文章编号：1003-0530(2023)04-0719-11

基于音色一致的语音克隆说话人特征提取方法

李嘉欣 张连海 李宜亭
（中国人民解放军战略支援部队信息工程大学信息系统工程学院，河南郑州 450001）

摘 要：当前基于预训练说话人编码器的语音克隆方法可以为训练过程中见到的说话人合成较高音色相似性的语音，但对于训练中未看到的说话人，语音克隆的语音在音色上仍然与真实说话人音色存在明显差别。针对此问题，本文提出了一种基于音色一致的说话人特征提取方法，该方法使用当前先进的说话人识别模型TitaNet作为说话人编码器的基本架构，并依据说话人音色在语音片段中保持不变的先验知识，引入一种音色一致性约束损失用于说话人编码器训练，以此提取更精确的说话人音色特征，增加说话人表征的鲁棒性和泛化性，最后将提取的特征应用端到端的语音合成模型VITS进行语音克隆。实验结果表明，本文提出的方法在2个公开的语音数据集上取得了相比基线系统更好的性能表现，提高了对未见说话人克隆语音的音色相似度。

关键词：语音克隆；说话人编码器；说话人表征；音色一致性
中图分类号：TN912.33 **文献标识码：**A **DOI：**10.16798/j.issn.1003-0530.2023.04.013

引用格式：李嘉欣，张连海，李宜亭. 基于音色一致的语音克隆说话人特征提取方法[J]. 信号处理，2023，39(4)：719-729. DOI：10.16798/j.issn.1003-0530.2023.04.013.
Reference format： LI Jiaxin，ZHANG Lianhai，LI Yiting. Speaker feature extraction based on timbre consistency for voice cloning[J]. Journal of Signal Processing，2023，39(4)：719-729. DOI: 10.16798/j.issn.1003-0530.2023.04.013.

Speaker Feature Extraction Based on Timbre Consistency for Voice Cloning

LI Jiaxin ZHANG Lianhai LI Yiting
（School of Information System Engineering，PLA Strategic Support Force Information Engineering University，Zhengzhou，Henan 450001，China）

Abstract: Current speech cloning methods based on pre-trained speaker encoders can synthesize speech with high timbre similarity for speakers seen during training，but for speakers not seen during training，the timbre of cloned speech is still significantly different from that of real speaker. Aiming at this problem，this paper proposes a speaker feature extraction method based on timbre consistency. This method uses the current advanced speaker recognition model TitaNet as the basic architecture of the speaker encoder. According to the prior knowledge that the speaker timbre remains unchanged in the speech segment，a timbre consistency constraint loss is introduced for speaker encoder training to extract more accurate speaker timbre features and increase the robustness and generalization of speaker representation. Finally，the extracted features are applied to the end-to-end speech synthesis model VITS for speech cloning. Experimental results show that the proposed method achieves better performance than the baseline system on two public speech datasets and improves the timbre similarity of the cloned speech of unseen speakers.

Key words: voice cloning；speaker encoder；speaker representation；timbre consistency

收稿日期：2022-10-08；修回日期：2022-11-11
基金项目：国家自然科学基金(62171470)

1 引言

语音合成[1]是依据文本合成对应语音的技术。在深度学习技术发展之前，拼接语音合成[2]和统计参数语音合成方法[3]是主流语音合成技术。在过去几年中，随着深度学习技术的发展，基于神经网络强大的非线性建模能力，神经语音合成模型生成语音的质量、自然度和可懂度都有了很大的提高。由于文本到语音跨模态预测具有较大的难度，现阶段的神经语音合成模型主要由声学预测模块和声码器模块级联而成，也称为两阶段语音合成方法。声学预测模块的功能是依据文本预测合成语音的声学特征，典型的模型有 Tacotron2[4]和 FastSpeech2[5]等。声码器模块的功能是将预测的声学特征转换成语音波形，经典模型有 WaveNet[6]，WaveGlow[7]，HiFi-GAN[8]等。这种两阶段语音合成方法存在声学预测模块的输出和声码器模块的输入不匹配问题，因此目前更先进的方法使用完全端到端的语音合成模型，直接由文本生成语音波形，例如 VITS[9]，通过线性谱引导中间特征的学习。

神经语音合成模型的训练需要使用大量的高质量语音文本对数据，当扩展新的说话人时需要该新说话人充足的语料数据，极大限制了该方法的个性化应用，因此研究依据少量目标说话人的语音来合成具有目标说话人音色的语音成为了一个引人注意的方向，也称为语音克隆[10]，能够极大降低个性化语音合成的门槛和应用成本。

语音克隆目前有两种主要方法，分别为说话人自适应和说话人编码方法，其核心在于如何表示目标说话人的特征。说话人自适应方法使用目标说话人语音微调多说话人语音合成模型，因此需要重新训练模型，由于语音合成模型过于复杂，当使用少量数据进行训练时，容易出现模型过度拟合。基于此问题，文献[11]将语音合成模块拆分为两个级联模块，前一个模块用于预测语音的声学特征，后一个模块用于建模说话人的音色，两者通过说话人无关的音素后验概率图作为中间特征相连接，该方法的优点是，当只对新说话人音色进行学习时，只需要对说话人音色相关的模块进行微调即可，有效缓解模型的过拟合问题。AdaSpeech[12]在 FastSpeech2 模型的基础上引入条件层归一化，和说话人特征向量联合微调，使得微调参数量极大减少，在不降低音质的情况下学习新说话人音色。

