计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(9):216−225 [doi: 10.15888/j.cnki.csa.009637]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多重互信息约束的高表现力语音转换①

王　光[1], 刘宗泽[1], 姜彦吉[1,3], 董　浩[2,3]

[1](辽宁工程技术大学 软件学院, 葫芦岛 125105)
[2](清华大学苏州汽车研究院, 苏州 215134)
[3](优策 (江苏) 安全科技有限公司 OpenSafe 实验室, 苏州 215100)
通信作者: 姜彦吉, E-mail: jyjvip@126.com

摘　要: 随着语音转换在人机交互领域的广泛应用, 对于获取高表现力语音的需求日益显著. 当前语音转换主要通过解耦声学特征实现, 侧重对内容和音色特征的解耦, 很少考虑语音中混合的情感特性, 导致转换音频情感表现力不足. 为解决上述问题, 本文提出一种基于多重互信息约束的高表现力语音转换模型 (MMIC-EVC). 在对内容和音色特征进行解耦的基础上, 引入表现力模块分别对话语级韵律和节奏特征进行建模, 以实现情感特性的传递; 随后通过最小化各特征之间的多重互信息变分对数上界, 约束各编码器专注于解耦对应的声学嵌入. 在 CSTR-VCTK 和 ESD 语音数据集上的实验表明, 本模型的转换音频语音自然度评分 (MOS) 达到 3.78, 梅尔倒谱失真为 5.39 dB, 最佳最差占比测试结果大幅领先于基线模型, MMIC-EVC 能够有效解耦韵律和节奏特征, 并实现高表现力语音转换, 为人机交互带来更加出色和自然的用户体验.

关键词: 语音转换; 特征解耦; 互信息约束; 韵律建模; 人机交互

引用格式: 王光,刘宗泽,姜彦吉,董浩.基于多重互信息约束的高表现力语音转换.计算机系统应用,2024,33(9):216–225. http://www.c-s-a.org.cn/1003-3254/9637.html

High Expressiveness Voice Conversion Based on Multiple Mutual Information Constraints

WANG Guang[1], LIU Zong-Ze[1], JIANG Yan-Ji[1,3], DONG Hao[2,3]

[1](Software College, Liaoning Technical University, Huludao 125105, China)
[2](Suzhou Automotive Research Institute, Tsinghua University, Suzhou 215134, China)
[3](OpenSafe Laboratory, Youce (Jiangsu) Safety Technology Co. Ltd., Suzhou 215100, China)

Abstract: As voice conversion technology becomes increasingly prevalent in human-computer interaction, the need for highly expressive speech continues to grow. Currently, voice conversion primarily relies on decoupling acoustic features, emphasizing the decoupling of content and timbre features, but often neglects the emotional features in speech, resulting in insufficient emotional expressiveness in converted audio. To address this problem, this study introduces a novel model for highly expressive voice conversion with multiple mutual information constraints (MMIC-EVC). On top of decoupling content and timbre features, the model incorporates an expressiveness module to capture discourse-level prosody and rhythm features, enabling the conveyance of emotional features. It constrains every encoder to focus on its acoustic embedding by minimizing the variational upper bounds of multiple mutual information between features. Experiments on the CSTR-VCTK and ESD speech datasets indicate that the converted audio of the proposed model achieves a mean opinion score of 3.78 for naturalness and a Mel cepstral distortion of 5.39 dB, significantly outperforming baseline models in the best-worst sensitivity test. The MMIC-EVC model effectively decouples rhythmic and prosodic features, facilitating high expressiveness in voice conversion, and thereby providing a more natural and better user experience in human-computer interaction.

① 基金项目: 辽宁省教育厅面上项目 (LJKZ0338); 葫芦岛市科技计划 (2023JH(1)4/02b); 广东省科技创新战略专项市县科技创新支撑项目 (STKJ2023071)
收稿时间: 2024-03-26; 修改时间: 2024-04-23; 采用时间: 2024-05-09; csa 在线出版时间: 2024-07-26
CNKI 网络首发时间: 2024-07-29

Key words: voice conversion; feature decoupling; mutual information constraint; prosody modeling; human-computer interaction

语音转换旨在将语音中源说话人的声音转换为目标说话人的声音, 同时保留源语音的语言内容[1,2]. 语音转换技术的广泛应用为人机交互领域注入了新的活力, 涵盖多媒体娱乐、医疗辅助、语音导航、智能家居等多个方面, 高效的语音转换技术为人机交互提供了更便捷、更智能的体验, 并推动科技的不断创新.

传统语音转换方法包括高斯混合模型 (Gaussian mixture model, GMM)[3]、示例方法[4]和稀疏表示[5]等, 各自采用不同策略处理语音信号. GMM 通过概率分布进行建模, 示例方法依赖大量样本捕获特征, 稀疏表示方法通过简化模型专注于提取关键信息. 但是这些统计方法对语音的表示能力相对有限, 难以捕捉语音中丰富的表现力. 随着深度学习的出现, 基于深度神经网络 (deep neural network, DNN)[6]和循环神经网络 (recurrent neural network, RNN)[7]等数据驱动方法相继被提出.

尽管这些网络模型拥有强大的拟合能力[8], 但其固定长度的输入输出序列在处理不同时长的语音数据时存在限制, 此外在进行并行训练时仍需要额外执行数据之间的对齐. 为此研究人员对非并行的语音转换进行探索, 如 CycleGAN[9]、StarGAN[10]等基于生成对抗网络 (generative adversarial network, GAN) 方法, 在生成器和判别器之间进行反复迭代以逼近目标语音. 然而生成对抗模型通常难以训练, 并且仅能对训练集中出现的说话人之间进行转换.

为实现跨说话人的语音转换, 近几年研究者采用特征解耦技术, 如 VQVC+[11]和 Again-VC[12]等, 分别从源语音和目标语音中分离出语言内容和说话人音色等关键声学特征, 进而利用这些独立的特征进行语音的重构. 在这个过程中, 自动编码器架构发挥着核心作用, 编码器专注于提取语音的内容嵌入和音色嵌入, 解码器则依据这些特征以及额外提供的目标音色信息重构梅尔频谱, 并通过声码器生成转换音频, 这种结构设计能够有效进行特征解耦. 尽管上述方法在实现跨说话人语音转换任务中表现出较好的效果, 但模型仅对音色和内容特征进行解耦, 忽略了语音中携带的情感特性, 这一局限性导致转换后的音频与源语音具有相似的音调和节奏[13], 使听觉感受上较为生硬.

