Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(1), 166-176
Published Online February 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.131018

Hans汉斯

一种节奏与内容解纠缠的语音克隆模型

王　萌，姜　丹，曹少中

北京印刷学院信息工程学院，北京

收稿日期：2024年1月22日；录用日期：2024年2月23日；发布日期：2024年2月29日

摘　要

语音克隆是一种通过语音分析、说话人分类和语音编码等算法合成与参考语音非常相似的语音技术。为了增强说话人个人发音特征转移情况，提出了节奏与内容解纠缠的MRCD模型。通过节奏随机扰动模块的随机阈值重采样将语音信号所传递的节奏信息解纠缠，使语音节奏相互独立；利用梅尔内容增强模块获取说话人的相似发言特征内容，同时增加风格损失函数及循环一致性损失函数衡量生成的语音与源语音的谱图及说话人身份之间差异，最后用端到端的语音合成模型FastSpeech2进行语音克隆。为了进行实验评估，将该方法应用于公开的AISHELL3数据集进行语音转换任务。通过客观和主观评价指标对该模型进行评估，结果表明，转换后的语音在保持自然度得分的同时，在说话人相似度方面优于之前的方法。

关键词

语音克隆，零样本，扬声器表示，内容增强

A Voice Cloning Model for Rhythm and Content De-Entanglement

Meng Wang, Dan Jiang, Shaozhong Cao

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: Jan. 22nd, 2024; accepted: Feb. 23rd, 2024; published: Feb. 29th, 2024

Abstract

Voice cloning is a technique for synthesizing speech that closely resembles a reference speech through algorithms such as speech analysis, speaker classification, and voice coding. To improve the transfer of individual speaker articulatory features, the MRCD model with rhythm and content de-entanglement is proposed. The rhythmic information carried by the speech signal is de-entangled by the random threshold resampling of the rhythmic random perturbation module, so that the speech rhythms are independent of each other; the content of the speaker's similar speech fea-

文章引用: 王萌, 姜丹, 曹少中. 一种节奏与内容解纠缠的语音克隆模型[J]. 人工智能与机器人研究, 2024, 13(1): 166-176. DOI: 10.12677/airr.2024.131018

tures is obtained by using the Meier content enhancement module, and at the same time the stylistic and cyclic consistency loss functions are added to measure the differences between the generated speech and the spectrograms of the source speech and the speaker's identity, and then finally the speaker is identified by an end-to-end speech synthesis model, FastSpeech2. Finally, an end-to-end speech synthesis model, FastSpeech2, is used for speech cloning. For experimental evaluation, the method was applied to the publicly available AISHELL3 dataset for the speech cloning task. The model is evaluated using objective and subjective evaluation metrics, and the results show that the converted speech outperforms the previous method in terms of speaker similarity while maintaining the naturalness score.

Keywords

Voice Cloning, Zero-Shot, Speaker Representation, Content Enhance

1. 引言

随着现代电子通讯技术的快速发展，语音交互成为机器人人机交互最直接高效、通俗易懂的方式。语音合成(Text-to-Speech, TTS) [1]是一种基于文本生成相应语音的技术，是人机交互的核心技术之一。在深度学习技术兴起之前，传统的语音合成方法包括拼接语音合成[2]和统计参数[3]语音合成方法。近年来，随着深度学习技术的迅猛发展，语音合成在自然度和可懂度方面取得了显著进展。当前的神经语音合成模型主要由声学模型和声码器模块级联组成，被称为两阶段语音合成方法。声学模型根据输入的文本预测合成语音的声学特征，代表性的模型有 Tacotron2 [4]等。声码器模块将声学特征转换成语音波形，典型的模型包括有 WaveNet [5]，WaveGlow [6]，HiFi-GAN [7]等。然而，两阶段语音合成方法由于分开训练模块情况，存在信息传递的不匹配与失真问题。为了解决这一问题，近年来研究者们逐渐倾向于采用端到端的语音合成方法，例如 FastSpeech2 [8]，一种采用非自回归方式的端到端语音合成模型。

语音克隆任务是语音合成领域的一个重要任务之一。语音克隆任务是一项旨在通过训练模型，使其能够生成与特定说话人相似的语音的任务。该任务可以被应用于多种场景和应用，例如恢复失去声音的用户自然交流能力，或者定制智能交互环境中的数字助手 Siri 等。在最近的文献中，现有的语音合成系统可以将目标说话人的语音风格、语调、音色等发音特征捕捉并转移到生成的语音中，例如 Attentron [9]提出了一种利用基于注意力的嵌入变长方法特性以接近原始参考语音，从而实现更好的泛化；Xiao 等人[10]专门为网络设计了六个损失函数以衡量说话人风格转移情况；CDFSE 模型[11]从参考语音中提取相应的局部内容嵌入和细粒度说话人嵌入来建模个人发音特征。

