基于 i 向量和变分自编码相对生成对抗网络的语音转换

李燕萍[1] 曹盼[1] 左宇涛[1] 张燕[2] 钱博[3]

摘 要 提出一种基于 i 向量和变分自编码相对生成对抗网络的语音转换方法, 实现了非平行文本条件下高质量的多对多语音转换. 性能良好的语音转换系统, 既要保持重构语音的自然度, 又要兼顾转换语音的说话人个性特征是否准确. 首先为了改善合成语音自然度, 利用生成性能更好的相对生成对抗网络代替基于变分自编码生成对抗网络模型中的 Wasserstein 生成对抗网络, 通过构造相对鉴别器的方式, 使得鉴别器的输出依赖于真实样本和生成样本间的相对值, 克服了 Wasserstein 生成对抗网络性能不稳定和收敛速度较慢等问题. 进一步为了提升转换语音的说话人个性相似度, 在解码阶段, 引入含有丰富个性信息的 i 向量, 以充分学习说话人的个性化特征. 客观和主观实验表明, 转换后的语音平均梅尔倒谱失真距离值较基准模型降低 4.80%, 平均意见得分值提升 5.12%, ABX 值提升 8.60%, 验证了该方法在语音自然度和个性相似度两个方面均有显著的提高, 实现了高质量的语音转换.

关键词 语音转换, 相对生成对抗网络, i 向量, 非平行文本, 变分自编码器, 多对多

引用格式 李燕萍, 曹盼, 左宇涛, 张燕, 钱博. 基于 i 向量和变分自编码相对生成对抗网络的语音转换. 自动化学报, 2022, **48**(7): 1824−1833

DOI 10.16383/j.aas.c190733

Voice Conversion Based on i-vector With Variational Autoencoding Relativistic Standard Generative Adversarial Network

LI Yan-Ping[1] CAO Pan[1] ZUO Yu-Tao[1] ZHANG Yan[2] QIAN Bo[3]

Abstract This paper proposes a novel voice conversion method based on i-vector and variational autoencoding relativistic standard generative adversarial network, which can realize high-quality many-to-many voice conversion for non-parallel corpora. A high performance voice conversion method should not only ensure speech naturalness, but also take into account speaker similarity of converted speech. Firstly, in order to improve the speech naturalness, the Wasserstein generative adversarial network in the voice conversion model based on variational autoencoding generative adversarial network is replaced by the relativistic standard generative adversarial network, which makes the output of the discriminator depend on the relativistic standard value between real and generated samples by constructing a relativistic standard discriminator, overcoming the unstable performance and slow convergence rate. Furthermore, i-vector representing speaker characteristics is adopted as speaker representation for many-to-many voice conversion in addition to traditional one-hot vector, thus significantly improving speaker similarity of converted speech. Sufficient objective and subjective experiments show that the average value of mel-cepstral distortion is decreased by 4.80%, the mean opinion score is increased by 5.12%, and ABX is increased by 8.60% compared with baseline variational autoencoding wasserstein generative adversarial network method which demonstrate that the proposed method has a great improvement on both speech naturalness and speaker similarity.

Key words Voice conversion, relativistic standard generative adversarial network, i-vector, non-parallel corpora, variational autoencoder, many-to-many

Citation Li Yan-Ping, Cao Pan, Zuo Yu-Tao, Zhang Yan, Qian Bo. Voice conversion based on i-vector with variational autoencoding relativistic standard generative adversarial network. *Acta Automatica Sinica*, 2022, **48**(7): 1824−1833

语音转换是在保持语音内容不变的同时, 改变一个人的声音, 使之听起来像另一个人的声音[1−2]. 根据训练过程对语料的要求, 分为平行文本条件下

收稿日期 2019-10-23 录用日期 2020-07-27
Manuscript received October 23, 2019; accepted July 27, 2020
国家自然科学青年基金 (61401227), 国家自然科学基金 (61872199, 61872424), 金陵科技学院智能人机交互科技创新团队建设专项 (218/010119200113) 资助
Supported by National Natural Science Foundation of Youth Foundation of China (61401227), National Natural Science Foundation of China (61872199, 61872424), and Special Project of Intelligent Human-Computer Interaction Technology Innovation Team Building of Jinling Institute of Technology (218/010119200113)
本文责任编委 贾磊
Recommended by Associate Editor JIA Lei