说话人编码方法的思想是训练一个单独的说话人编码模块，用来提取反映目标说话人特征的编码说话人特征向量，将说话人特征向量作为控制条件作用于多说话人语音合成模型。这种方法不需要重新训练模型，因此对目标说话人语音数据量要求更低，只需要少量目标说话人语音即可提取说话人特征向量，便于实际应用。根据说话人编码器的训练方式可将说话人编码方法分为两类：预训练说话人编码器方法和参考编码器方法（联合训练）。预训练说话人编码器方法是将应用在其他语音应用任务（如说话人验证）中的说话人特征提取模块迁移过来作为说话人编码器。该方法的典型代表是 SV2TTS[13]，它使用 GE2E 损失[14]预训练的说话人验证模型作为说话人编码模块，提取说话人特征向量作为条件输入到语音合成模型 Tacotron2，使得模型能合成和目标说话人相似的语音。文献[15]在 Tacotron2 上测试了不同说话人验证模型，表明 LDE（Learnable Dictionary Encoding）向量方法[16]提高了新目标说话人的合成语音的自然度和相似度。参考编码器方法是将说话人编码器与语音合成模型联合训练。Attentron[17]参考多条语音并使用两种不同粒度的编码器提取说话人信息，其中细粒度编码器通过注意力机制提取表示局部说话人信息的向量序列，粗粒度编码器提取整句语音的全局说话人信息向量，从而捕获更多说话人音色相关的信息。StyleSpeech[18]将说话人音色视为语音风格，使用风格编码器提取风格嵌入向量，用来控制基于 FastSpeech2 的自适应层正则化模块，从而控制语音合成的音色。

使用参考编码器的方法受限于语音合成模型训练的数据，对训练数据外的语音泛化不强。而预训练说话人编码器方法泛化能力强，使得合成模型能对训练数据外的语音音色更具有泛化性。这种方法关键在于提取能恰当表示说话人音色特征的说话人特征向量。目前的编码器模型均源于其他语音应用任务，提取的特征向量在表示说话人音色特征方面还存在欠缺，导致合成语音的音色相似度比前一种方法差。对于训练不可见说话人，合成出的语音在相似度上存在明显的差距。

针对上述问题，本文提出一种基于音色一致性约束损失的说话人编码器训练方法，该方法借助目前性能优异的说话人识别模型 TitaNet[19]作为说话人编码器基本架构，在训练说话人编码器时，依据说话人音色在语音片段中是不变的先验知识，通过引入音色一致性约束，减轻语音片段中的其他因素干扰，从而获得更鲁棒的说话人音色表示。为了验证特征的性能，将此特征应用于端到端语音合成模型 VITS 中，通过对克隆语音的性能分析，表明本文提出的方法在 2 个公开的语音数据集上取得了相比 TitaNet 基线更好的性能表现。

2　音色一致说话人编码器

在语音克隆领域，不少工作[15,17,20-21]使用说话人识别模型作为说话人编码器，其中性能比较好的有 X-vector[22]，LDE-vector[16]等。但是对于新出现的目标说话人，合成语音的相似度仍然与真实语音有较大的差距，其中原因之一是当前使用的说话人编码器提取的特征对说话人音色信息描述不足。这是因为文本无关的说话人识别任务其目标是消除语音中语义内容的影响，关注区分说话人差异的特征属性，因此说话人识别任务中的说话人特征提取模块可以作为表征说话人特性的提取模型，但说话人识别模型的训练任务是识别不同说话人，学习的说话人表征是区分性特征，而不同说话人的语音中除了音色的不同，同时也受到其他变量的影响，如韵律、情感、环境噪声、语音录制条件等。在训练数据集中为了将不同说话人分类，可能将与说话人音色无关但是又能区分说话人的信息（如明显独特的韵律变化，环境噪声等）编码到说话人特征向量中。这些无关变量给说话人音色的表征引入不可控的噪声，直接影响语音克隆的克隆质量。

为了减轻其他无关因素的影响，基于同一段语音中说话人音色是一致不变，而韵律、噪声等在一段语音中是局部变化的先验假设，本文将足够长的语音片段和整句语音提取的说话人特征向量保持一致，通过约束整段语音提取的说话人特征向量和随机选取的局部语音片段提取的说话人特征向量尽量保持相同，使得模型更加关注全局一致性的说话人音色信息，减轻其他局部无关因素的影响。

对比分析 X-vector，LDE-vector 和 TitaNet 模型在文本无关说话人识别和说话人验证任务上的性能，TitaNet 模型在提取说话人特征方面性能更好，本文采用 TitaNet 模型作为基本说话人特征提取架构，加入音色约束提取用于语音克隆的说话人特征。

2.1　TitaNet

TitaNet 网络结构如图 1 所示，由三个部分组成，

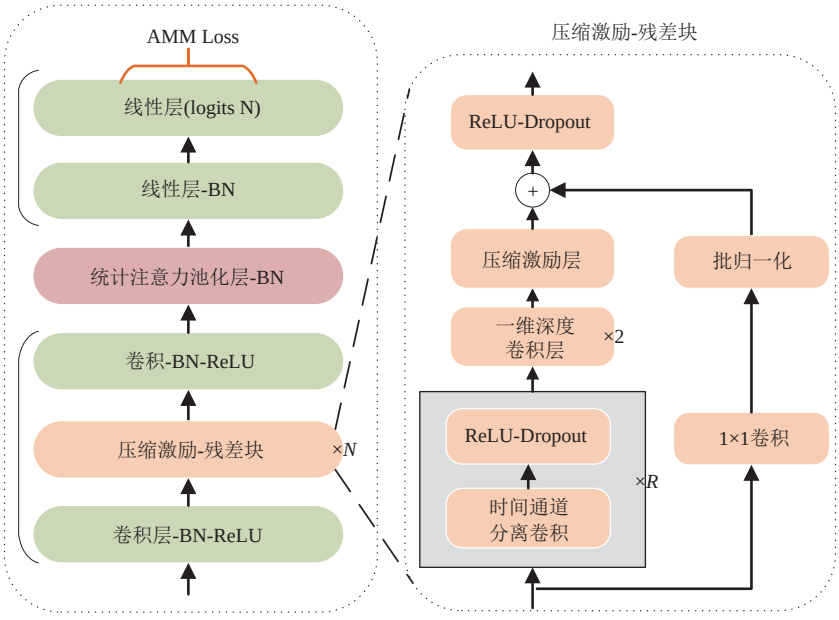

图 1　TitaNet 结构图

Fig. 1　Schematic diagram of TitaNet

从下到上依次为编码器、池化层和分类器。编码器提取语音帧级特征序列；池化层使用基于通道注意力的统计池化层，将帧级特征序列聚合为固定维度的句段级嵌入向量；分类器根据嵌入向量进行说话人分类，将其倒数第二层线性层的输出中间特征作为说话人特征向量。编码器使用具有全局上下文的压缩-激励层，显性建模通道间的依赖性，扩大帧级特征的感受野，捕捉全局相关性，提高了对频谱特征的提取能力。该模型将一维卷积替换成深度可分离卷积，降低网络参数量的同时保持了模型的表达能力。