遗憾的是，现有模型中的大多数往往难以转移个人音素发音特点相关内容的明确意义，以及难以捕捉独立语音信息，这使得合成的语音本质上是单一重复的。针对以上问题，本文引入了不同的特征处理方式，并使用了一种语音内容增强模块。此外，使用了一种不同的策略来调节语音合成与之前未观察到的说话者的声音，并比较了几种说话者编码器模型的神经结构。具体来说，本文提出了一个模型 MRCD (Model with Rhythm and Content Disentanglement)，(1) 首先设计节奏随机扰动模块，引入一个风格编码器将语音单独分解出节奏、音高，利用经典语音特征梅尔倒谱系数 MFCC，按照时间维度进行语音信息扰乱；(2) 设计语音梅尔内容增强模块，结合参考注意力机制[12]指导细粒度说话人特征生成。将学习参考语音中梅尔谱图

信息内容，改进了说话人个体发音特征的内容转移情况，使语音解码器生成更接近参考说话人的语音。(3) 设计两个损失函数，以控制对风格和说话人特征的解耦，作为解码器的条件来迁移语音的风格。

主要创新点是：(1) 通过引入风格编码器，将语音中独立提取的节奏和音高信息与文本信息以及音频信息融合。这使得生成的音色风格更加接近参考说话人，提高了语音合成的音质和自然度；(2) 采用两种解纠缠技术，其中一种增加了参考说话人自身发音节奏信息，另一种增加了参考语音与内容的相关性。这些技术的应用能够保证生成语音的质量，使其更适用于实际应用场景。这种方法可以有效提升语音合成系统在应对不同说话者和内容情境时的适应性和表现效果。

2. 相关工作

传统的语音克隆技术[12]通常采用自适应方式进行训练，不仅存在处理效率低问题，而且获取目标说话人的语音样本难度大的问题。目前主流的语音克隆技术通常采用说话人编码方法[13]，该方法可以直接从目标说话人的语音样本中提取说话人的嵌入向量，克隆速度快。例如 Li 等人[14]尝试通过引入关注机制，以捕捉更多来自演讲的信息，从而使扬声器嵌入具有更细致的纹理。

在获取语音信息方面，当前研究主要集中在纯净语音生成的解缠、控制和去噪等方面[15]，例如，Fang 等人[16]采用增强的词嵌入及文本预处理来改进文本的表示学习；AUTOPST [17]模型通过自动编码器从语音中提取韵律信息；Gibiansky 等人[18]的多扬声器 TTS 系统通过扬声器查找表或扬声器编码器显式地建模扬声器表示。虽然神经方法显著提高了生成语音的质量，但仍有很大的空间进行进一步的改进。

本文目标是模拟、控制和转移言语的风格、韵律，并学习良好的文本和语音特征以产生表达性的言语，参考 AUTOPST [17]中的风格编码器中随机阈值重采样技术，设计了节奏随机扰动模块。同时为利用特定不可见说话者的语音的发音特征，设计梅尔内容编码器，结合 CDFSE [11]模型中的参考注意力机制获取细粒度说话人表征，以更好地将参考语音中的语音内容信息应用到模型生成的结果中。

2.1. 随机阈值重采样

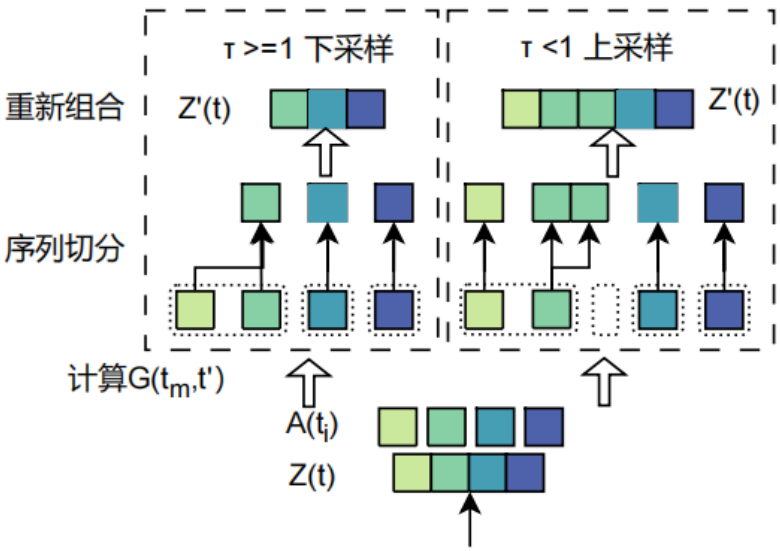

Figure 1. Random resampling. Squares with different colors represent sequences of frames with different similarities
图 1. 随机阈值重采样结构。不同颜色的方块表示具有不同相似度的帧序列

语音信号表示为声谱图 $X(t)$，其中 t 表示帧索引。用 S 表示包含在语音中的音素符号向量。将 b 表

示为S中每个音素符号重复次数的向量，也将其称为节奏信息。随机阈值重采样的任务是改变音素的重复b，更好地保留语音的身份信息。如图1所示，$Z(t)$是输入的13维MFCC特征(梅尔频率倒谱系数)经过风格编码器后输出的特征序列，随机阈值重采样过程如下。