随着个性化人机交互的不断发展, 对语音转换效果的需求已不仅限于满足高自然度, 同时也对在多种场景中能够展现出语音的情感特性提出更高的要求. Xie 等人[14]的研究表明, 通过将部分韵律特征进行建模, 并整合入语音转换模型中, 不仅能够丰富转换音频的多样性, 还能为执行情感语音处理、情感语音识别等下游任务提供基本框架[15]. 语音的韵律特性, 包括频谱、音调和节奏等元素, 与内容和音色等声学特征紧密相连, 它们的相互作用和纠缠使得特征解耦变得复杂且具有挑战性. 因此, 进一步解耦这些元素, 并全方位地对韵律进行建模, 对于提升语音转换的表现力至关重要. Cheng 等人[16]的研究表明, 互信息损失在特征解耦方面展现出良好的效果, 为本研究提供了新的思路.

针对上述问题, 本文提出一种基于多重互信息约束的高表现力语音转换模型 (expressive voice conversion with multiple mutual information constraints, MMIC-EVC). 结合对语音情感特性的分析, 将语音分解为内容、音色、音高和节奏 4 个因素, 通过自动编码器结构进行特征提取及重构. 首先对音高进行沿时间维度的随机重采样[17]操作, 以去除音高中混杂的节奏信息; 随后引入内容编码器、说话人编码器、音高编码器以及节奏编码器分别对各特征进行建模; 在训练过程中, 计算多重互信息的变分对数上界 (variational contrastive log-ratio upper bound, vCLUB)[18], 并通过最小化该上界的方式减小各特征矢量之间的相关性, 即实现内容、音色、音高及节奏等特征的解耦. 最后将声学嵌入送入语音解码器和 WaveNet 声码器[19]重建梅尔频谱图并输出语音.

本文的主要贡献如下: (1) 针对转换音频表现力不足的问题, 提出一种高表现力语音转换模型, 通过构建表现力模块对音高和节奏进行精细建模, 从而在转换过程中有效地传递语音的韵律特性, 显著提升转换音频的自然度和表现力. (2) 为进一步解耦语音中的声学特征, 首次提出一种基于多重互信息约束的特征解耦方法, 约束编码器专注于学习各自对应的声学嵌入, 有效降低不同声学特征之间的相互依赖性.

1 特征解耦中的韵律建模

韵律作为语音表达的关键要素, 直接影响着听众对语音的理解和情感的传递. 在进行特征解耦的过程中, 韵律特征通常与音素及其他声学特征紧密相连[20], 因此难以完全从语音中学习韵律信息并单独建模, 所以如何更好地进行韵律建模显得尤为关键.

Tacotron[21]首次提出通过无监督学习进行韵律建模, 将目标语音编码成一个矢量, 该矢量代表语句级的韵律嵌入. CHiVE[22]使用分层结构的条件变分自动编码器, 每一层分别对基频、能量和持续时间进行特征提取. IQDUBBING[23]使用一个韵律提取器和两个韵律过滤器来提取韵律特征. 然而韵律特征并不能够简单地表示为一个固定长度的矢量或者特征集合.

AutoVC[24]应用信息瓶颈结构, 迫使编码器丢弃与输入语音冗余的信息, 在不需要任何文本转录的情况下进行特征解耦. 研究人员受此启发, 开始尝试在 AutoVC 模型的基础上进一步控制语音的韵律属性. AutoPST[25]将韵律信息定义为音素的持续时间, 通过基于相似性的重采样对节奏进行解耦. Qian 等人采用 SpeechSplit[26]模型重点关注音高和节奏的建模问题, 在无监督的方式下使用多个自动编码器将语音分解为内容、音色、音高和节奏, 这一过程需要对自动编码器的瓶颈特征进行精细调整, 瓶颈的质量直接影响到模型对情感特征的学习和再现能力. 上述方法都需要对特征矢量的维度施加严格的限制, 以适当平衡语音质量和解耦效果, 鲁棒性较差.

SpeechSplit2.0[27]沿用与 SpeechSplit 相似的架构, 并通过应用多种高效的信号处理技术, 实现韵律特征在自动编码器上的解耦并建模, 避免复杂的瓶颈特征调整过程. 但 SpeechSplit2.0 在进行训练集之外的说话人韵律建模时效果不佳, 各特征矢量之间解耦的充分性难以衡量. 为进一步提高语音转换的表现力, 本文提出 MMIC-EVC 模型, 融合多重互信息约束策略, 对语音中不同声学特征矢量之间进行充分解耦, 并引入表现力模块, 分别对音高和节奏进行精细建模, 保留语音的韵律特征.

2 基于多重互信息的表现力语音转换

2.1 模型架构

在本节中, 详细介绍基于多重互信息约束的高表现力语音转换模型框架, 即 MMIC-EVC 模型. 该模型采用端到端结构, 将语音转换过程划分为两个关键阶段: 解耦阶段和合成阶段. MMIC-EVC 模型的整体架构如图 1 所示. 在解耦阶段, MMIC-EVC 模型包括 3 个主要部分: 表现力模块、内容编码器 *Ec* 和音色编码器 *Es*. 其中表现力模块用以捕捉音频的情感特性和时长信息, 以便后续合成阶段能够较好地传递语音的情感特性, 该模块由音高编码器 *Ep* 和节奏编码器 *Er* 组成, 各编码器详细结构如图 2 所示. 使用上述 4 种编码器分别对语音的内容、音色、音高和节奏特征进行建模, 并引入互信息进行约束, 通过最小化多重互信息的变分对比上界降低各特征矢量之间的依赖关系. 在合成阶段, MMIC-EVC 模型包括解码器 *D* 和 WaveNet 声码器两部分, 解码器将各编码器的输出的声学嵌入映射为梅尔谱, 并通过 WaveNet 声码器输出语音.

