成。通过 Adam 优化器对整个网络模型进行训练, 共训练 500 轮, 批处理大小设定为 64。

3.2 评价指标

实验采用主观评价指标 MOS 评分、客观评价指标梅尔谱失真距离 (MCD)、$f0$ 的均方根误差[27]($f0$−RMSE) 以及字符错误率 (CER)/字错误率 (WER)[28] 对转换语音质量进行评价。此外, 还将不同转换方法输出的语音进行语谱图可视化, 对模型的转换性能进行辅助评价。

主观评价指标 MOS 评分检测是让测试人员对转换后的语音进行评分。通常采用五分制 (5 分: 优秀, 4 分: 良好, 3 分: 一般, 2 分: 差, 1 分: 很差), 分数越高表明转换方法性能越好, 转换语音与目标说话人的语音更为接近。MOS 测试共有 20 名听众参与, 其中 10 名听众学习过语音信号处理相关知识, 另外 10 名听众则为随机选取。在每个测试场景中, 抽取 2 名男性说话人和 2 名女说话人进行随机转换, 每人生成 10 条转换语音。20 名听众采用 MOS 评分对各个转换方法得出的转换语音与目标说话人语音的自然度及相似度进行评价。

客观评价指标梅尔谱失真主要计算转换语音频谱和真实目标语音频谱之间的距离, 其计算公式为

$$\mathrm{MCD}=\frac{10}{\ln 10}\sqrt{2\sum_{n=1}^{N}(m_c-m_t)^2}, \tag{10}$$

其中, N 表示梅尔倒谱的维数, 而 m_c 和 m_t 分别表示第 n 维的转换特征和目标特征的梅尔倒谱系数。转换语音的 $f0$ 与其对应的目标语音 $f0$ 之间的均方根误差为

$$\mathrm{RMSE}=\sqrt{\frac{1}{M}\sum_{i=1}^{M}(\log_{10}(f0_{ci})-\log_{10}(f0_{ti}))^2}, \tag{11}$$

其中, $f0_{ci}$ 和 $f0_{ti}$ 分别表示第 i 维的转换语音和目标语音的 $f0$, 该均方根误差 ($f0-\mathrm{RMSE}$) 的单位为 Hz。以上两种评价指标与语音质量呈负相关, 即数值越小, 表明转换后语音的失真度越小, 语音质量越高。由 ASR 系统评估的 CER 和 WER 评价了被转换语音是否保持了源语音的语言内容, 为了获得 CER 和 WER, 使用了来自 WeNet[28] 的预训练 ASR 模型, 该模型使用 librispeech 语料库进行训练。

3.3 实验结果及分析

以下实验均在源说话人语音为干净语音、目标说话人语音为含噪语音的转换场景下进行。另外, 在数据预处理阶段, 对语音内容进行随机重采样操作[29]。具体而言, 将输入的梅尔谱随机分割成不同长度的片段, 然后将每个片段在时间维度上随机地拉伸或压缩, 最后再将它们拼接到一起。随机重采样操作的引入可以使内容编码器专注于输出语音内容信息, 近一步实现语音内容和节奏信息的特征解耦。

3.3.1 噪声鲁棒语音转换模型结果对比

实验对比对象选取 AutoVC[9]、AdaINVC[11]、基线 VQMIVC[13]、VAED-CN-C[15]、NoiseVC[16] 以及 FlowVC[19] 六种模型。其中, 前三种模型均采用干净的语音数据集进行训练, 并在理想的实验室测试环境下取得了不错的转换效果, 后三种模型则为目前最先进的噪声鲁棒语音转换模型。实验的主观评价结果对比如图 7 所示, 当目标说话人语音含噪时 (实验时使用信噪比为 5 dB 的目标说话人含噪语音), 自然度和相似度方面均优于其他模型, 表明 MENR-VC 模型具有较高的抗噪性。

使用 CER/WER 指标来衡量上述所有转换模型所得语音的可理解性, 用对数尺度 $f0$ 的均方根误差来衡量 $f0$ 失真。实验选择了两名女性和两名男性说话人进行随机转换, 对每个说话人的 30 组转换语音对进行客观评价, 结果如表 1 所示。