首先将特征序列$Z(t)$分成连续的片段$A(t)$，利用Gram矩阵来记录相邻帧之间的相似性。

$$G(t,t')=\frac{A^T(t)A(t')}{\|A(t)\|_2\|A(t')\|_2} \tag{1}$$

其中t表示当前片段索引，t'表示$Z(t)$其他片段索引。

设置一个随机阈值，目的是在分割序列时，确定t_m是否是下一段序列的边界，分为上采样和下采样两部分。当随机阈值$\tau(t)<1$时，根据公式(3)将$G(t_m,t')<\tau(t)$下的序列划分。划分后子序列通过平均池化成独立编码，例如公式2。

$$\tilde{Z}(m)=meanpool\left(Z\left(t_m:t_{m+1}+1\right)\right) \tag{2}$$

$$\forall t\in[t:t+1],G(t_m,t')\le\tau(t) \tag{3}$$

$$\forall t'\in[t:t+1],G(t_m,t')\ge1-\tau(t) \tag{4}$$

当随机阈值$\tau(t)\ge1$时，根据公式(4)将$G(t_m,t')\ge1-\tau(t)$下的序列划分。在此t_m，t_{m+1}后添加一个t_{m+2}空白序列，在平均池化简化成独立编码时t_{m+2}序列复制左边t_{m+1}段编码，作为韵律风格编码器输出。

Table 1. Example of correlation between references speech and input text
表1. 参考语音内容与输入文本相关性示例

	音素
参考语音内容	“zh ang1 x iou4 c ai2 y ie3 m ei2 y iou3 d uo1 x iang3 sp”
输入文本	“b ai3 d u4 y iong1 y iou3 sh u4 w uan4 m ing2 y ian2 f a1 g ong1 ch eng2 sp”
相似度较高的内容	“sp”与“sp”、“ai2”与“ai3”

2.2. 细粒度说话人嵌入特征

在零样本语音克隆任务中，应充分利用说话人表征和语音发音特征。在语音发音特征中，理想状态是在语音发音特征中，理想状态是直接使用参考语音中的音素段落。例如表1。

首先，通过预处理网络提取帧级特征，使用内容编码器用于将这些特征转化成更具信息量的帧级内容嵌入。将帧级内容嵌入传递给下采样内容编码器，将原始的音素序列传递给下采样说话人编码器中。这两个下采样编码器包括4个一维卷积层和一个256维的全连接层，形成一个层次化的特征提取结构。同时通过自注意力机制构建了参考注意模块。该模块以参考语音的音素编码器输出为查询(Q)，而来自参考语音的局部内容嵌入作为键(K)，从而引导细粒度说话人嵌入的生成。

3. 基于FastSpeech2改进的MRCD模型及网络

为了验证本文提出的两种解纠缠技术，结合FastSpeech2模型搭建了语音克隆系统MRCD模型，MRCD模型网络结构如图2所示。模型引入节奏风格扰乱模块，用于获取音频的MFCC特征嵌入进行建模。同时，采用带有语音梅尔内容增强模块的参考注意编码器，以获得与内容相关的细粒度说话人嵌入。将这两个编码器的输出添加到音素编码器的输出中，并传递给FastSpeech 2的方差适配器，以生成与参考语音中的说话人相同的语音。

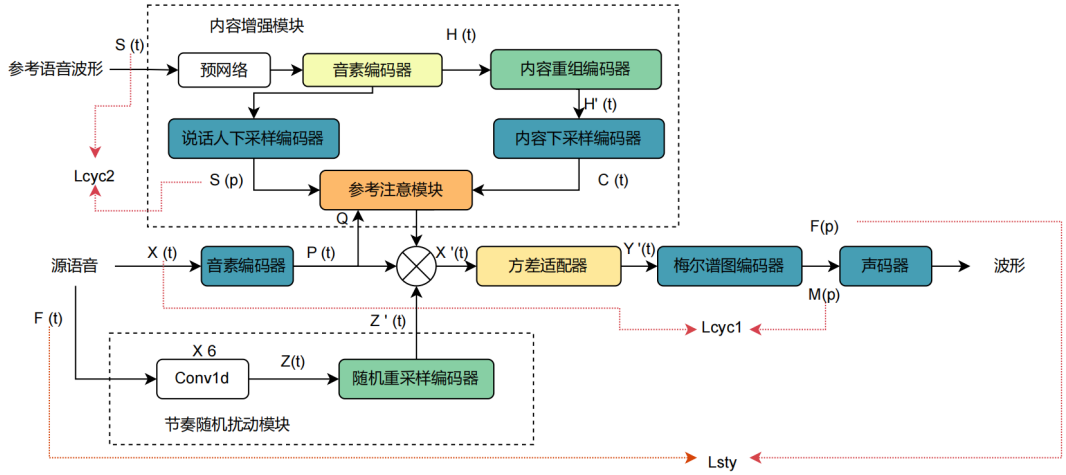

Figure 2. The overall architecture of the MRCD model
图 2. MRCD 模型结构

3.1. 节奏随机扰动模块

在 FastSpeech2 模型中，增加 MFCC 特征处理主要是为了有效提高系统的识别性能，MFCC 是一种常用于语音信号处理的特征表示方法，可以有效地捕捉到语音信号的频谱特性和时域特性。本文 MRCD 模型将对语音中获取的 80 维 MFCC 特征相似度较高的片段采用了随机阈值重采样方法，通过对语音信号的时序扰动提供更多信息。