TitaNet 模型选择加性角度间隔损失[23]（Additive Angular Margin loss）进行说话人分类训练，损失公式如式（1）所示。

$$L_{\mathrm{AMM}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\mathrm{e}^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}}{\mathrm{e}^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}+\sum_{j=1,j\neq y_i}^{n}\mathrm{e}^{s\left(\cos\left(\theta_{y_i}\right)\right)}} \tag{1}$$

其中 m 是边缘间隔因子，s 是尺度因子，θ_{y_i} 是最后一层线性层权重 W_j 和输入特征 x_i 的夹角。m 和 s 都是预先定义的超参数。

2.2 基于 TitaNet 的音色约束特征提取

语音克隆系统合成的语音音色相似度取决于提取的说话人特征向量对音色的表征能力。为了分析使用 TitaNet 提取数据集的说话人特征的性能，本文对整个数据集的说话人特征向量分布进行分析。在实验中发现，由相同说话人的不同语音片段提取的一些说话人特征向量间的相似度不高。并且，在有些较长的语音片段截取固定长度的语音片段得到的说话人特征向量和由整段语音片段提取的说话人特征向量相似度有较大差别。上述实验现象表明模型提取的说话人特征向量对音色的表征不稳定，可能编码较多音色无关的声学信息。这会影响语音克隆系统对音色的克隆性能。

如图 3 所示，TitaNet 通过基于通道注意力的统计池化层将帧级特征 $h_1h_2h_3\cdots$ 聚合成单一句级特征 H_{all}，句级特征和帧级特征的相似度体现了模型提取说话人特征向量时对各帧的关注程度。本文选取一段语音示例，计算句级特征和各帧级特征的余弦相似度，将模型对各帧的关注度进行可视化。如图 2 黑线部分所示，模型提取说话人特征向量的过程中，聚焦在局部的语音段，对其他语音段的关注度低，如对前面 30 帧和第 150 帧到第 180 帧的语音信息关注度低。并且说话人特征向量中对除音色信息的其他声学信息也进行了编码，如第 50 帧到 120 帧的静音部分也有较高的关注度。通过以上分析得出，TitaNet 模型提取说话人表征时存在两种问题：一是模型对局部的语音段关注度高，对其他语音段的关注度低。二是特征提取时容易受到局部的其他信息影响，如明显独特的韵律变化，环境噪声等。这些问题共同导致出现同一说话人的语音片段提取的说话人特征向量存在较大差别的现象。

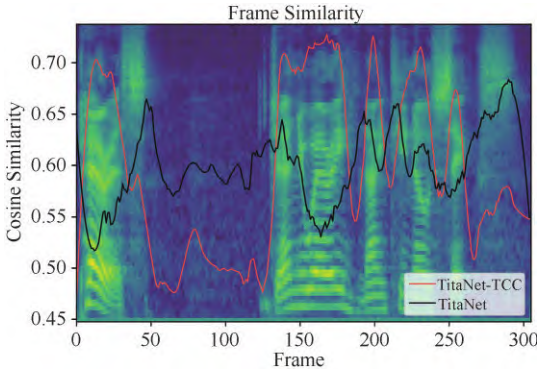

图 2 帧级特征相似度图

Fig. 2 Diagram of Frame-level feature similarity

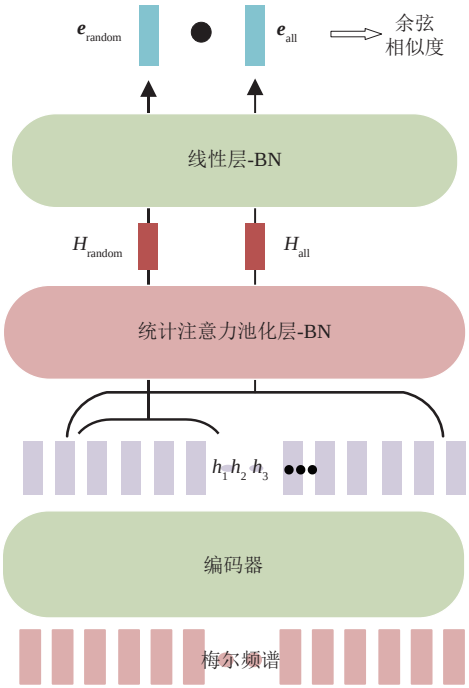

图 3 音色一致性约束示意图

Fig. 3 Schematic diagram of speaker consistency constraint

音色是指说话人声音的品质，由每个人的声带

及其共鸣器官的结构特征决定。音色代表了说话人发音特有的品质，可以在说话人语音中体现。说话人发音特有的品质一般不会改变，因此在单说话人的一整段语音中不同片段所体现的音色信息应该是一致不变的，即在同一段语音中音色具有全局不变性的特点。