1. 南京邮电大学通信与信息工程学院 南京 210003 2. 金陵科技学院 南京 211169 3. 南京电子技术研究所 南京 210039
1. School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003 2. Jinling Institute of Technology, Nanjing 211169 3. Nanjing Institute of Electronic Technology, Nanjing 210039

的语音转换和非平行文本条件下的语音转换. 在实际应用中, 预先采集大量平行训练文本不仅耗时耗力, 而且在跨语种转换和医疗辅助系统中往往无法采集到平行文本, 因此非平行文本条件下的语音转换研究具有更大的应用背景和现实意义.

性能良好的语音转换系统, 既要保持重构语音的自然度, 又要兼顾转换语音的说话人个性信息是否准确. 近年来, 为了改善转换后合成语音的自然度和说话人个性相似度, 非平行文本条件下的语音转换研究取得了很大进展, 根据其研究思路的不同, 大致可以分为 3 类, 第 1 类思想是从语音重组的角度, 在一定条件下将非平行文本转化为平行文本进行处理[3−4], 其代表算法包括两种, 一种是使用独立于说话人的自动语音识别系统标记音素, 另一种是借助文语转换系统将小型语音单元拼接成平行语音. 该类方法原理简单, 易于实现, 然而这些方法很大程度上依赖于自动语音识别或文语转换系统的性能; 第 2 类是从统计学角度, 利用背景说话人的信息作为先验知识, 应用模型自适应技术, 对已有的平行转换模型进行更新, 包括说话人自适应[5−6] 和说话人归一化等. 但这类方法通常要求背景说话人的训练数据是平行文本, 因此并不能完全解除对平行训练数据的依赖, 还增加了系统的复杂性; 前两类通常只能为每个源−目标说话人对构建一个映射函数, 即一对一转换, 当存在多个说话人对时, 就需要构建多个映射函数, 增加系统的复杂性和运算量; 第 3 类是解卷语义和说话人个性信息的思想, 转换过程可以理解为源说话人语义信息和目标说话人个性信息的重构, 其代表算法包括基于条件变分自编码器 (Conditional variational auto-Encoder, C-VAE)[7] 方法、基于变分自编码生成对抗网络 (Variational autoencoding wasserstein generative adversarial network, VAWGAN)[8] 方法和基于星型生成对抗网络 (Star generative adversarial network, StarGAN)[9] 方法. 这类方法直接规避了非平行文本对齐的问题, 实现将多个源−目标说话人对的转换整合在一个转换模型中, 提供了多说话人向多说话人转换的新框架, 即多对多转换, 成为目前非平行文本条件下语音转换的主流方法.

基于 C-VAE 模型的语音转换方法, 其中的编码器对语音实现语义和个性信息的解卷, 解码器通过语义和说话人身份标签完成语音的重构, 从而解除对平行文本的依赖, 实现多说话人对多说话人的转换. 但是由于 C-VAE 基于理想假设, 认为观察到的数据通常遵循高斯分布, 导致解码器的输出语音过度平滑, 转换后的语音质量不高. 基于循环一致生成对抗网络的语音转换方法[10] 可以在一定程度上解决过平滑问题, 但是该方法只能实现一对一的语音转换.

Hsu 等[8] 提出的 VAWGAN 模型通过在 C-VAE 中引入 Wasserstein 生成对抗网络 (Wasserstein generative adversarial network, WGAN)[11], 将 VAE 的解码器指定为 WGAN 的生成器来优化目标函数, 一定程度上提升转换语音的质量, 然而 Wasserstein 生成对抗网络仍存在一些不足之处, 例如性能不稳定, 收敛速度较慢等. 同时, VAWGAN 使用说话人身份标签 one-hot 向量建立语音转换系统, 而该指示标签无法携带更为丰富的说话人个性信息, 因此转换后的语音在个性相似度上仍有待提升.