节奏随机扰动模块首先将源语音数据通过卷积网络转换成 MFCC 特征 $Z(t)$。通过随机阈值重采样将相邻帧之间的相似度较高的片段重新排列，从而生成一个新的时间序列 $Z'(t)$，以改变其节奏和韵律。根据随机重采样的结果，节奏随机扰动模块将按照缩短和延长分为上采样和下采样两部分。上采样是指将 MFCC 特征 $Z(t)$ 中相似音素信息 S 的帧数 b 增加，以加快其节奏；而下采样则是指将 MFCC 特征 $Z(t)$ 中相似音素信息 S 的帧数 b 减少，以减慢其节奏。通过上采样和下采样的操作，可以对源语音数据的节奏进行随机扰动，从而生成具有不同节奏和韵律的新语音样本。

最终将随机重采样编码器的输出，添加到 FastSpeech2 编码器输出中重构语音。

3.2. 说话人发音特征转移与建模

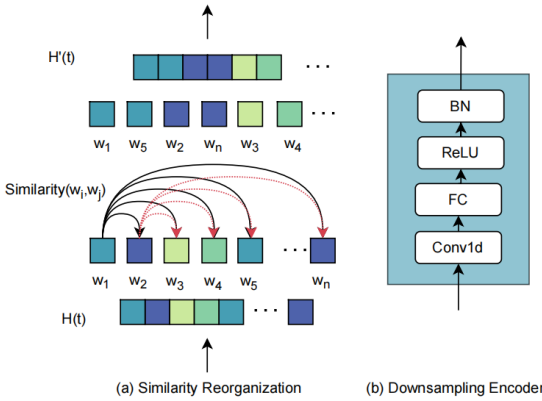

Figure 3. (a) Similarity recombination in the content enhancement module. (b) Speaker embedding/content downsample encoder
图 3. (a) 内容增强模块中的相似性重组。(b) 说话人嵌入/内容下采样编码器

为了更好打乱参考语音中音素之间的关系，提高参考语音和输入文本之间的内容相关性，约束内容编码器对内容信息的编码，MRCD 模型采用语音梅尔内容增强器将具有相似发音内容和关系的单词或子序列放在一起，放大说话人发音特征。如图 3 所示，将输入语音的音素序列 $H(t)$ 进行重新排序和补充。其中 $H(t)$ 包含 n 个词或字符，表示为 $H(t)=[w_1,w_2,\cdots,w_n]$。每个子序列 w_i 对应的嵌入表示为 $E(w_i)$。使用余弦相似度来计算每个子序列的相似性，对于任意两个子序列 w_i、w_j 的相似性定义为：$similarity(w_i,w_j)=\frac{E(w_i)\cdot E(w_j)}{\|E(w_i)\|\|E(w_j)\|}$。其中 $E(w_i)\cdot E(w_j)$ 是点积，$\|E(w_i)\|$ 和 $\|E(w_j)\|$ 分别是子序列的长度。随后，基于计算的相似性结论，在 w_i 子序列后选择相似度最高的子序列 w_j，例如：$H'(t)=[w_1,w_3,w_2,w_6,\cdots,w_n]$。这样，输入序列就完成了内容增强操作。

3.3. 风格损失函数和循环一致性损失函数

语音转换任务中如何进行风格转移是至关重要的，这将影响说话人的表现。为了解决这个问题，同时约束特征模块学习，本文引入余弦相似度风格损失函数来测量风格失真，约束并指导模型的训练过程。在学习风格表示 Z_t 时，强制这种风格表示 Z_s 和目标风格 Z_t 之间的失真最小化。

$$L_{sty}=\frac{\sum_{i=1}^{n}(z_s z_t)}{\sqrt{\sum_{i=1}^{n}(z_s)^2}\sqrt{\sum_{i=1}^{n}(z_t)^2}} \tag{5}$$

当然，余弦风格损失函数 L_{sty} 只能约束生成的话语和目标风格保持不变，但是不能保证源说话人的身份不变，为了保证源说话人身份完整，本文参考了循环一致性损失函数[19]设计了 L_{cyc}。在这里循环一致性损失函数首先计算了生成语音和源语音梅尔谱图之间 M_P、M_S 的 L1 损失，同时计算生成语音和目标语音的说话人信息 S_P、S_t 之间的 L1 损失。

$$L_{cyc}=\frac{\sum_{i=1}^{n}\left|f(M_p)-M_s\right|}{n}+\frac{\sum_{i=1}^{n}\left|f(S_p)-S_t\right|}{n} \tag{6}$$