说话人音色是整段语音的全局特征，在语音段中的音色信息是一致不变，而韵律、噪声等在一段语音中是局部变化。基于以上认识，为了使模型能充分利用每处语音片段中的音色信息，本文提出一种基于音色一致的音色约束损失（TCC，Timbre Consistency Constrain），鼓励随机选取适当长的语音片段和整句语音提取的说话人特征向量相似度高。由于随机选取语音片段，模型需要编码全局一致性的说话人音色信息才能让局部片段和整段语音提取的说话人特征向量尽量相似。这样使得模型关注每处语音段信息，更加关注全局一致性的说话人音色信息，减轻其他局部无关因素的影响，图 2 中 TitaNet-TCC 曲线展示了应用音色约束损失训练后模型对各语音帧关注程度。可以看到相比原 TitaNet 模型，TitaNet-TCC 对每处语音段均有较高关注，并减轻静音段对局部噪声的关注度。这说明说话人音色约束损失的实际作用符合理论分析。

本文提出的用于说话人编码器模型训练的音色约束损失的具体损失公式如式（2）所示，* 表示向量点积，· 表示标量积，$\|\boldsymbol{e}\|$ 表示取向量 $\boldsymbol{e}$ 的二范数。

$$L_{\mathrm{TCC}} = 1 - \frac{\boldsymbol{e}_{\mathrm{all}} * \boldsymbol{e}_{\mathrm{random}}}{\|\boldsymbol{e}_{\mathrm{all}}\| \cdot \|\boldsymbol{e}_{\mathrm{random}}\|} \tag{2}$$

如图 3 所示，模型输入梅尔频谱序列，经过编码器编码成等长的帧级特征序列 $h_1 h_2 h_3 \cdots$，选取整段帧级特征序列经过注意力池化层和线性层得到说话人特征向量 $\boldsymbol{e}_{\mathrm{all}}$，同时从帧级特征序列随机选取一段连续的特征序列同样输入到后续网络中得到说话人特征向量 $\boldsymbol{e}_{\mathrm{random}}$，计算两者的余弦相似度。

整个 TitaNet 模型训练损失公式如式（3）所示。

$$L_{\mathrm{TitaNet}} = L_{\mathrm{AMM}} + \lambda L_{\mathrm{TCC}} \tag{3}$$

λ 是权重超参数，权重超参数的设置对说话人编码器的训练至关重要，实验中对权重超参数的取值进行了实验。

3　基于 VITS 的语音克隆系统

为验证本文提取特征的有效性，搭建了基于 VITS 的语音克隆系统。该系统使用当前先进的说话人识别模型 TitaNet 作为说话人编码器提取更精确的说话人音色特征，扩展端到端的语音合成模型作为语音克隆模型。两阶段语音合成模型需要分别训练后进行微调，流程复杂，存在前后模块不匹配的问题，影响合成音质的质量。VITS 直接以端到端的方式从文本预测生成语音波形，摒弃了中间特征梅尔频谱，而由模型自行学习中间特征，提高了合成语音的自然度。本文在原 VITS 模型上进行了修改，如图 4 所示。

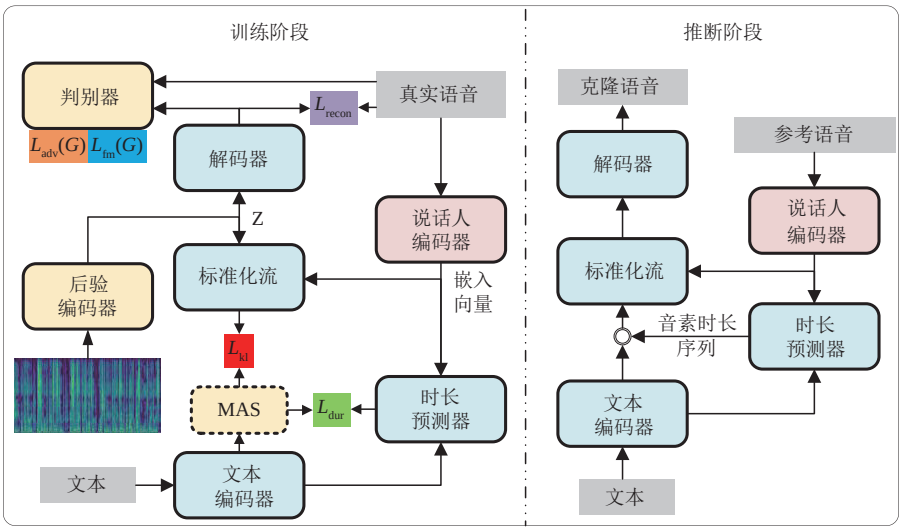

图 4　基于 VITS 的改进结构原理图

Fig. 4　Schematic diagram of improved structure based on VITS

整个模型由文本编码器、标准化流、后验编码器、解码器、时长预测器、判别器六个部分组成。原VITS模型是通过说话人查阅表的方法实现多说话人的语音合成，在训练中学习说话人的说话人特征向量。本文使用说话人编码器替代说话人查阅表，直接从参考语音中提取说话人特征向量。在训练阶段，后验编码器将语音线性谱编码成潜变量Z。文本编码器将输入的音素序列映射生成文本特征。标准化流是由一系列基于流的可逆变换模块组成，在仿射耦合层嵌入说话人特征向量。标准化流将潜变量Z复杂分布转换成潜变量Z_p的简单分布。模型使用单调对齐方法[24]（MAS，Monotonic Alignment Search），通过Z_p和文本特征的最大化似然计算得到对齐信息，即单一音素对应的语音帧数。时长预测器以文本特征和说话人特征向量为条件，预测音素的持续时长信息。时长预测器的训练以MAS得到的对齐信息为目标标签，计算时长预测损失，如公式(4)所示。d_i为由对齐信息得到的第i个音素的持续长度，$\hat{d}_i$为时长预测器预测的第i个音素的持续长度，计算两者绝对值距离。