表 1 的转换实验中, 统一对目标说话人的干净语音添加了噪声, 加噪后目标说话人语音的信噪比为 5 dB。观察实验结果可知, MENR-VC 模型在各项客观评价指标上均优于基线系统 VQMIVC, 在所有

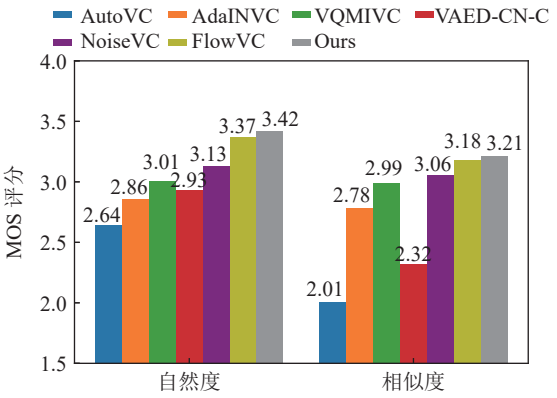

图 7　噪声鲁棒转换模型语音自然度和相似度 MOS 评分对比

方法中取得了最低的 CER 和 WER, 说明 MENR-VC 模型在保留源语音内容的同时具有更好的可理解性。同时, MENR-VC 模型实现了更低的 $f0-\mathrm{RMSE}$, 这表明模型有能力将源语音的详细语调变化进行转换并保留到转换后的语音中。

同时, 与先进的噪声鲁棒模型对比发现, MENR-VC 模型得到的转换语音均取得了较好的实验效果。分析原因为, VAED-CN-C 及 FlowVC 模型中对于说话人身份信息的提取主要依赖声学 ASR 模型的性能, 而 MENR-VC 模型通过融合梅尔谱增强模块的说话人编码器提取噪声鲁棒的说话人矢量, 对模型整体进行训练, 避免因预训练模型性能不足导致的转换模型性能降低的问题。NoiseVC 只对内容编码器进行噪声抑制, 并未考虑语音其他分量在语音转换中的重要作用。

另外, 抽取上述实验中的 4 组实验结果, 对源说话人语音以及转换语音的语谱图进行可视化, 其中源说话人均为男性, 目标说话人为一名男性和一名女性, 含噪目标语音的信噪比为 5 dB, 结果如图 8 所示。

图 8(a)(b)(c) 是男性说话人语音转换为女性说话人语音的实验结果语谱图, 观察结果可知, 对比

表 1　各模型所得转换语音客观评价结果

方法	MCD (dB)	CER (%)	WER (%)	$f0-\mathrm{RMSE}$ (Hz)
AutoVC	11.12	30.39	55.72	30.75
AdaINVC	10.28	28.03	43.75	32.46
VQMIVC	10.05	21.90	37.14	27.31
VAED-CN-C	9.18	14.71	29.31	19.33
NoiseVC	8.95	20.73	34.88	21.05
FlowVC	8.49	16.29	31.14	16.89
Ours	**8.29**	**10.11**	**27.49**	**12.17**