最终通过联合训练将上述损失函数加权值最小。

$$L=L_{\text{pitch}}+L_{\text{energy}}+L_{\text{duration}}+\alpha L_{\text{phone}}+\beta L_{\text{sty}}+\chi L_{\text{cyc}} \tag{7}$$

其中 L_{pitch}、L_{energy}、$L_{duration}$、L_{phone}、L_{sty}、L_{cyc} 分别是 pitch 损失函数，能量损失函数，持续时间损失函数，音素分类器损失函数、风格损失函数、循环一致性损失函数。

4. 实验结果和分析

4.1. 实验设置

数据集。AISHELL-3 数据集[20]是希尔贝壳公司创建的中文普通话语音数据库。数据集包含 218 名来自中国不同口音区域的发言人参与录制。本实验将 218 名说话者的话语，95%的话语用于训练，5%用于验证。同时选取了 7 名未在数据集中出现的说话人的日常话语录音，以便在训练后使用它们来评估模型在未见过的源说话人上的表现。

训练设置。在训练之前，将数据集中的所有语音以 22.05 kHz 的采样率将波形转换为 80 维梅尔谱图。帧大小为 1024，跳数为 256。同时本实验原始文本通过汉语字形音素转换工具包转换为拼音声母和韵母组成的音素序列[21]。在 NVIDIA RTX3090 GPU 上训练所有模型，迭代次数 250 K，批处理大小为 16。

采用 Adam 优化器，$\beta 1 = 0.9$，$\beta 2 = 0.98$，$\varepsilon = 10 - 9$。在 4000 次迭代前采用热身策略。使用训练良好的 HiFi-GAN [7]作为神经声码器来生成波形。

4.2. 基线模型

CDFSE [11]：是一个零样本说话人自适应模型，提出了一种内容相关的细粒度说话人表征方法。以直接迁移目标说话人个人发音习惯，来提高和改善合成语音的质量。

Attentron [9]：是一个少样本语音合成模型用于克隆在训练过程中未出现的说话人的声音，提出了一种基于注意力的微调嵌入方法。此外，模型还能扩展到任意长度的参考音频以此来改善合成语音的质量。

4.3. 对比实验

实验采用语音克隆方法通用的主观评价[9]：平均主观意见得分(MOS)和相似度平均主观意见得分(SMOS)来比较不同模型对训练中已见说话人和不可见说话人的语音合成质量和音色相似度。实验选择的 8 个未见的说话者和从训练集中随机选择的 7 个已见的说话者作为参考声音。文本句子来自测试集，长度和内容各不相同。对于每个说话人，只使用一个话语作为参考语音来指导语音合成。由 15 名精通中文的听众通过耳机试听给出主观评分，根据语音的质量和参考语音音色的相似度进行打分。MOS 和 SMOS 评分如表 2 所示。

4.3.1. 语音自然度

根据表 2 的结果，MRCD 模型在自然度方面表现良好，优于两个基线模型。具体来说，对于已见的说话人，MRCD 模型的 MOS 为 3.25，而对于不可见的说话人，MOS 达到 3.73。值得注意的是，MRCD 模型对不可见说话人的改进尤为显著，它实现了 0.3 点以上的差距。这进一步说明了利用 MFCC 特征删除节奏信息对于提高合成语音的自然度是非常有用的。

Table 2. MOS, SMOS, and speaking similarity scores for different models
表 2. 不同模型的 MOS、SMOS、说话相似度分数

类别	模型	MOS	SMOS	说话人相似度(↑)
已见说话人	CDFSE [11]	3.21	3.36	77.19
	Attentron [9]	3.16	3.30	75.00
	MRCD	**3.25**	**3.51**	**84.6**
不可见说话人	CDFSE [11]	3.43	3.36	78.00
	Attentron [9]	3.40	3.30	75.40
	MRCD	**3.73**	**3.72**	**80.5**

4.3.2. 说话人相似度

在说话人相似度方面，同时为了评估说话人的相似度，采用说话人验证系统[22]来提取话语级说话人向量，并计算合成语音与真实语音说话人之间的余弦相似度。根据表 2 的结果，MRCD 模型的 SMOS 对于已见说话人为 3.51，对于不可见说话者为 3.72，优于两种基线模型。在客观评价中，MRCD 模型生成的语音与真实语音的相似度在可见说话人类别上高于基线模型 2.5 个百分点以上，在不可见说话人类别上高于基线模型 6 个百分点以上。

4.3.3. 风格相似度

实验采用风格 ABX 偏好测试进行评估迁移后语音风格相似度。ABX 偏好测试中受试者要求选择转换后的话语 A 与 B (两种方法)听起来更像目标说话者的真实语音 X 或者没有偏好。每一对 A 与 B 都是被

打乱避免优先偏向。实验主要比较了 MRCD 模型与 CDFSE 模型，结果如图 4 所示。对于可见说话人，MRCD 模型的选择频率比 CDFSE 模型高 3 个百分点，对于不可见说话人，MRCD 模型的优势更为显著，选择频率比 CDFSE 模型高 12 个百分点。