$$L_{\mathrm{dur}}=\sum_i\left|\hat{d}_i-d_i\right| \tag{4}$$

依据MAS得到的对齐信息，文本特征序列扩展到长度与语音帧长度一致。使用KL散度（Kullback-Leibler Divergence）损失拉近Z_p分布和文本特征分布的距离，如公式(5)所示，x_{lin}是语音线性谱，后验编码器将x_{lin}编码成潜变量Z。KL散度拉近先验编码器在给定文本信息c_{text}和对齐信息A后的文本特征z_i的条件分布$p_\theta\left(z|c_{\mathrm{text}},A\right)$和后验编码器的在给定线性谱$x_{\mathrm{lin}}$的潜变量$Z$的条件分布$q_\varphi\left(z|x_{\mathrm{lin}}\right)$，使得文本信息能准确控制潜变量$Z$的生成。

$$L_{\mathrm{kl}}=\mathbb{E}_{q_\varphi\left(z|x_{\mathrm{lin}}\right)}\left[\log q_\varphi\left(z|x_{\mathrm{lin}}\right)-\log p_\theta\left(z_i|c_{\mathrm{text}},A\right)\right] \tag{5}$$

解码器使用了HiFi-GAN V1生成器的网络架构，依据潜变量Z预测生成语音波形。生成语音和真实语音计算重构损失如式(6)所示，计算真实语音的梅尔频谱x_{mel}和合成语音的梅尔频谱$\hat{x}_{\mathrm{mel}}$间的绝对值距离。

$$L_{\mathrm{recon}}=x_{\mathrm{mel}}-\hat{x}_{\mathrm{mel}} \tag{6}$$

同时生成语音和真实语音输入到判别器中，解码器G和判别器D进行对抗训练，判别器的目标是区分真实语音和解码器恢复生成的语音，而解码器需要生成能欺骗判别器的语音，通过两者的相互对抗，使得解码器能生成与真实语音难以分辨的语音。判别器D对语音的判别结果为1或0，判定为真则输出1，反之输出0。对抗损失公式如式(7)、(8)所示，y表示真实语音，z为潜变量。

$$L_{\mathrm{adv}}(G)=\mathbb{E}_z\left[\left(D\left(G(z)\right)-1\right)^2\right] \tag{7}$$

$$L_{\mathrm{adv}}(D)=\mathbb{E}_{(y,z)}\left[\left(D\left(y\right)-1\right)^2+\left(D\left(G(z)\right)\right)^2\right] \tag{8}$$

为了给生成器提供更多的引导信息，模型使用特征匹配损失（Feature Matching Loss，FML）引导模型在特征空间学习鉴别真实和虚假样本的相似尺度，计算真实语音和合成语音输入到判别器得到的每层网络中间特征的差异，使用绝对值距离来衡量这种差异，特征匹配损失函数如式(9)所示。

$$L_{\mathrm{fm}}(G)=\mathbb{E}_{(y,z)}\left[\sum_{j=1}^{T}\frac{1}{N_j}\left|D^j\left(y\right)-D^j\left(G(z)\right)\right|_1\right] \tag{9}$$

其中T代表判别器网络层数，D^j和N_j分别表示鉴别器的第j层网络的输出中间特征和特征数量。

综上VITS的总损失如式(10)所示，忽略了各个损失的权重超参数。说话人编码器经预训练后参数固定不动，为语音合成模型的训练提供说话人特征向量。

$$L_{\mathrm{total}}=L_{\mathrm{recon}}+L_{\mathrm{kl}}+L_{\mathrm{dur}}+L_{\mathrm{adv}}(G)+L_{\mathrm{adv}}(D)+L_{\mathrm{fm}}(G) \tag{10}$$

VITS模型训练完成后进行推演时不需要后验编码器和判别器模块，直接由文本编码器编码文本特征，依据时长预测器提供的时长信息，将文本特征序列扩展到语音帧级表征序列。之后语音帧级表征序列经过标准化流增加说话人音色的变换得到潜变量Z。最终解码器依据潜变量Z预测生成语音。

4 实验结果及分析

4.1 实验设置

实验数据采用公开的英语数据集，其中说话人编码器在大型人声识别数据集Voxceleb1[25]和Voxceleb2[26]上训练。Voxceleb1有1211个说话人，Voxceleb2上有5994个说话人，音频采样率为16 kHz，训练时进行数据增广，增加混响和加性噪声。说话人编码器的输入为梅尔频谱，梅尔频谱的维度、傅里叶变换长度、窗长和窗移分别设置为80、512、400

和160个采样点。因为1.6 s语音足够表示音色信息，梅尔频谱的帧移为10 ms，设置随机选取的帧级特征序列片段长度为160帧刚好对应1.6 s时长的语音。

语音克隆模型在VCTK[1]数据集和LibriTTS[27]数据集上训练。VCTK数据集包括109个说话人的英语音频片段和相应的文本，每个说话人的音频数据量较平均，单个说话人总音频时长有二十多分钟，音频总长度约为44个小时，采样率为44 kHz，在训练中本文保留3名男性和3名女性作为测试的未见过的新说话人。LibriTTS数据集有2456个说话人，音频总长度为585小时，采样率为24 kHz，说话人语音数据不平均，其中有的说话人音频总时长有三十多分钟，也有说话人音频总时长不到一分钟。在训练中，本文使用子集train-clean-100和train-clean-360的1151个说话人作为训练集，选取来自子集clean-test的10名说话人语音作为未见过的新说话人的测试集。因为VCTK的音频起始点静音片段比较多，实验中利用开源的自动对齐工具Montreal Forced Aligner[2]将过长的静音片段消除，并将所有的语音下采样到22050 Hz。线性谱傅里叶变换长度、窗长和窗移分别设置为1024、1024和256个采样点。

本文在Voxceleb1和Voxceleb2上训练x-vector，LDE-vector，TitaNet，和TitaNet-TCC作为说话人编码器，并且在Vox1-test测试集上的等错误率（EER，Equal Error Rate）分别为3.42%、3.22%、0.82%、0.89%。模型的输入均为80维的梅尔频谱，输出的说话人特征向量维度均为256维。语音合成模型使用开源的端到端语音合成模型VITS，其损失权重超参数和原模型一致。所有模型均已在两张GPU（NVIDIA Tesla V100，内存为16 GB）上基于PyTorch框架训练200K步，批处理大小为32，采用ADAM优化器，其中$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\varepsilon = 1e^{-6}$，学习率设置为$1e^{-3}$。