● 节奏编码器. 节奏特征描述声音的语速、停顿和重音分布等, 节奏的变化不仅影响语句的流畅性和可理解性, 还能传达说话人的情感和意图. 为了有效地捕捉这种复杂的时间序列数据, MMIC-EVC 模型对话语级的节奏特征进行单独建模, 输入音频 X 转换为梅尔谱 (*M*), 并将其输入节奏编码器中. 该编码器由 1 个 128-5-1 的卷积层构成, 采用 ReLU 激活函数增强非线性表达能力, 随后进行组归一化处理, 该操作可提高编码器的训练速度, 并增强模型的泛化能力. 随后被送入双向长短期记忆网络 (bidirectional long short-term memory, Bi-LSTM) 层, 使得模型能够捕捉序列数据中的长期依赖关系, 同时保留重要信息并减少特征维度, 为转换音频提供准确而生动的节奏变化. 节奏编码器的详细设计如图 2(a) 所示, 提取的节奏嵌入 (*Zr*) 表示如下:

$$Zr = Er(M) \tag{1}$$

● 音高编码器. 音高特征描述声音高低程度的关键声学属性, 可以表达不同的情感和语气及单词意义变化的决定因素. 音高编码器的输入为语音中获取的音高轮廓 (*P*), 由于音高轮廓携带节奏信息, 在馈送至该编码器之前, 沿时间维度执行随机重采样 (*RR*) 操作, 以去除节奏信息. 随机重采样涉及两个操作步骤. 首先将输入音高轮廓并分割成随机长度的片段, 其次, 沿时间维度对每个片段进行随机拉伸或挤压, 从而导致节奏信息的丢失或变形. 音高编码器的结构由 3 个 256-5-1 的卷积层构成, 采用 ReLU 函数进行激活, 随后进行组

归一化处理, 最后送入 1 个 Bi-LSTM 层. 音高编码器的详细设计如图 2(b) 所示, 提取的音高嵌入 (*Zp*) 表示如下:

$$Zp = Ep(RR(P)) \tag{2}$$

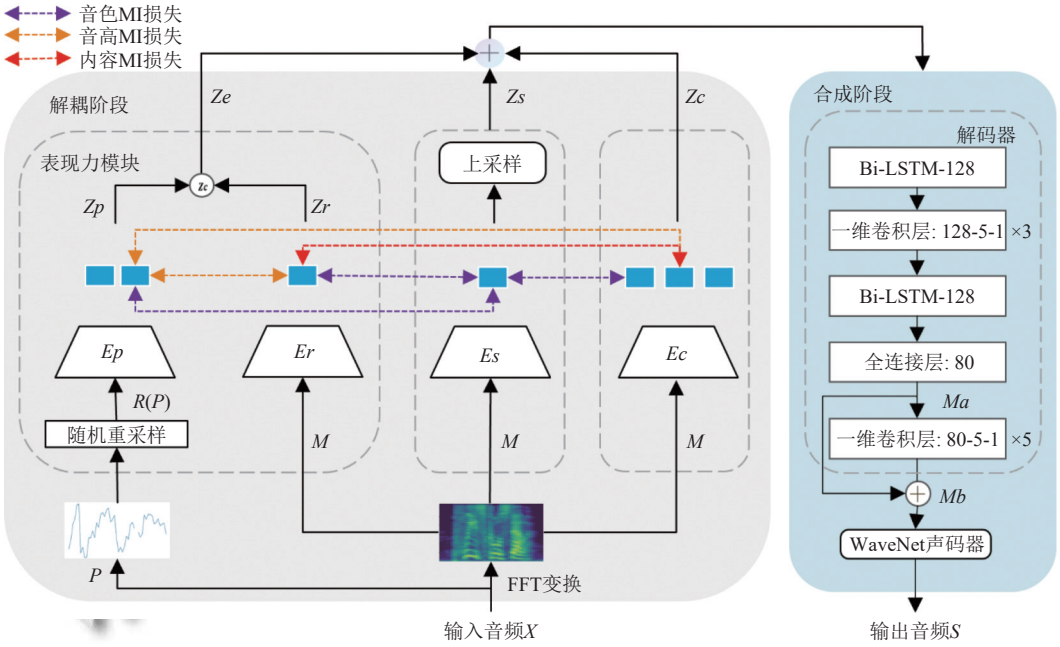


图 1 基于多重互信息约束的高表现力语音转换模型框架图

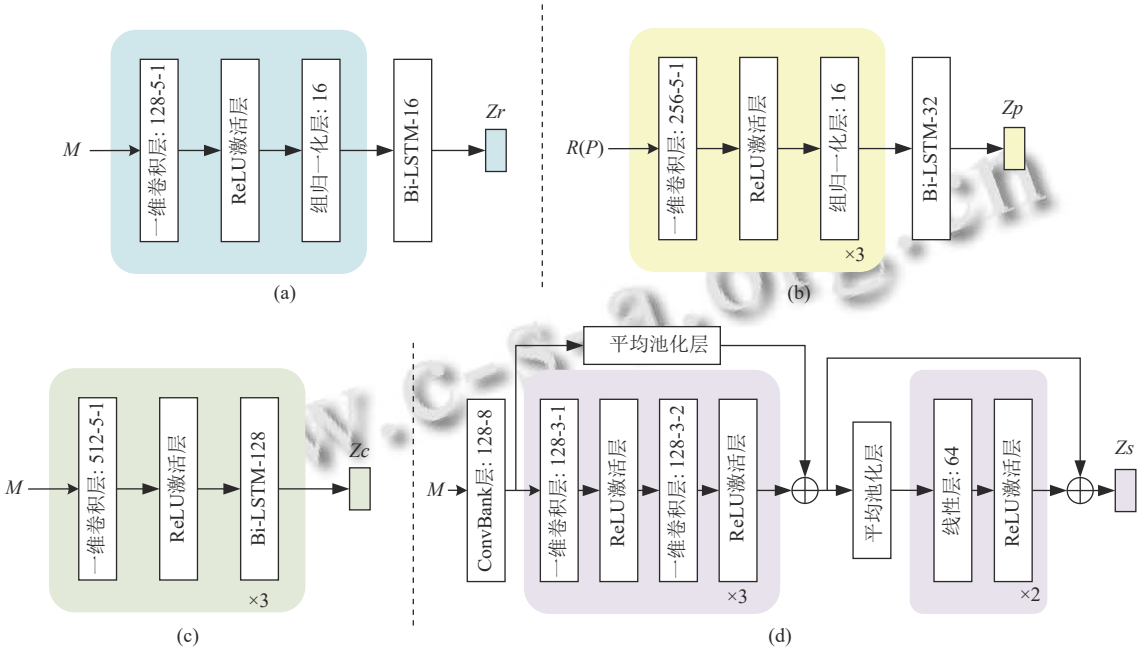


图 2 各编码器结构图

● 内容编码器. 内容特征描述语音中的语义信息, 即实际传达的文字或语言信息, 确保语音传达的意图和信息的准确性. 内容编码器由 1 个 512-5-1 的卷积层构成, 负责从语音信号中提取初步的特征, 采用 ReLU 激活函数, 模型能够在非线性变换中保持信息的丰富性, 同时减少梯度消失的问题, 随后, 这些特征被送入 3 个 Bi-LSTM 层, 使得模型能够同时考虑到序列数据中的前向和后向信息, 可以更加全面地学习语音信号中的语义内容. 音高编码器的详细设计如图 2(c) 所示, 提取的内容嵌入 (*Zc*) 表示如下:

$$Zc = Er(M) \tag{3}$$

● 音色编码器. 音色特征用于区分不同说话人之间声音的独特属性, 音色编码器由 ConvBank 层、3 个一维卷积与激活函数的组合、2 个线性层与激活函数的组合, 以及平均池化层等 4 个关键模块组成. ConvBank 层用于从输入序列中精确提取局部特征和上下文信息, 扩大感受野范围, 从而更好地捕捉长时序信息; 一维卷积与 ReLU 激活函数组合优化特征提取的非线性能力; 平均池化层加强关注全局信息, 有助于准确捕获音色特征; 最后线性层与 ReLU 激活函数的串联组合, 进一步提升音色编码器的映射能力. 音色编码器的详细设计如图 2(d) 所示, 提取的音色嵌入 (Zs) 表示如下:

$$Zs = Es(M) \tag{4}$$

● 解码器. 首先将学习到的节奏嵌入、音高嵌入、内容嵌入以及音色嵌入按通道维度连接在一起, 形成一个综合的特征向量. 并将其送入 Bi-LSTM 层学习上下文信息, 随后将特征向量送入 3 个 128-5-1 的卷积层, 在卷积操作之后, 特征向量再次被送入 1 个 Bi-LSTM 层, 经过最后 1 个全连接层后, 输出为转换后语音的 80 维梅尔谱 Ma. 将 Ma 再次用作卷积层的输入, 经过 5 个附加卷积层, 与 Ma 叠加后得到转换的梅尔谱 Mb, 附加卷积层进一步提取特征向量表示, 以获得更好的转换效果, 解码器 D 生成梅尔谱 Mb 的表示如下:

$$Mb = D(Zc, Zs, Zr, Zp) \tag{5}$$

最后将梅尔谱 Mb 送入 WaveNet 声码器输出转换音频 S.

2.2 融合双重互信息

互信息 (mutual information, MI) 用来描述随机变量 X 与随机变量 Y 之间相互依赖程度, 公式表达如下:

$$\begin{aligned} I(X,Y) &= \int P(X,Y)\log_{10}\frac{P(X,Y)}{P(X)P(Y)}\mathrm{d}X\mathrm{d}Y \\ &= E_{P(X,Y)}\left[\log_{10}\frac{P(X,Y)}{P(X)P(Y)}\right] \end{aligned} \tag{6}$$

其中, $P(X)$ 和 $P(Y)$ 分别是 X 和 Y 的边缘分布, $P(X, Y)$ 表示 X 和 Y 的联合分布. 互信息的变分对数上界 (vCLUB) 则用于计算两种随机变量的互信息上限. 变分对数上界定义为:

$$\begin{aligned} \hat{I}(X,Y) =& E_{P(X,Y)}[\log q_\theta(X|Y)] \\ & - E_{P(X)}E_{P(Y)}[\log q_\theta(X|Y)] \end{aligned} \tag{7}$$

其中, $X, Y \in \{Zc, Zs, Zr, Zp\}$, $q_\theta(X|Y)$ 则为变分逼近网络, 使用参数 θ 来近似 $P(X|Y)$. 对于无偏估计样本集 $\{x_i, y_i\}$, 其变分对数上界的无偏估计量表示为:

$$\hat{I}(X,Y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\log q_\theta(x_i|y_i) - \log q_\theta\left(x_j|y_i\right)\right] \tag{8}$$

其中, $x_i, y_i \in \{Zc_i, Zs_i, Zr_i, Zp_i\}$. 通过最小化互信息的变分对数上界, 可以减少不同声学特征之间的相互依赖.

在 MMIC-EVC 模型中, 使用多重互信息损失以提高解耦效果, 在训练过程中, 引入音色互信息损失, 编码器被约束以不断降低音色嵌入与内容嵌入、节奏嵌入和音高嵌入等特征之间的相关性, 以实现音色特征解耦; 同时引入音高互信息损失, 将音高嵌入与内容嵌入及节奏嵌入进行解耦, 以实现音高特征的迁移; 最后为去除转换音频内容特征中杂糅的节奏特征, 引入内容互信息损失, 对内容嵌入和节奏嵌入进特征解耦.

(1) 音色互信息损失 (L_{SMI}): 计算音色嵌入 (Zs) 与内容嵌入 (Zc)、节奏嵌入 (Zr) 以及音高嵌入 (Zp) 之间的互信息, 其公式如下:

$$L_{\mathrm{SMI}} = \hat{I}(Zs,Zc) + \hat{I}(Zs,Zr) + \hat{I}(Zs,Zp) \tag{9}$$

(2) 音高互信息损失 (L_{PMI}): 计算音高嵌入 (Zp) 与内容嵌入 (Zc) 以及节奏嵌入 (Zr) 之间的互信息, 其公式如下:

$$L_{\mathrm{PMI}} = \hat{I}(Zp,Zc) + \hat{I}(Zp,Zr) \tag{10}$$

(3) 内容互信息损失 (L_{CMI}): 计算节奏嵌入 (Zr) 与内容嵌入 (Zc) 之间的互信息, 其公式如下:

$$L_{\mathrm{CMI}} = \hat{I}(Zc,Zr) \tag{11}$$

2.3 损失函数

在训练过程中, 解码器 D 的输出旨在学习准确地重构输入的语音 X, 为实现这一目标, 解码器与编码器联合训练, 并通过最小化重建损失来优化模型性能. 重建损失 L_D 被定义为重建语音 S 与输入语音 X 之间的 L1 范数和 L2 范数差的平方的期望值:

$$L_D = E\left[\| Xab - X \|_1^2 + \| Xab - X \|_2^2\right] \tag{12}$$

在每次迭代的训练中, 首先进行变分近似网络的优化, 目的是最大化似然对数 $\log q_\theta(X|Y)$, 多重互信息损失 L_{MI} 表示为音色互信息损失 L_{SMI}、音高互信息损失 L_{PMI} 和内容互信息损失 L_{CMI} 的加权和:

$$L_{\mathrm{MI}} = \lambda_{\mathrm{SMI}}L_{\mathrm{SMI}} + \lambda_{\mathrm{PMI}}L_{\mathrm{PMI}} + \lambda_{\mathrm{CMI}}L_{\mathrm{CMI}} \tag{13}$$

其中, λ_{SMI}、λ_{PMI}、λ_{CMI} 和是用于调节互信息损失以

增强解耦的权重的超参数.