针对上述问题, 本文从以下方面提出改进意见: 1) 通过改善生成对抗网络[12] 的性能, 进一步提升语音转换模型生成语音的清晰度和自然度; 2) 通过引入含有丰富说话人个性信息的表征向量, 提高转换语音的个性相似度. 2019 年, Baby 等[13] 通过实验证明, 相比于 WGAN, 相对生成对抗网络 (Relativistic standard generative adversarial networks, RSGAN) 生成的数据样本更稳定且质量更高. 此外, 在说话人确认[14−16] 和说话人识别[17] 领域的相关实验证明, i 向量 (Identity-vector, i-vector) 可以充分表征说话人个性信息. 鉴于此, 本文提出基于 i 向量和变分自编码相对生成对抗网络的语音转换模型 (Variational autoencoding RSGAN and i-vector, VARSGAN + i-vector), 该方法将 RSGAN 应用在语音转换领域, 利用生成性能更好的相对生成对抗网络替换 VAWGAN 模型中的 Wasserstein 生成对抗网络, 同时在解码网络引入含有丰富说话人个性信息的 i 向量辅助语音的重构. 充分的客观和主观实验表明, 本文方法在有效改善合成语音自然度的同时进一步提升了说话人个性相似度, 实现了非平行文本条件下高质量的多对多语音转换.

1　基于 VAWGAN 的语音转换基准方法

基于 VAWGAN 语音转换模型利用 WGAN[11] 提升了 C-VAE 的性能, 其中 C-VAE 的解码器部分由 WGAN 中的生成器代替. VAWGAN 模型由编码器、生成器和鉴别器 3 部分构成. 完整的语音转换模型可表示为:

$$\hat{\boldsymbol{x}} = \hat{f}(\boldsymbol{x}, \boldsymbol{y}) = f_\theta(\boldsymbol{z}, \boldsymbol{y}) = f_\theta(f_\phi(\boldsymbol{x}), \boldsymbol{y}) \quad (1)$$

式中, $f_\phi(\cdot)$ 表示编码过程, 通过编码过程将输入语音 $\boldsymbol{x}$ 转换为独立于说话人的隐变量 $\boldsymbol{z}$, 认为是与说话人个性特征无关的语义信息. $f_\theta(\cdot)$ 表示解码过程, 将说话人标签 $\boldsymbol{y}$ 拼接至隐变量 z 上构成联合特征 $(\boldsymbol{z}, \boldsymbol{y})$, 在解码过程中利用联合特征 $(\boldsymbol{z}, \boldsymbol{y})$ 重构出特定说话人相关的语音, 然后将真实语音 $\boldsymbol{x}$ 和生成语音 $\hat{\boldsymbol{x}}$ 送入鉴别器判别真假. 同时, 利用表征说话人身份的 one-hot 标签 $\boldsymbol{y}$, VAWGAN 模型可以根据 $\boldsymbol{y}$ 的数值对其表示的特定说话人进行语音转换, 从而实现多说话人对多说话人的语音转换.

为实现语音转换, WGAN 通过 Wasserstein 目标函数[8] 来代替生成对抗网络中的 JS(Jensen-Shannon) 散度来衡量生成数据分布和真实数据分布之间的距离, 在一定程度上改善了传统生成对抗网络[18] 训练不稳定的问题.

综上分析可知, VAWGAN 利用潜在语义内容 $\boldsymbol{z}$ 和说话人标签 $\boldsymbol{y}$ 重构任意目标说话人的语音, 实现了非平行文本条件下多对多的语音转换. 该基准模型中 WGAN 采用权重剪切操作来强化 Lipschitz 连续性限制条件, 但仍存在训练不易收敛, 性能不稳定等问题, 在数据生成能力上仍存在一定的改进空间. 此外, VAWGAN 利用 one-hot 标签表征说话人身份, 而 one-hot 标签只是用于指示不同说话人, 无法携带更为丰富的说话人个性信息. 通过提升 WGAN 的性能或找到生成性能更加强大的生成对抗网络, 有望获得更好自然度的语音, 进一步引入含有丰富说话人个性信息的表征向量能够有助于提升说话人个性相似度.