为了实验对比音色约束损失的有效性，同时对比TitaNet和当前语音克隆方法使用的说话人编码器模型的性能。本文使用以下方法进行语音克隆方法评估：

（1）真实语音：从训练集和测试集中提取的语音片段。

（2）VITS：使用说话人查阅表训练的VITS模型，只能合成查阅表中所对应说话人音色的语音。

（3）VITS+X-vector：使用预训练的X-vector说话人编码器提取的说话人特征向量训练VITS模型。

（4）VITS+LDE-vector：使用预训练的LDE-vector说话人编码器提取的说话人特征向量训练VITS模型。

（5）VITS+TitaNet：使用预训练的TitaNet说话人编码器提取的说话人特征向量训练VITS模型，作为验证本文音色约束损失有效性的基线系统。

（6）VITS+TitaNet-TCC：使用预训练的TitaNet-TCC说话人编码器提取的说话人特征向量训练VITS模型。

4.2　克隆语音相似度评价

实验采用语音克隆方法通用的主观评价：平均主观意见得分（Mean Opinion Score，MOS）和相似度平均主观意见得分（Similarity Mean Opinion Score，SMOS）来比较不同模型对训练中已见说话人和新说话人的语音合成质量和音色相似度。从训练集中选取20条语音作为可见说话人的评估集，从测试集中选取20条语音作为不可见说话人的评估集，由10名精通英语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，分数值在从1至5的区间内。统计打分取平均后计算95%的置信区间，结果如表1所示。从表中可以看到，在四种语音克隆方法中，VITS+TitaNet方法在可见说话人中取得了最好的自然度和相似度，VITS+TitaNet-TCC方法在可见说话人中效果稍微比VITS+TitaNet方法差，但是VITS+TitaNet-TCC方法在不可见说话人中相比其他语音克隆明显提高了克隆语音的自然度和相似度。本文分析认为这是因为TitaNet使用压缩-激励层的全局上下文信息，相比X-vector和LDE-vector能提取更丰富的说话人信息从而获取更好的说话人音色表征。TitaNet-TCC进一步通过说话人音色一致性约束损失去除了说话人特征向量中与音色无关的信息，所以对不可见说话人中泛化能力更强，验证了音色约束损失的有效性。而TitaNet提取的说话人特征向量相比前者含有语音的更多音色之外的信息，在训练中模

[1]https://datashare.ed.ac.uk/download/DS_10283_3443.zip

[2]https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner

表1 不同模型的MOS和SMOS得分

Tab. 1 MOS and SMOS scores of different models

模型	可见说话人		不可见说话人	
	MOS	SMOS	MOS	SMOS
真实语音	4.46±0.08	4.58±0.07	4.46±0.08	4.58±0.07
VITS	4.32±0.07	4.42±0.07	-	-
VITS+X-vector	4.21±0.08	4.12±0.08	3.95±0.08	3.34±0.08
VITS+LDE-vector	4.24±0.08	4.2±0.07	4.02±0.08	3.43±0.08
VITS+TitaNet	**4.28±0.06**	**4.24±0.08**	4.10±0.09	3.54±0.08
VITS+TitaNet-TCC	4.26±0.06	4.23±0.06	**4.21±0.08**	**3.80±0.06**

型学习到这些信息，使得在可见说话人中效果略好些。此外，从表中可知，语音克隆方法对不可见说话人音色的克隆音色相似度还是存在差距。

4.3 说话人编码器的损失权重超参数设置

权重超参数的设置对说话人编码器的训练至关重要，本文对六种权重超参数取值下说话人编码器训练结果进行了对比实验，记录最后模型在Vox1-test的说话人确认等错误率(EER)和同一说话人的说话人向量余弦相似度的平均值。实验结果如图5所示，随着超参数取值增大，同说话人的说话人向量余弦相似度值先快速增大再缓慢增加，等错误率逐渐增大。这说明权重超参数取值过小时，对说话人向量一致性约束效果不明显。权重超参数取值过大时，模型过分关注说话人向量的一致性，可能关注于提取相同的特征，过滤掉音色间的辨别信息，影响了说话人向量对不同说话人的区分能力。综上考虑，权重超参数取值为1时效果最好，既强调了说话人向量的一致性又对说话人向量的区分能力影响可以忽略不计。

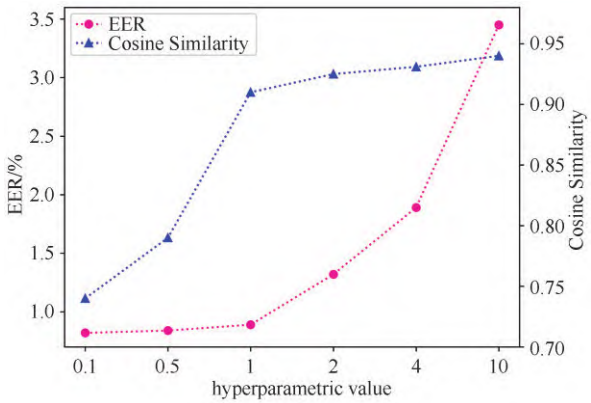

图5 超参数取值实验图

Fig. 5 Experimental diagram of hyperparametric value

4.4 说话人特征向量分布

语音克隆的音色相似度的关键是对参考语音中音色特征的提取，本文直接在测试集上对是否使用音色一致性约束损失进行对比，分别使用TitaNet和TitaNet-TCC从测试语音中提取说话人特征向量，计算测试集中两两语音片段的说话人特征向量的余弦相似度，其余弦相似度分布直方图如图6所示。