随后进行表现力语音转换模型的优化, 其总损失 L 可以表示为:

$$L = L_D + L_{\mathrm{MI}} \tag{14}$$

3 实验及结果分析

3.1 实验设置

实验选取语音转换研究主流使用的 CSTR-VCTK 数据集[28]和 ESD 数据集[29], CSTR-VCTK 数据集包括 109 位英语说话人的语音样本, 每位说话人阅读约 400 个句子. 随机选取 89 位说话人作为训练集, 10 位说话人作为验证集, 其余 10 位说话人用于测试集. ESD 数据集中包括 10 位母语为英语的说话人和 10 位母语为汉语的说话人录制的音频, 每种语言的说话人包含 5 位男性和 5 位女性, 每位说话人使用自然、高兴、愤怒、悲伤和惊讶的 5 种情感, 分别录制 350 个音频, 平均持续时间为 2.9 s. 随机选取 12 位说话人作为训练集, 4 位说话人作为验证集, 其余 4 位说话人用于测试集.

将所有语音下采样至 16 kHz, 对音频进行预加重、分帧, 并进行加窗操作, 采用窗口大小为 1 024, 滑动大小为 256 的 Hanning 窗, 并进行短时傅里叶变换后, 随后采用 FFT 大小为 1 024 的短时傅里叶变换, 用来计算梅尔谱图, 使用跨度为 90 Hz–7.6 kHz 的 80 通道梅尔滤波器组将 STFT 幅度转换为梅尔刻度.

MMIC-EVC 模型在单个 NVIDIA 3060 GPU 上进行训练, 并使用 ADAM 优化器, 学习率为 1E–4, $\beta 1$=0.9, $\beta 2$=0.99, 批量大小为 16, 设置 λ_{SMI}=0.1, λ_{PMI}=0.2, λ_{CMI}=0.1, 并使用预训练的 WaveNet 声码器, 将梅尔谱解码成语音.

在实验中选取 VQMIVC[30]、SRDVC[17]等先进的基线模型进行比较, 所有基线模型均使用与 MMIC-EVC 相同的训练集、验证集和测试集.

3.2 语音转换效果评价

3.2.1 主观评价

主观测试采用两种评价指标, 分别为语音自然度平均意见得分 (mean opinion score, MOS) 和最佳最差占比方法 (best-worst sensitivity, BWS).

在语音自然度 MOS 实验中, 测试人员使用五分制进行对转换后的语音进行评分, 评分标准为 1–5 分, 分数越高表明转换方法性能越好, 转换音频更加接近真实语音. 测试过程中, 测试人员需听取每个语音对, 并对音频质量进行打分, 随后计算 MOS 评分, 实验采用 95% 的置信区间.

在 BWS 实验中, 该方法用于评估一组实验中的不同模型进行语音转换效果. 测试人员被要求从一组语音对中选择他们认为情感表现力是最佳或最差的语音, 从而可以计算出每种模型情感表现力的最佳最差占比, 以验证进行转换后的语音情感表现力.

主观测试共有 16 名 20–30 岁之间的研究生听众参与, 其中 12 人 (6 男 6 女) 具有语音测评经验, 另外 4 人 (2 男 2 女) 则为随机选取. 转换场景为自然-自然 (N-N)、自然-愤怒 (N-A)、自然-快乐 (N-H)、自然-惊讶 (N-SUP) 和自然-悲伤 (N-SAD) 的测试中, 每种场景分别随机选取 10 个语音对, 每个语音对中包含源音频、目标说话人音频, 以及使用 MMIC-EVC 与 VQMIVC、SRDVC 基线模型分别对同一条语音进行转换后得到的音频. 16 名听众采用 MOS 评分与 BWS 评分依次进行评价.

语音自然度测试的实验结果如表 1 所示: 在 5 种不同的场景中, MMIC-EVC 的 MOS 评分分别达到了 3.78、3.72、3.75、3.66 和 3.68, 在自然-自然 (N-N) 情境中, 评分均超越基线模型 0.19 以上; 在自然-惊讶 (N-SUP) 场景中, 均超越基线模型 0.26 以上. 表明 MMIC-EVC 在生成转换音频质量方面具备显著优势.

表 1 语音自然度实验结果

方法	N-N	N-H	N-SUP	N-SAD	N-A
VQMIVC	3.35±0.07	3.37±0.08	3.24±0.12	3.39±0.08	3.29±0.11
SRDVC	3.59±0.09	3.61±0.09	3.49±0.11	3.59±0.07	3.56±0.09
Ours	**3.78±0.07**	**3.72±0.08**	**3.75±0.09**	**3.66±0.10**	**3.68±0.11**

BWS 实验结果如图 3 所示, 在进行自然-自然 (N-N) 场景测试时, 听众认为有 136 组语音对中 MMIC-EVC 表现最佳, 占比高达 85%, 明显优于基线模型 VQMIVC 3% 和 SRDVC 12% 的最佳占比结果; 同时在 10 组语音对中, 听众认为 MMIC-EVC 表现最差, 占比 6%.

在进行自然-快乐 (N-H) 和自然-愤怒 (N-A) 场景测试时表现相似, 听众认为分别有 123 组和 125 组语音对中 MMIC-EVC 表现最佳, 占比为 77% 和 78%; 另外在自然-快乐 (N-H) 场景中, 最差占比为 15%, 在自然-愤怒 (N-A) 场景中, 最差占比为 8%.