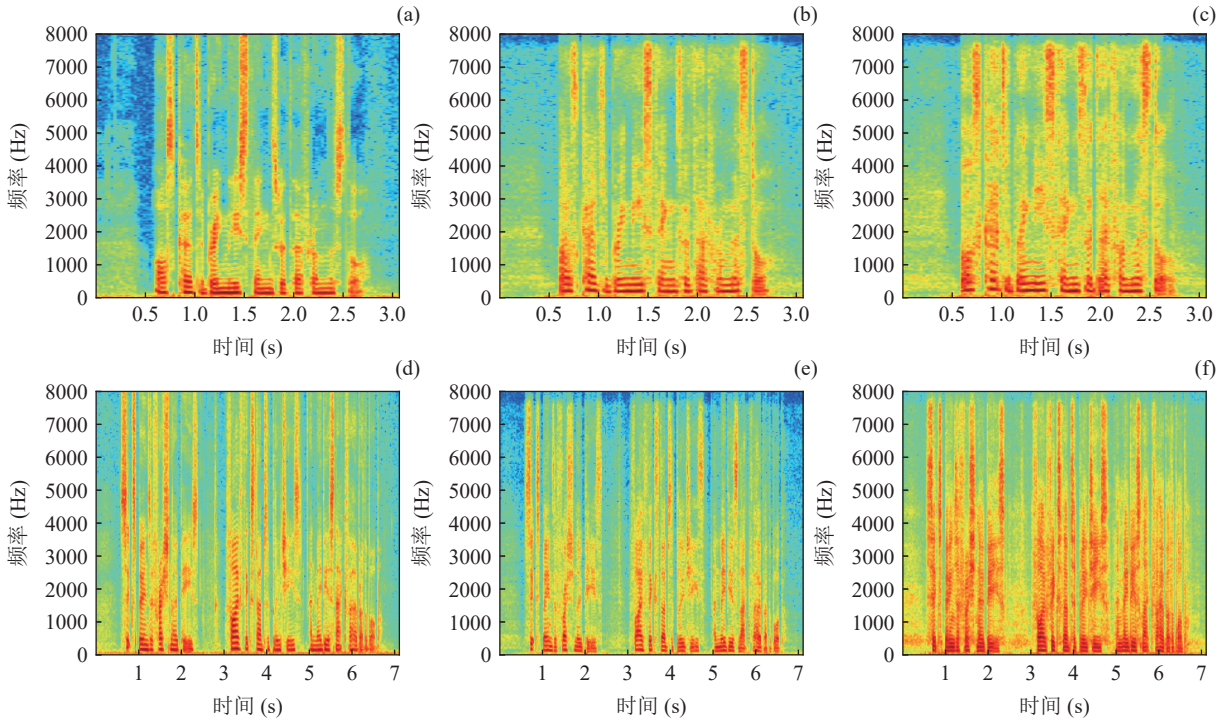

图 8　语谱图可视化对比结果 (a) 源语音 (男性); (b) 转换语音 (女性, 本文方法); (c) 转换语音 (女性, VQMIVC 方法); (d) 源语音 (男性); (e) 转换语音 (男性, 本文方法); (f) 转换语音 (男性, VQMIVC 方法)

VQMIVC 方法, MENR-VC 模型在高频细节上, 与源说话人的语谱图更为接近; 图 8(d)(e)(f) 是男性说话人到男性说话人之间的转换, MENR-VC 模型不仅在低频内容上的表现优于 VQMIVC 方法, 高频部分同样发挥稳定。同时通过人耳的主观判断, 模型得到的转换语音在自然度和相似度上相较 VQMIVC 方法均有显著提高。

3.3.2 不同信噪比及不可见噪声环境下语音转换结果对比

为了验证所提方法对含有不同信噪比噪声的目标说话人语音, 特别是在低信噪比环境下同样具有有效性。对目标说话人语音进行加噪处理, 加入噪声后, 含噪语音的信噪比分别为−5 dB, 0 dB, 5 dB, 10 dB, 20 dB, 计算目标说话人干净语音和转换后语音的梅尔谱失真距离, 同时也分析比较了不同性别转换之间的差异性, 结果如图 9 所示。

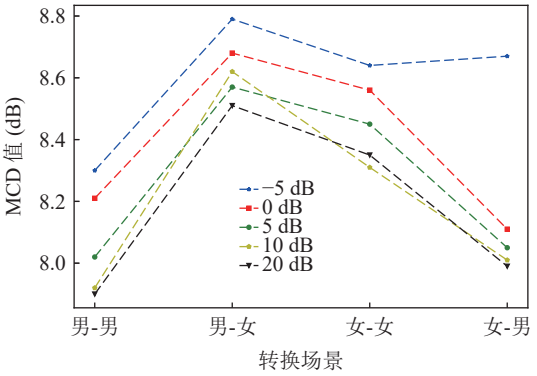

图 9 不同信噪比环境下转换语音的梅尔谱失真距离

由实验结果可知, 无论是在高信噪比或是在−5 dB 这种低信噪比的情况下, MENR-VC 模型均具备一定的鲁棒性, 并且随着信噪比的提高, MCD 值也是逐步降低。另外, 观察性别间的转换发现, 当目标说话人是男性时, 其转换效果普遍优于目标说话人为女性的情况。分析原因可能是, 模型采用梅尔谱作为语音特征进行增强以及转换操作, 相较于男性说话人, 女性说话人的梅尔谱包含更多的高频信息, 因此对女性目标说话人的含噪语音的梅尔谱进行操作的难度要高于男性说话人, 因此导致下游的转换效果男性目标说话人的语音质量要高于女性目标说话人。