相同说话人的特征向量余弦相似度值普遍比VITS+TitaNet高，说明通过音色一致性约束强调模型从同一语音片段中提取音色表征向量的一致性，使得从VITS+TitaNet-TCC提取的说话人特征向量更加鲁棒，减少了其他无关信息对特征的干扰。

4.5 说话人特征向量聚类

为了探究在客观层面上未见说话人的真实语音和克隆语音的音色相似程度，在未见说话人测试集中每个说话人随机选取十条真实语音，并分别作为参考语音，分别使用VITS+TitaNet和VITS+TitaNet-TCC合成对应的克隆语音。虽然训练的说话人编码器均能提取可表示说话人音色特征的说话人特征向量，但是为了方便统一比较，本文使用开源说话人特征向量提取工具[3]提取参考语音和克隆语音的说话人特征向量，使用UMAP[4]进行降维可视化，如图7所示，同一说话人用同一颜色表示，圆形图标表示真实语音，交叉图标表示对应的克隆语音。从图中可以看出，两种方法的同一说话人的说话人向量明显聚集在一起，并且真实语音和克隆语音有部分已重叠，说明克隆语音在说话人特征向量空间中与真实语音

[3]https://github.com/resemble-ai/Resemblyzer

[4]https://github.com/lmcinnes/umap

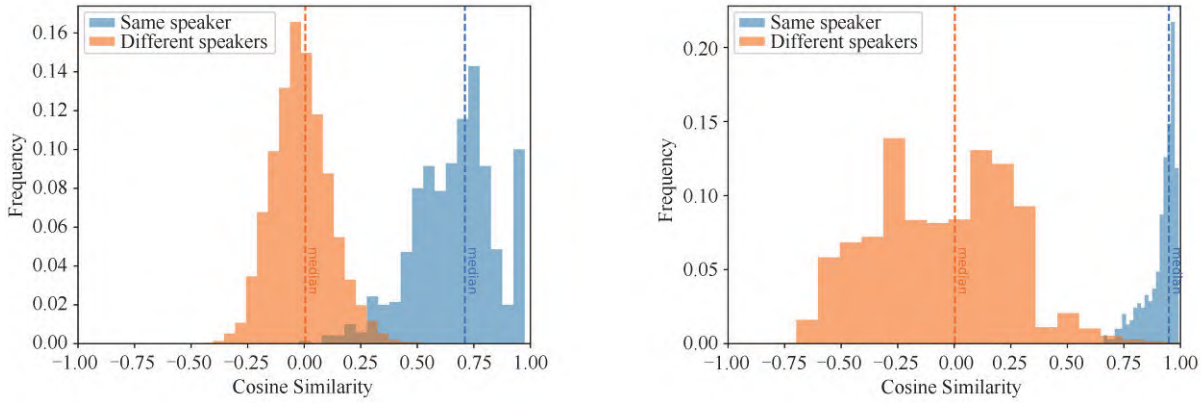

图 6　余弦相似度分布直方图，左图为 TitaNet 方法，右图为 TitaNet-TCC 方法

Fig. 6　Histogram of Cosine similarity distribution, the left is TitaNet and the right is TitaNet-TCC

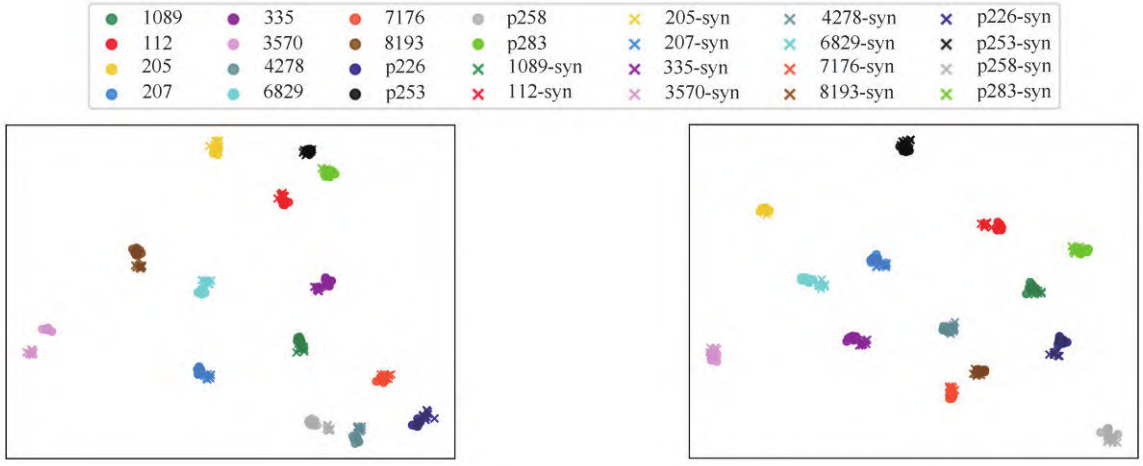

图 7　说话人向量聚类图，左图为 TitaNet 方法，右图为 TitaNet-TCC 方法

Fig. 7　Diagram of Speaker embed clustering, the left is TitaNet and the right is TitaNet-TCC

的相似程度较高。但是可以观察到说话人 3570、8193 和 p258，VITS+TitaNet-TCC 方法的克隆语音更加与真实语音接近，客观上说明音色一致性约束提高了对未见说话人的语音克隆的音色相似度。

5　结论

当前基于预训练说话人编码器的语音克隆方法中提取的说话人特征关注区分说话人差异的特征属性，容易受除音色外的其他变量影响，导致对音色信息描述不足，影响语音克隆的音色相似度。本文针对此问题，提出了一种基于音色一致的音色约束损失，在说话人编码器模型训练时引导模型关注全局说话人特征信息，减轻语音片段中的其他因素干扰，从而获得更鲁棒的说话人特征。实验结果表明，音色约束损失去除了说话人特征向量中与音色无关的信息，提高对不可见说话人的泛化能力，最终在不损失合成语音的自然度的同时，提高了语音克隆模型对不可见说话人音色的克隆相似度。本文改进的说话人特征可以用于其他语言任务，如说话人识别、说话人验证和说话人日志任务，具有对训练集外的说话人更加鲁棒的特点。

参考文献

[1] TAYLOR P. Text-to-Speech Synthesis [M]. Cambridge: Cambridge University Press, 2009: 2-3.