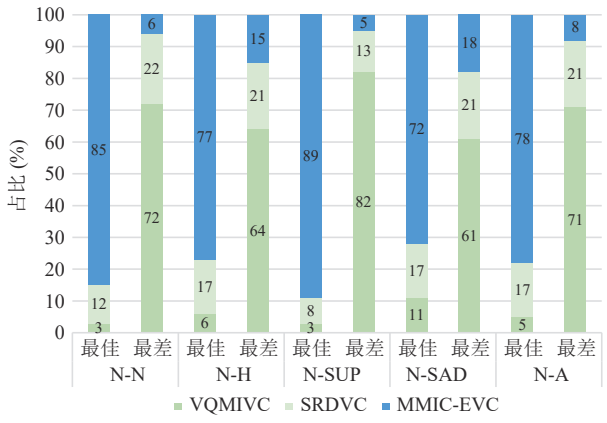


图 3 BWS 实验结果

在自然-惊讶 (N-SUP) 场景中, MMIC-EVC 模型在最佳类别中获得了压倒性的认可, 占比高达 89%. 最差类别的占比仅为 5%. 在自然-悲伤 (N-SAD) 场景中, MMIC-EVC 模型的表现较为不佳, 最佳占比仅有 72%, 同时最差占比为 18%, 这是由于悲伤情感通常伴随着较低的音调和较慢的节奏, 以及更多的情感细节, 如颤抖或断续, 这些因素都增加了语音转换的难度.

BWS 实验结果表明: MMIC-EVC 模型在多种场景中表现出优越性能, 特别是在“最佳”类别中, 其得分显著高于 VQMIVC 和 SRDVC 模型, 转换音频更加自然, 在传达语义和情感方面也更为有效和富有表现力.

3.2.2 客观评价

在客观实验中采用以下 3 种评价指标, 分别为梅尔倒谱失真 (Mel cepstral distortion, *MCD*)、单词错误率 (word error rate, WER) 和 logF0 的皮尔逊相关系数 (Pearson correlation coefficient, PCC).

MCD: 梅尔倒谱失真衡量两个频谱之间的欧几里德距离, 用来评估合成语音和目标语音之间的差异程度. 计算公式为:

$$MCD = \frac{10}{\ln 10}\sqrt{2\sum_{n=1}^{N}(S_c - S_t)^2} \tag{15}$$

其中, S_t 和 S_c 分别表示转换音频与目标语音的梅尔倒谱系数. *MCD* 的值越低, 表示两个音频片段之间的失真越小, 即它们的梅尔倒谱特征越相似, 音频片段之间的差异越小.

WER: 验证转换音频能否对源语音的内容特征进行保留, 使用 ESPnet2 模型中的自动语音识别模块[31]计算得到, 该模块由 LibriSpeech 语料库进行训练.

logF0 PCC: 通过计算源语音和转换音频的 F0 之间的皮尔逊相关系数, 并将其进行对数转换, 取值范围控制在[−1, 1], 反映转换音频的语调变化.

在转换场景为自然-自然 (N-N)、自然-愤怒 (N-A)、自然-快乐 (N-H)、自然-惊讶 (N-SUP) 和自然-悲伤 (N-SAD) 的测试中, 每个场景分别随机选取 20 个语音对, 并计算 3 种客观指标的平均值, 实验结果如图 4 所示.

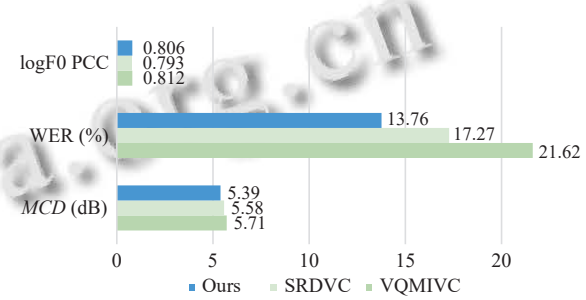


图 4 客观实验结果

在 *MCD* 对比实验中, MMIC-EVC 模型的 *MCD* 值仅为 5.39 dB, 相较于两个基线模型, MMIC-EVC 的 *MCD* 值分别减少 0.19 dB 和 0.32 dB, 结果表明 MMIC-EVC 模型的转换音频与目标语音之间的频谱距离相近, 失真较低, 可以有效提高转换音频的质量.

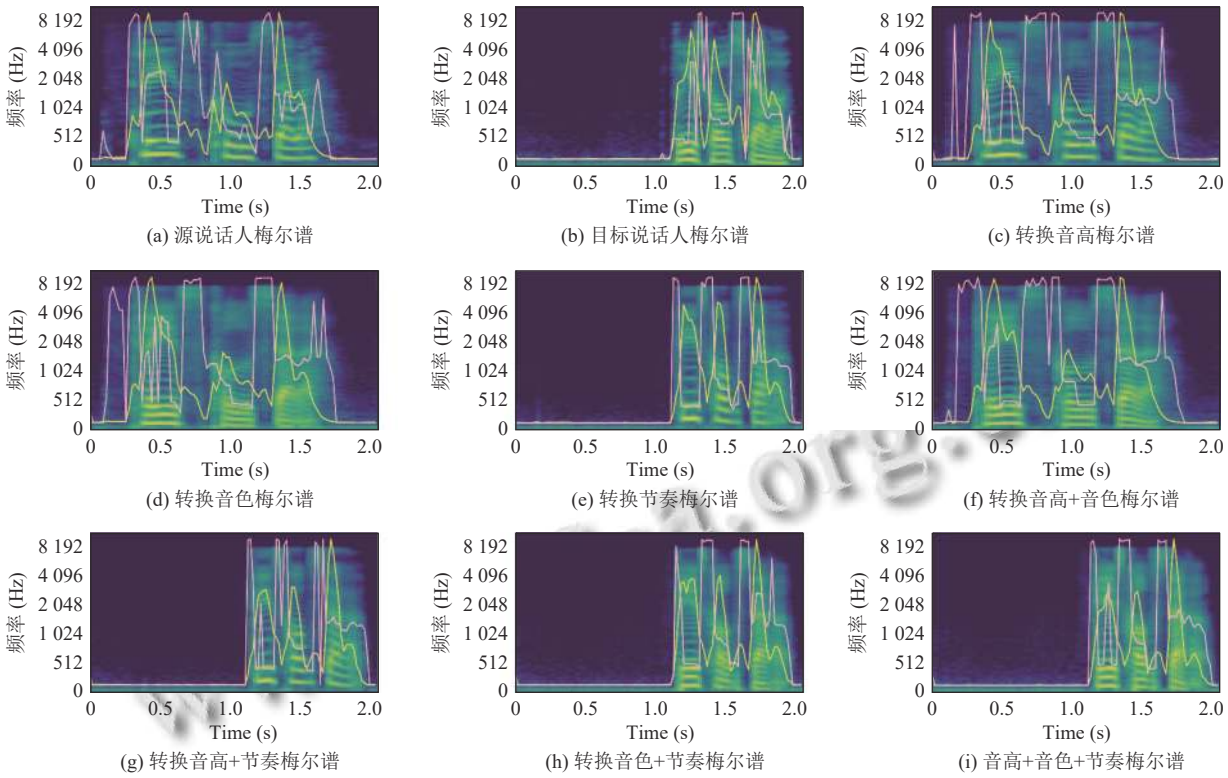
在 WER 测试中, MMIC-EVC 与 SRDVC 相比, 单词错误率下降 3.51%, 并远低于 VQMIVC 模型, 进一步证明 MMIC-EVC 能够对声学特征进行有效的解耦, 可以有效地保留源语音的内容信息.