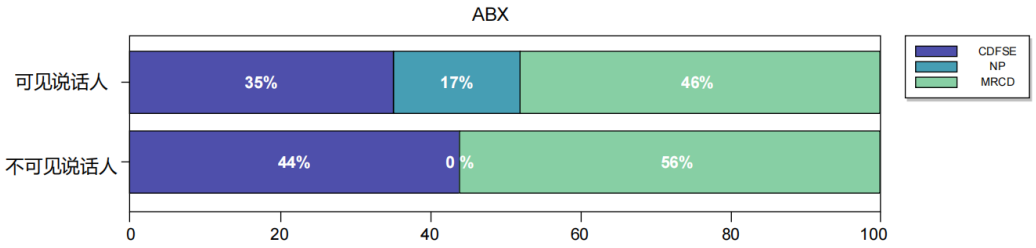

Figure 4. ABX preference test results of the CDFSE model and the MRCD model
图 4. CDFSE 模型和 MRCD 模型的 ABX 偏好测试结果

同时利用人工中文标注方法，对每个文字所需时间进行划分。输入文本与参考语音内容相同，对比生成语音和参考语音的每个文字对应音素时间之差。实验将生成语音和真实参考语音的对应文字对齐的音素时间差总和 $\sum_{i=1}^{n}|x_i - y_i|$ 进行对比。

其中 x_i 是模型中生成语音单个文字对齐因素对应的时间，y_i 是参考语音中的每个文字对齐音素对应的时间。根据图 5(a)显示，MRCD 模型的音素时间差为 0.38，CDFSE 模型的音素时间差为 0.43。根据图 5(b)，MRCD 模型的音素时间差为 0.22，CDFSE 模型的音素时间差为 0.26。由此结果显示，MRCD 模型的音素时间差明显低于 CDFSE 模型，这表明 MRCD 模型在音素级别上更接近参考语音。

4.4. 消融实验

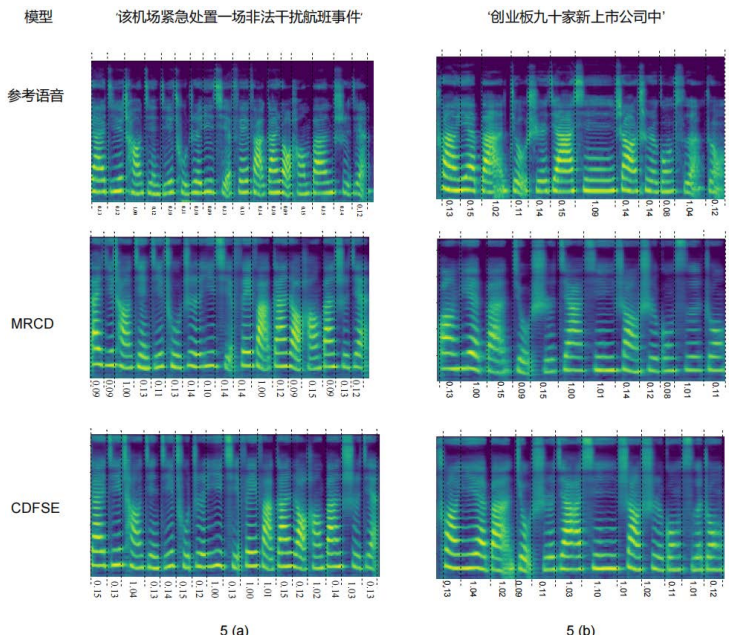

Figure 5. Comparison of text corresponding phoneme time between CDFSE model and MRCD model. The time marked on the abscissa is the time corresponding to each phoneme of the text
图 5. CDFSE 模型和 MRCD 模型的文本对应音素时间对比。横坐标标注时间是每个文字音素对应的时间

为了验证不同模块对于已见和不可见语音说话人特征转移的有效性，进行了一系列实验，并分别考察了不同模块的影响，包括 BM + RR、BM + RR + CE、BM + RR + CE + L_{sty}、以及最终的模型 MRCD = BM + RR + CE + L_{sty} + L_{cyc}，其中 RR 是节奏随机扰动模块，CE 是梅尔内容增强模块。通过主观和客观评分来衡量生成语音的自然度和说话人相似度。

根据表 3 结果可知，对于可见说话人，在说话人验证系统说话人相似度方面，BM 模型的分值是 77.19%，增加 RR 模块的分值是 75.09%略有下降，但在此基础上增加 CE 模块、L_{sty}、L_{cyc} 损失函数后说话人相似度分数为 80.7，有超过 2 个百分点的提升。对于不可见说话人，在语音自然度和主观判断说话人相似度上，加入 RR 模块在自然度和说话人相似度都有积极影响，其他指标未有明显提升。根据整体模型和实验结果，可以看出在方差适配器前期增加了多个特征学习模块，导致预测值与真实值之间存在较大的差异。因此，需要引入了循环一致性损失函数和风格损失函数来解决这个问题。

Table 3. MOS, SMOS with increasing different modules, and speaker similarity
表 3. 增加不同模块的 MOS、SMOS，说话人相似度