验证所提方法在不可见噪声 (即测试所用的噪声类型不属于训练中的噪声类型) 环境下语音转换的有效性。对于不可见噪声条件, 随机抽取了 Noise-92 噪声库中的 babble 和 hfchannel 两种噪声类型, 分别以 5 dB 和 20 dB 信噪比添加到测试语音中, 利用上述含噪测试语音对 MENR-VC 模型及其他模型进行了测试。实验选择了两名女性和两名男性说话人进行随机转换, 对每个说话人的 20 组话语转换对进行客观评价, 结果如表 2 所示。

可以观察到, 随着信噪比水平的增加, 所有系统的 MCD 和 WER 评分都在下降。此外, 在相同噪声类型和信噪比水平下, 由于从干净语音和噪声语音中提取的说话人表征分布不同, 而解码器输出的是干净语音, 这种不匹配降低了解码器的性能。MENR-VC 模型通过使用梅尔谱增强模块获取含噪语音的增强梅尔谱, 以此缓解上述解码器性能下降的问题, 所取得的 MCD 和 WER 评分均优于基线系统 VQMIVC。最后, 比较了可见和不可见噪声类型下的系统性能。就 WER 而言, NoiseVC 和 FlowVC 模型在所有评估的信噪比条件下, 不可见噪声类型相较于可见噪声类型, 性能表现均有所下降, 而 MENR-VC 模型在可见和不可见噪声类型之间的表现更加稳定。

通过实验, 验证了在室内混响的条件下所提模型的有效性。为了快速构建 3D 房间中涉及多个声源和传声器的不同仿真场景, 实验中通过使用 python 包 (pyroomacoustic) 实现房间脉冲响应 (RIR), 给干净的语音增加混响, 得到实验所需的混响语音数据。具体的实验参数设置如下: 房间尺寸为 8 m × 6 m × 5 m (长 × 宽 × 高); 人 (声源) 的位置为 (2 m, 2 m); 传声器的位置为 (3 m, 3 m); 混响时间 T60 分别设置为 200 ms, 400 ms, 600 ms, 800 ms。本实验只在单一传声器、不同混响时间的条件下进行对比实验,

表 2 不可见噪声场景下各模型所得转换语音客观评价结果

场景	可见噪声				不可见噪声			
	MCD (dB)		WER (%)		MCD (dB)		WER (%)	
	5 dB	20 dB	5 dB	20 dB	5 dB	20 dB	5 dB	20 dB
VQMIVC	10.02	9.88	37.14	30.55	10.05	9.85	37.19	30.59
VAED-CN-C	8.65	8.42	29.96	20.06	8.70	8.40	36.42	24.19
NoiseVC	9.25	9.19	34.87	27.75	9.49	9.27	40.26	30.13
FlowVC	9.18	8.79	31.29	23.64	9.23	8.85	37.13	29.34
Ours	**8.30**	**8.21**	**27.52**	**18.10**	**8.30**	**8.19**	**27.89**	**17.99**

实验结果如表 3 所示。

由表 3 结果可知, 相较于目标语音含有加性噪声下的转换结果, 在混响环境中, MENR-VC 模型所得的转换语音在 MCD 和 WER 值上均有所增加, 转换语音质量有所下降。与基线模型 VQMIVC 相比, MENR-VC 模型所得转换语音在 MCD 与 WER 等评价指标上仍具有优势, 但差距并不大, 说明所提模型在目标语音含有混响情况下的转换结果有较大的提升空间, 这也将成为下一步的研究重点。

表 3　室内混响场景下转换语音客观评价结果

场景	200 ms	400 ms	600 ms	800 ms
MCD (dB)	9.88	9.92	9.98	10.03
WER (%)	33.71	34.12	34.89	36.45

3.3.3　不同语音增强方法的语音转换结果对比

本组实验对比了添加不同的语音波形增强方法的噪声鲁棒语音转换方法的性能, 同时, 对比了语音增强、语音转换两个模块级联训练和联合训练的转换效果。主要针对男−男、男−女、女−男、女−女 4 种转换场景, 分别计算了不同方法的梅尔谱失真距离, 实验结果如图 10 所示。几种对比模型介绍如下:

SEGAN-VC: SEGAN[30] 方法主要在波形域上对语音进行操作, 基于生成对抗模型提出了一种端到端的语音增强框架。该对比方法将 SEGAN 模型与 VQMIVC 模型级联训练。

Wiener-VC: 维纳滤波一直作为语音增强的经典方法加入到对比实验中, 实验采取文献 [30] 中的设置, 同样也是与 VQMIVC 模型级联训练。

DCCRN-CL-VC: DCCRN-CL 是文献 [20] 实验中效果较好的一种模式, DCCRN- CL 的通道数为 {32, 64, 128, 256, 256, 256}, 该对比实验是将增强后的语音波形再输入到优化后的 VQMIVC 模型中继续联合训练。

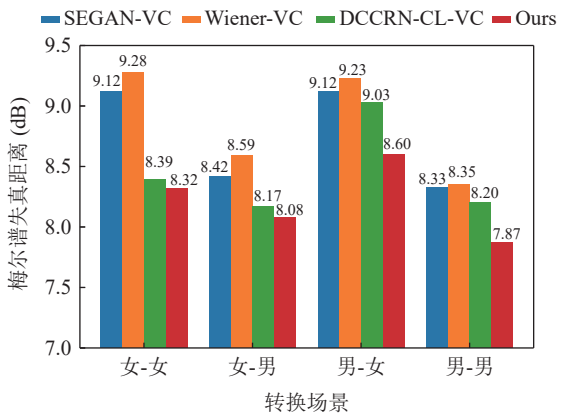

图 10　不同的语音增强方法的梅尔谱失真距离

MENR-VC 模型在 4 种场景下的 MCD 值均优于其他 3 种方法, 分析原因为, 以上 3 种方法均是在波形域对语音进行增强, 得到增强后的语音后, 再提取增强语音的梅尔谱作为说话人编码器的输入来获取说话人矢量。与所提模型直接在频谱上进行增强不同, 中间增加的重构增强语音的环节必然会带来语音的失真, 因此会对噪声鲁棒语音转换模型的训练带来影响。另外, MENR-VC 模型以及同样采用联合训练的 DCCRN-CL-VC 方法, 在实验结果上优于另外两种与 VQMIVC 模型进行级联训练的转换方法, 表明了改进的整体联合损失函数的有效性。

4　结论

为提高含噪场景下所得转换语音的质量, 提出了一种融合梅尔谱增强和特征解耦的噪声鲁棒语音转换方法。该模型引入梅尔谱增强模块, 直接在频域对含噪语音进行增强处理, 避免语音在合成波形过程中的失真。充分利用互信息降低语音内容、音高、说话人身份 3 个特征矢量相互间的依赖性, 通过联合训练增强模块和转换模块提高模型整体的噪声鲁棒性。实验结果表明, 在目标说话人语音含噪以及在多种信噪比的场景下, MENR-VC 模型得到的语音在自然度和说话人相似度方面与基线模型相比分别提高了 0.12 和 0.07。

参　考　文　献

1　张雄伟, 孙蒙, 杨吉斌, 等. 智能语音处理. 北京: 机械工业出版社, 2020

2　Sisman B, Yamagishi J, King S, *et al*. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2021; **29**: 132—157

3　李阳春, 俞一彪. 倒谱本征空间结构化高斯混合模型语音转换方法. 声学学报, 2015; **40**(1): 14—21

4　Erro D, Moreno A, Bonafonte A. Voice conversion based on weighted frequency warping. *IEEE Trans. Audio Speech Lang. Process.*, 2010; **18**(5): 922—931

5　Erro D, Alonso A, Serrano L, *et al*. Interpretable parametric voice conversion functions based on Gaussian mixture models and constrained transformations. *Comput. Speech Lang.*, 2015; **30**(1): 3—15

6　谷东, 简志华. 面向少量语料的语音转换算法. 声学学报, 2018; **43**(5): 138—146

7　Kaneko T, Kameoka H, Tanaka K, *et al*. CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion. IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Brighton, UK, 2019: 6820—6824

8　Kaneko T, Kameoka H, Tanaka K, *et al*. MaskCycleGAN-VC: Learning non-parallel voice conversion with filling in frames. IEEE International Conference on Acoustics, Speech and Signal