此外, 在 logF0 PCC 测试中, MMIC-EVC 模型与 VQMIVC 拥有相似的性能, 都能有效地转换语音的情感特性, 但相较于 VQMIVC, MMIC-EVC 在数值上降低 0.006, 因为 VQMIVC 的基音信息直接馈送到解码器, 而未参与编码与解耦的过程.

MMIC-EVC 在多个方面均领先于两种基线模型. 实验表明 MMIC-EVC 模型特征解耦效果出色, 转换音频拥有较好的语音自然度, 同时能够有效地保持说话人的情感特性, 使得转换音频更加富有表现力.

3.3 频谱分析

图 5(a)–(i) 展示 MMIC-EVC 模型针对不同声学特征进行转换, 并将转换音频生成的梅尔谱图, 其中图 5(a) 是来自女性源说话人的梅尔谱图, 图 5(b) 是男性目标说话人的梅尔谱图, 横轴代表音频的时长, 纵轴代表音频的频率. 粉色为音高曲线, 黄色为能量曲线.



(a) 源说话人梅尔谱 (b) 目标说话人梅尔谱 (c) 转换音高梅尔谱

(d) 转换音色梅尔谱 (e) 转换节奏梅尔谱 (f) 转换音高+音色梅尔谱

(g) 转换音高+节奏梅尔谱 (h) 转换音色+节奏梅尔谱 (i) 音高+音色+节奏梅尔谱

图 5 不同声学特征的梅尔谱

在音高方面, 源说话人的音高曲线具有明显的尖锐特点, 而经过 MMIC-EVC 转换后的音频音高曲线与目标说话人音高曲线趋势相似, 均出现较为平缓的曲线, 符合源说话人与目标说话人的性别特性.

在节奏方面, 可以观察到源说话人的梅尔频谱图的峰值间隔相对较长, 而目标说话人的峰值间隔则较为短暂. 经过 MMIC-EVC 转换后的梅尔谱图显示, 其峰值间隔更接近于目标说话人. 同时, 源说话人的语速集中在整个梅尔谱图中, 而目标说话人的语速较快, 时长集中在梅尔谱图的后半段. 经过转换的音频符合这些特征. 能量曲线的变化趋势与节奏相关, 目标说话人的能量曲线前中端低后端高, 而源说话人的能量曲线为中间低两侧高, 转换音频的变化趋势与目标说话人相似.

在音色方面, 经过转换的语音在低频区域的条纹宽度和波动趋势更接近目标说话人, 表明在特定频率范围内的声学特征得到了有效的转换. 这表明转换后的语音能够很好地保留目标说话人的音色特征.

通过频谱分析, MMIC-EVC 模型能够在转换语音时成功解耦混合的声学特征, 并保持源语音的内容信息, 确保转换后的音频在表达意义和语境上保持一致. 同时学习目标说话人的音色、节奏和音调变化, 从而使得转换音频更加生动且富有的表现力.

3.4 消融实验

本节对 MMIC-EVC 模型中表现力模块和多重互信息损失进行消融实验, 验证该模块对模型性能的影响. 消融实验在自然-自然 (N-N) 场景中进行, w/o *Ep* 表示去除音高编码器, w/o *Er* 表示去除节奏编码器, w/o L_{SMI} 表示去除音色互信息损失, w/o L_{PMI} 表示去除音高互信息损失, w/o L_{CMI} 表示去除内容互信息损失, w/o L_{MI} 表示去除多重互信息损失, 结果如表 2 所示.

表 2 消融实验结果

方法	*MCD* (dB)
w/o *Ep*	6.12
w/o *Er*	5.87
w/o *Er*+*Ep*	**6.49**
w/o L_{SMI}	5.68
w/o L_{PMI}	5.63
w/o L_{CMI}	5.56
w/o L_{PMI}+L_{SMI}	5.87
w/o L_{PMI}+L_{CMI}	5.75
w/o L_{CMI}+L_{SMI}	5.82
w/o L_{MI}	**6.04**

在 *MCD* 对比实验中, 去除表现力模块后, *MCD* 达到 6.49 dB, 这表明了表现力模块在语音转换任务十分

重要, 能够扩大转换语音与真实目标语音之间的距离, 进而影响转换音频的质量.

另一方面, 去除多重互信息损失中任意一部分, *MCD* 均上升; 去除整体的多重互信息损失, 与 MMIC-EVC 相比, *MCD* 上升 0.65 dB. 表明该模块有助于进一步进行特征解耦, 提高模型的重构能力, 从而使得转换后的语音更加接近于真实目标语音.

综上所述, 表现力模块和多重互信息损失模块在 MMIC-EVC 模型中发挥了重要作用, 能够有效地影响模型的性能, 提高转换语音的质量和表现力.

4 结论与展望

本研究针对现有语音转换模型在转换音频表现力方面存在不足的问题, 提出 MMIC-EVC 模型. 实验结果表明, 精细的音调建模和节奏处理对于提升语音转换的表现力至关重要. MMIC-EVC 模型中引入表现力模块, 能够有效地传递语音的情感特性, 保持时长的合理性, 从而提升转换音频表现力. 同时将语音声学特征进行充分的解耦, 对于提高语音质量具有重要作用, MMIC-EVC 模型采用多重互信息约束, 进一步减少不同特征之间的相关性, 增加转换音频的自然度和真实感. 在未来的工作中, 我们将专注于进一步探索模型的解耦能力, 以实现对语音中关键声学特征的精确提取和独立操控, 重构出更加自然且富有表现力的语音, 为未来的人机交互真实场景带来更加出色的体验.