类别	模型	MOS	SMOS	说话人相似度(↑)
可见说话人	BM	3.21	3.36	77.19
	BM + RR	3.11	3.30	75.09
	BM + RR + CE	3.11	3.32	**80.70**
不可见说话人	BM	3.43	3.36	78.00
	BM + RR	3.54	3.41	78.95
	BM + RR + CE	**3.50**	**3.60**	76.34

Table 4. MOS, SMOS with increasing loss function, and speaker similarity
表 4. 增加损失函数的 MOS、SMOS，说话人相似度

类别	模型	MOS	SMOS	说话人相似度(↑)
可见说话人	BM	3.21	3.36	77.19
	BM + RR + CE + L_{sty}	3.00	3.26	81.9
	BM + RR + CE + L_{sty} + L_{cyc}	**3.25**	**3.51**	**84.6**
不可见说话人	BM	3.43	3.36	78.00
	BM + RR + CE + L_{sty}	3.43	3.68	80.5
	BM + RR + CE + L_{sty} + L_{cyc}	**3.73**	**3.72**	**80.5**

根据表 4 结果可知，在 MOS、SMOS、说话人验证系统说话人相似度方面，加入 L_{sty}、L_{cyc} 损失函数后，表现出显著优势。其中可见说话人的说话人相似性及和不可见说话人的 MOS、SMOS 提升效果最为明显。

4.5. 结论

表 2 实验结果显示，在自然度的 MOS 对比结果中可以看出与 CDFSE 模型、Attentron 模型相比，对于未见的说话者，MRCD 模型在 MOS 上的改进较为为显著，差距为 0.3。时间节奏相似度上，MRCD 模型的音素时间差明显低于 CDFSE 模型。这些结果表明在节奏随机扰动模块上对于 MFCC 特征的有效提取，有利于 MRCD 模型更好的学习参考语音中的特征信息在语音合成任务中发挥了关键作用，对于模型性能的改进非常重要。

在说话人相似度方面，SMOS 结果表明本文提出的 MRCD 模型在说话人相似度方面优于两个基线。MRCD 对可见说话人的 SMOS 为 3.51，对不可见说话人的 SMOS 为 3.72。对于未见的说话者，MRCD

模型在 SMOS 上的改进更为显著，差距超过 0.3。同时采用客观说话人相似度对比，MRCD 模型在可见说话人任务表现中优于其他模型。在 ABX 偏好中，受试者在基线模型 CDFSE 与设计的 MRCD 模型中，更偏向于设计的模型。这些结果说明 CE 模块可能通过增强梅尔频谱改进说话人特征提取，从而提高了说话人相似度使生成的语音更好地匹配参考语音的声学特征，使生成的语音更接近参考语音的说话人特征，从而提高说话人相似度。

在最后的消融实验表明，增加节奏随机扰动模块、语音内容增强模块对模型性能产生了正向作用。这可能是因为节奏随机扰动模块有助于模型更好地捕获语音的节奏和韵律，从而使生成的语音更加自然和流畅。CE 模块可能通过梅尔频谱内容聚集说话人的发音特征，有助于生成语音更好地匹配参考语音的声学特征。实验中加入不同损失函数的结果显示，这些损失函数同样起到了增益作用。这表明引入损失函数可以提高模型的训练稳定性和性能。特别是在可见说话人情况下，所有损失函数的组合效果最好，这可能是因为它们能够有效地引导模型学习正确的声学和说话人特征。结合以上实验结果，最优模型选择了具有所有模块和损失函数的模型作为最终的 MRCD 模型，该模型在自然度、说话人相似度和风格相似度方面表现最佳。

5. 结论

本文提出了结合节奏和内容解缠的 MRCD 模型用于语音转换任务，它提高了生成语音的自然度和与真实语音的说话人相似度。该方法除了利用梅尔谱图外还利用了 MFCC 特征增强时间依赖性用以节奏信息随机扰乱。同时设计了语音内容增强模块，将聚集说话人发音特征更加有效的迁移说话人的发音特征。此外增加了风格损失函数和循环一致性损失函数约束 MRCD 模型训练过程。进行了大量的实验来研究所提出的模块的有效性，实验分析表明，该方法具有模拟个人语音特征的能力，在语音自然度、说话人相似度和风格相似度方面表现最佳。

基金项目

北京市自然基金项目－北京市教委科技计划重点项目(KZ202010015021)，专业学位研究生联合培养基地建设－电子信息(21090223001)，北京印刷学院博士启动金(27170123036)。

参考文献

[1] Sproat, R.W. and Olive, J.P. (1995) Text-to-Speech Synthesis. *AT&T Technical Journal*, **74**, 35-44. https://doi.org/10.1002/j.1538-7305.1995.tb00399.x

[2] Olive, J.P. (1977) Rule Synthesis of Speech from Dyadic Units. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hartford, 9-11 May 1977, 568-570. https://doi.org/10.1109/ICASSP.1977.1170350

[3] Zen, H., Tokuda, K. and Black, A.W. (2009) Statistical Parametric Speech Synthesis. *Speech Communication*, **51**, 1039-1064. https://doi.org/10.1016/j.specom.2009.04.004