Processing, IEEE, Toronto, ON, Canada, 2021: 5919—5923
9 Qian K, Zhang Y, Chang S, *et al*. AutoVC: Zero-shot voice style transfer with only autoencoder loss. International Conference on Machine Learning, IMLS, Long Beach, California, 2019: 5210—5219
10 Wu D Y, Chen Y H, Lee H Y. VQVC + : One-shot voice conversion by vector quantization and U-Net architecture. Interspeech 2020, ISCA, Shanghai, China, 2020: 4691—4695
11 Chou J C, Lee H Y. One-shot voice conversion by separating speaker and content representations with instance normalization. Interspeech 2019, ISCA, Graz, Austria, 2019: 664—668
12 Qian K, Zhang Y, Chang S, *et al*. Unsupervised speech decomposition via triple information bottleneck. International Conference on Machine Learning, IMLS, Vienna, Austria, 2020: 7836—7846
13 Wang D, Deng L, Yu T Y, *et al.* VQMIVC: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. Interspeech 2021, ISCA, Brno, Czechia, 2021: 1344—1348
14 Cong J, Yang S, Xie L, *et al*. Data efficient voice cloning from noisy samples with domain adversarial training. Interspeech 2020, ISCA, Shanghai, China, 2020: 811—815
15 Du H, Xie L, Li H. Noise-robust voice conversion with domain adversarial training. *Neural Networks*, 2022; **48**(4): 74—84
16 Wang S, Borth D. NoiseVC: Towards high quality zero-shot voice conversion. arXiv preprint: 2104.06074, 2021
17 Alejandro M, Jaime L, Sri V K, *et al*. Voicy: Zero-shot non-parallel voice conversion in noisy reverberant environments. 11th ISCA Speech Synthesis Workshop, ISCA, Budapest, Hungary, 2021: 113—117
18 Cong J, Yang S, Xie L, *et al*. Glow-WaveGAN: Learning speech representations from GAN-based variational auto-encoder for high fidelity flow-based speech synthesis. Interspeech 2021, ISCA, Brno, Czechia, 2021: 2182—2186
19 Xue L M, Yang S, Hu N, *et al*. Learning noise-independent speech representation for high-quality voice conversion for noisy target speakers. Interspeech 2022, ISCA, Incheon, Korea, 2022: 2548—2552
20 Hu Y, Liu Y, Lv S, *et al*. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. Interspeech 2020, ISCA, Shanghai, China, 2020: 2472—2476
21 Yamamoto R, Song E, Kim J M. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Barcelona, Spain, 2020: 6199—6203
22 Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. ArXiv preprint: 1807.03748, 2018
23 Wang Y, Skerry-Ryan R J, Stanton D, *et al*. Tacotron: Towards end-to-end speech synthesis. Interspeech 2017, ISCA, Stockholm, Sweden, 2017: 4006—4010
24 Cheng P, Hao W, Dai S, *et al*. CLUB: A contrastive log-ratio upper bound of mutual information. International Conference on Machine Learning, IMLS, Vienna, Austria, 2020: 1779—1788
25 Veaux C, Yamagishi J, MacDonald K, *et al*. Superseded-CSRT VCTK Corpus: English multi-speaker corpus for CSRT voice cloning toolkit. The Centre for Speech Technology Research (CSTR), University of Edinburgh, 2016
26 Vincent E, Watanabe S, Nugraha A A, *et al*. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Comput. Speech Lang.*, 2017; **46**: 535—557
27 Ming H, Huang D, Dong M, *et al*. Fundamental frequency modeling using wavelets for emotional voice conversion. International Conference on Affective Computing & Intelligent Interaction, IEEE, Xi'an, China, 2015: 804—809
28 Yao Z, Wu D, Wang X, *et al*. WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. Interspeech 2021, ISCA, Brno, Czechia, 2021: 4054—4058
29 Polyak A, Wolf L. Attention-based WaveNet autoencoder for universal voice conversion. IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Brighton, UK, 2019: 6800—6804
30 Pascual S, Bonafonte A, Serrà J. SEGAN: Speech enhancement generative adversarial network. Interspeech 2017, ISCA, Stockholm, 2017: 3642—3646