参考文献

1 杨帅, 乔凯, 陈健, 等. 语音合成及伪造、鉴伪技术综述. 计算机系统应用, 2022, 31(7): 12–22. [doi: 10.15888/j.cnki.csa.008641]

2 陈乐乐, 张雄伟, 孙蒙, 等. 融合梅尔谱增强与特征解耦的噪声鲁棒语音转换. 声学学报, 2023, 48(5): 1070–1080. [doi: 10.15949/j.cnki.0371-0025.2023.05.012]

3 Toda T, Black AW, Tokuda K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(8): 2222–2235. [doi: 10.1109/TASL.2007.907344]

4 Takashima R, Takiguchi T, Ariki Y. Exemplar-based voice conversion in noisy environment. Proceedings of the 2012 IEEE Spoken Language Technology Workshop. Miami: IEEE, 2012. 313–317. [doi: 10.1109/slt.2012.6424242]

5 Sisman B, Zhang MY, Li HZ. Group sparse representation with WaveNet vocoder adaptation for spectrum and prosody conversion. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(6): 1085–1097. [doi: 10.1109/TASLP.2019.2910637]

6 Xie FL, Soong FK, Li HF. A KL divergence and DNN-based approach to voice conversion without parallel training sentences. Proceedings of the 17th Annual Conference of the International Speech Communication Association. San Francisco: ISCA, 2016. 287–291. [doi: 10.21437/interspeech.2016-116]

7 Nakashika T, Takiguchi T, Ariki Y. High-order sequence modeling using speaker-dependent recurrent temporal restricted Boltzmann machines for voice conversion. Proceedings of the 15th Annual Conference of the International Speech Communication Association. Singapore: ISCA, 2014. 2278–2282. [doi: 10.21437/interspeech.2014-447]

8 Jafaryani M, Sheikhzadeh H, Pourahmadi V. Parallel voice conversion with limited training data using stochastic variational deep kernel learning. Engineering Applications of Artificial Intelligence, 2022, 115: 105279. [doi: 10.1016/j.engappai.2022.105279]

9 Bao F, Neumann M, Vu NT. CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz: ISCA, 2019. 2828–2832. [doi: 10.21437/interspeech.2019-2293]

10 Rizos G, Baird A, Elliott M, *et al*. Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020. 3502–3506. [doi: 10.1109/ICASSP40776.2020.9054579]

11 Wu DY, Chen YH, Lee HY. VQVC+: One-shot voice conversion by vector quantization and U-Net architecture. Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai: ISCA, 2020. 4691–4695. [doi: 10.21437/interspeech.2020-1443]

12 Chen YH, Wu DY, Wu TH, *et al*. Again-VC: A one-shot voice conversion using activation guidance and adaptive instance normalization. Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021. 5954–5958. [doi: 10.1109/icassp39728.2021.9414257]

13 Li T, Wang XS, Xie QC, *et al*. Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 1448–1460. [doi: 10.1109/taslp.2022.3164181]

14 Xie QC, Tian XH, Liu GH, *et al*. The multi-speaker multi-style voice cloning challenge 2021. Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021. 8613–8617. [doi: 10.1109/icassp39728.2021.9414001]

15 Tang HB, Zhang XL, Wang JZ, *et al*. QI-TTS: Questioning intonation control for emotional speech synthesis. Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: IEEE, 2023. 1–5. [doi: 10.1109/icassp49357.2023.10095623]

16 Cheng PY, Min MR, Shen DH, *et al*. Improving disentangled text representation learning with information-theoretic guidance. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Dan Jurafsky: ACL, 2020. 7530–7541. [doi: 10.18653/v1/2020.acl-main.673]

17 Yang SC, Tantrawenith M, Zhuang HL, *et al*. Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion. Proceedings of the 23rd Annual Conference of the International Speech Communication Association. Incheon: ISCA, 2022. 2553–2557. [doi: 10.21437/interspeech.2022-571]

18 Cheng PY, Hao WT, Dai SY, *et al*. CLUB: A contrastive log-ratio upper bound of mutual information. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 1779–1788.

19 Polyak A, Wolf L. Attention-based WaveNet autoencoder for universal voice conversion. Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton: IEEE, 2019. 6800–6804. [doi: 10.1109/icassp.2019.8682589]

20 Zhou K, Sisman B, Liu R, *et al*. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021. 920–924. [doi: 10.1109/icassp39728.2021.9413391]

21 Skerry-Ryan RJ, Battenberg E, Xiao Y, *et al*. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. Proceedings of the 35th International Conference on Machine Learning. Stockholm: PMLR, 2018. 4693–4702.

22 Kenter T, Wan V, Chan CA, *et al*. CHiVE: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 3331–3340.

23 Gan WD, Wen BL, Yan Y, *et al*. IQDUBBING: Prosody modeling based on discrete self-supervised speech representation for expressive voice conversion. arXiv:2201.00269, 2022.

24 Qian KZ, Zhang Y, Chang SY, *et al*. AutoVC: Zero-shot voice style transfer with only autoencoder loss. Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019. 5210–5219.

25 Qian KZ, Zhang Y, Chang SY, *et al*. Global prosody style transfer without text transcriptions. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 8650–8660.

26 Qian KZ, Zhang Y, Chang SY, *et al*. Unsupervised speech decomposition via triple information bottleneck. Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. 7836–7846.

27 Chan CH, Qian KZ, Zhang Y, *et al*. SpeechSplit2.0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022. 6332–6336. [doi: 10.1109/icassp43922.2022.9747763]

28 van den Oord A, Dieleman S, Zen H, *et al*. WaveNet: A generative model for raw audio. Proceedings of the 9th ISCA Speech Synthesis Workshop. Sunnyvale: ISCA, 2016. 125.

29 Zhou K, Sisman B, Liu R, *et al*. Emotional voice conversion: Theory, databases and ESD. Speech Communication, 2022, 137: 1–18. [doi: 10.1016/j.specom.2021.11.006]

30 Wang DS, Deng LQ, Yeung YT, *et al*. VQMIVC: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. Proceedings of the 22nd Annual Conference of the International Speech Communication Association. Brno: ISCA, 2021. 1344–1348. [doi: 10.21437/interspeech.2021-283]

31 Li CD, Shi J, Zhang WY, *et al*. ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration. Proceedings of the 2021 IEEE Spoken Language Technology Workshop. Shenzhen: IEEE, 2021. 785–792. [doi: 10.1109/slt48900.2021.9383615]

(校对责编: 张重毅)