[4] Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z. and Wu, Y. (2018) Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Calgary, 15-20 April 2018, 4779-4783. https://doi.org/10.1109/ICASSP.2018.8461368

[5] Wu, Y.C., Hayashi, T., Tobing, P.L., Kobayashi, K. and Toda, T. (2021) Quasi-Periodic WaveNet: An Autoregressive Raw Waveform Generative Model with Pitch-Dependent Dilated Convolution Neural Network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29**, 1134-1148. https://doi.org/10.1109/TASLP.2021.3061245

[6] Prenger, R., Valle, R. and Catanzaro, B. (2019) Waveglow: A Flow-Based Generative Network for Speech Synthesis. *ICASSP* 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 3617-3621. https://doi.org/10.1109/ICASSP.2019.8683143

[7] Kong, J., Kim, J. and Bae, J. (2020) Hifi-gan: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Advances in Neural Information Processing Systems*, Vol. 33, 17022-17033.

[8] Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T.Y. (2020) Fastspeech 2: Fast and High-Quality

End-to-End Text to Speech.

[9] Choi, S., Han, S., Kim, D. and Ha, S. (2020) Attentron: Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding. *Proceedings Interspeech* 2020, Shanghai, 25-29 October 2020, 2007-2011. https://doi.org/10.21437/Interspeech.2020-2096

[10] An, X., Soong, F.K. and Xie, L. (2022) Disentangling Style and Speaker Attributes for TTS Style Transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **30**, 646-658. https://doi.org/10.1109/TASLP.2022.3145297

[11] Zhou, Y., Song, C., Li, X., Zhang, L., Wu, Z., Bian, Y. and Meng, H. (2022) Content-Dependent Fine-Grained Speaker Embedding for Zero-Shot Speaker Adaptation in Text-to-Speech Synthesis. *Proceedings Interspeech* 2022, Incheon, 8-22 September 2022, 2573-2577. https://doi.org/10.21437/Interspeech.2022-10054

[12] Miao, Y. and Metze, F. (2015) On Speaker Adaptation of Long Short-Term Memory Recurrent Neural Networks. In 16*th Annual Conference of the International Speech Communication Association*, Dresden, 6-10 September 2015, 1101-1105. https://doi.org/10.21437/Interspeech.2015-290

[13] Cooper, E., Lai, C.I., Yasuda, Y., Fang, F., Wang, X., Chen, N. and Yamagishi, J. (2020) Zero-Shot Multi-Speaker Text-to-Speech with State-of-the-Art Neural Speaker Embeddings. *ICASSP* 2020-2020 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Barcelona, 4-8 May 2020, 6184-6188. https://doi.org/10.1109/ICASSP40776.2020.9054535

[14] Li, X., Song, C., Li, J., Wu, Z., Jia, J. and Meng, H. (2021) Towards Multi-Scale Style Control for Expressive Speech Synthesis. *Proceedings Interspeech* 2021, Brno, 30 August-3 September 2021, 4673-4677. https://doi.org/10.21437/Interspeech.2021-947

[15] Hsu, W.N., Zhang, Y., Weiss, R.J., *et al*. (2019) Disentangling Correlated Speaker and Noise for Speech Synthesis via Data Augmentation and Adversarial Factorization. *ICASSP* 2019-2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 5901-5905. https://doi.org/10.1109/ICASSP.2019.8683561

[16] Fang, W., Chung, Y.A. and Glass, J. (2019) Towards Transfer Learning for End-to-End Speech Synthesis from Deep Pre-Trained Language Models.

[17] Qian, K., Zhang, Y., Chang, S., Xiong, J., Gan, C., Cox, D. and Hasegawa-Johnson, M. (2021) Global Prosody Style Transfer without Text Transcriptions. *Proceedings of Machine Learning Research*, **139**, 8650-8660.

[18] Gibiansky, A., Arik, S., Diamos, G., Miller, J., Peng, K., Ping, W. and Zhou, Y. (2017) Deep Voice 2: Multi-Speaker Neural Text-to-Speech. *NIPS*'17: *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 3-8 December 2018, 2966-2974.

[19] Xue, L., Pan, S., He, L., Xie, L. and Soong, F.K. (2021) Cycle Consistent Network for End-to-End Style Transfer TTS Training. *Neural Networks*, **140**, 223-236. https://doi.org/10.1016/j.neunet.2021.03.005

[20] Shi, Y., Bu, H., Xu, X., Zhang, S. and Li, M. (2020) Aishell-3: A Multi-Speaker Mandarin TTS Corpus and the Baselines. *Proceedings Interspeech* 2021, Brno, 30 August-3 September 2021, 2756-2760. https://doi.org/10.21437/Interspeech.2021-755

[21] Pypinyin. https://pypi.org/project/pypinyin

[22] Wan, L., Wang, Q., Papir, A. and Moreno, I.L. (2018) Generalized End-to-End Loss for Speaker Verification. 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Calgary, 15-20 April 2018, 4879-4883. https://doi.org/10.1109/ICASSP.2018.8462665