[2] ADELL J, ESCUDERO D, BONAFONTE A. Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence [J]. Speech Communication, 2012, 54(3): 459-476.

[3] ZEN Heiga, TOKUDA K, BLACK A W. Statistical parametric speech synthesis[J]. Speech Communication, 2009, 51(11): 1039-1064.

[4] SHEN J, PANG Ruoming, WEISS R J, et al. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions[C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, Canada: IEEE, 2018 : 4779-4783.

[5] REN Yi, HU Chenxu, TAN Xu, et al. FastSpeech 2: fast and high-quality end-to-end text to speech[C]//Proceedings of the 9th International Conference on Learning Representations. Virtual Event, Austria : OpenReview.net, 2021: 4670-4675.

[6] VAN DEN OORD A, DIELEMAN S, ZEN H, et al. WaveNet: a generative model for raw audio[C]//Proceedings of the 9th ISCA Speech Synthesis Workshop. Sunnyvale, USA: ISCA, 2016: 113-117.

[7] PRENGER R, VALLE R, CATANZARO B. Waveglow: a flow-based generative network for speech synthesis[C]// Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Brighton, UK: IEEE, 2019: 3617-3621.

[8] SU Jiaqi, JIN Zeyu, FINKELSTEIN A. HiFi-GAN: high-fidelity denoising and dereverberation based on speech deep features in adversarial networks[C]//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA, 2020: 4506-4510.

[9] KIM J, KONG J, SON J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech[C]//Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: PMLR, 2021: 5530-5540.

[10] ARIK S Ö, CHEN J, PENG K, et al. Neural voice cloning with a few samples[C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2018. Montréal, Canada: NeurIPS, 2018: 10040-10050.

[11] WANG Tao, TAO Jianhua, FU Ruibo, et al. Spoken content and voice factorization for few-shot speaker adaptation[C]//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA, 2020: 796-800.

[12] CHEN Mingjian, TAN Xu, LI Bohan, et al. AdaSpeech: adaptive text to speech for custom voice[C]//Proceedings of the 9th International Conference on Learning Representations, Virtual Event, Austria: OpenReview.net, 2021: 1334-1338.

[13] JIA Y, ZHANG Yu, WEISS R J, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis[C]//Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018. Montréal, Canada: NeurIPS, 2018: 4485-4495.

[14] WAN Li, WANG Quan, PAPIR A, et al. Generalized end-to-end loss for speaker verification[C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, Canada: IEEE, 2018: 4879-4883.

[15] COOPER E, LAI C I, YASUDA Y, et al. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings[C]// ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain. IEEE, 2020: 6184-6188.

[16] CAI Weicheng, CHEN Jinkun, LI Ming. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system[C]//Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2018). Les Sables d'Olonne, France:ISCA, 2018: 74-81.

[17] CHOI S, HAN S, KIM D, et al. Attentron: few-shot text-to-speech utilizing attention-based variable-length embedding[C]//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA, 2020: 2007-2011.

[18] MIN D, LEE D B, YANG E, et al. Meta-StyleSpeech: Multi-speaker adaptive text-to-speech generation[C]//Proceedings of the 38th International Conference on Machine Learning, Vienna, Austria: PMLR, 2021: 7748-7759.

[19] KOLUGURI N R, PARK T, GINSBURG B. TitaNet: neural model for speaker representation with 1D depth-wise separable convolutions and global context[C]//Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, Singapore: IEEE, 2022: 8102-8106.

[20] NEEKHARA P, HUSSAIN S, DUBNOV S, et al. Expressive neural voice cloning[C]//Proceedings of the 2021 Asian Conference on Machine Learning. Virtual Event: PMLR, 2021: 252-267.

[21] SONG Wei, YUAN Xin, ZHANG Zhengchen, et al. Dian: duration informed auto-regressive network for voice cloning [C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, Canada: IEEE,2021: 8598-8602.

[22] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: robust DNN embeddings for speaker recognition [C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, Canada: IEEE, 2018: 5329-5333.

[23] DENG Jiankang, GUO Jia, XUE Niannan, et al. ArcFace: Additive angular margin loss for deep face recognition [C]//Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 4690-4699.

[24] KIM J, KIM S, KONG J, et al. Glow-TTS: a generative flow for text-to-speech via monotonic alignment search [C]//Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020. Virtual Event: NeurIPS, 2020: 8067-8077.

[25] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: A large-scale speaker identification dataset [C]//Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm. Sweden: ISCA, 2017: 2616-2620.

[26] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: Deep speaker recognition [C]//Proceedings of the 19th Annual Conference of the International Speech Communication Association. Hyderabad, India: ISCA, 2018: 1086-1096.

[27] ZEN H, DANG V, CLARK R, et al. LibriTTS: A corpus derived from LibriSpeech for text-to-speech [C]//20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019: 1526-1530.

作者简介

李嘉欣　男，1998 年生，湖南湘乡人。中国人民解放军战略支援部队信息工程大学硕士研究生，主要研究方向为智能信息处理、人工智能、语音合成。
E-mail：414171817@qq. com

张连海　男，1971 年生，山东单县人。中国人民解放军战略支援部队信息工程大学教授，主要研究方向为语音信号处理、智能信息处理、人工智能、信号分析等。
E-mail：llhhzz163@163. com

李宜亭　男，1993 年生，甘肃兰州人。中国人民解放军战略支援部队信息工程大学硕士研究生，主要研究方向为智能信息处理、语音识别、模型压缩。
E-mail：904890536@qq. com