2 改进的基于 VARSGAN + i-vector 的语音转换方法

2.1 RSGAN 的原理

为进一步提升 VAWGAN 的性能, 通过找到一个生成性能更加强大的 GAN 替换 WGAN 是本文的一个研究出发点. 2019 年 Baby 等[13] 通过实验证明相比于最小二乘 GAN[19] 和 WGAN[11], RSGAN 生成的数据样本更稳定且质量更高. RSGAN 由标准生成对抗网络发展而来, 通过构造相对鉴别器的方式, 使得鉴别器的输出依赖于真实样本和生成样本间的相对值, 在训练生成器时真实样本也能参与训练. 为了将鉴别器的输出限制在 [0, 1] 中, 标准生成对抗网络常常在鉴别器的最后一层使用 sigmoid 激活函数, 因此标准生成对抗网络鉴别器定义为:

$$D(x) = \text{sigmoid}(C(x)) \tag{2}$$

式中, $C(x)$ 为未经过 sigmoid 函数激励的鉴别器输出. 由于鉴别器的输出由真实样本和生成样本共同决定, 因此可以使用下述的方法构造相对鉴别器:

$$D(\tilde{x}) = \text{sigmoid}(C(x_r) - C(x_f)) \tag{3}$$

$$D_{rev}(\tilde{x}) = \text{sigmoid}(C(x_f) - C(x_r)) \tag{4}$$

式中, x_r 表示真实样本, $x_r \in P$, x_f 表示生成样本, $x_f \in Q$, $D(\tilde{x})$ 表示真实样本比生成样本更真实的概率, $D_{rev}(\tilde{x})$ 表示生成样本比真实样本更真实的概率. 经过如下推导:

$$\begin{aligned} 1 - D_{rev}(\tilde{x}) =& 1 - \text{sigmoid}(C(x_f) - C(x_r)) = \\ & \text{sigmoid}(C(x_r) - C(x_f)) = D(\tilde{x}) \end{aligned} \tag{5}$$

可得

$$\ln(D(\tilde{x})) = \ln(1 - D_{rev}(\tilde{x})) \tag{6}$$

进而可得 RSGAN 的鉴别器和生成器的目标函数:

$$L_D = -\text{E}_{(x_r, x_f)\sim(P,Q)}[\ln(\text{sigmoid}(C(x_r) - C(x_f)))] \tag{7}$$

$$L_G = -\text{E}_{(x_r, x_f)\sim(P,Q)}[\ln(\text{sigmoid}(C(x_f) - C(x_r)))] \tag{8}$$

式中, sigmoid 表示鉴别器最后一层使用 sigmoid 激活函数.

综上分析可知, 相比于 WGAN, RSGAN 生成的数据样本更稳定且质量更高, 若将 RSGAN 应用到语音转换中, 通过构造相对鉴别器的方式, 使得鉴别器的输出依赖于真实样本和生成样本间的相对值, 在训练生成器时真实样本也能参与训练, 从而改善鉴别器中可能存在的偏置情况, 使得训练更加稳定, 性能得到提升, 并且把真实样本引入到生成器的训练中, 可以加快 GAN 的收敛速度. 鉴于此, 本文提出利用 RSGAN 替换 WGAN, 构建基于变分自编码相对生成对抗网络 (Variational autoencoding RSGAN, VARSGAN) 的语音转换模型, 并引入可以充分表征说话人个性信息的 i 向量特征, 以期望在改善合成语音自然度的同时, 进一步提升转换语音的个性相似度.

2.2 i 向量的原理和提取

通过引入含有丰富说话人个性信息的表征向量, 从而提升转换语音的个性相似度是本文在上述研究基础上进一步的探索. Dehak 等[14] 提出的说话人身份 i 向量, 可以充分表征说话人的个性信息. i 向量是在高斯混合模型–通用背景模型 (Gaussian

mixture model-universal background model, GMM-UBM)[15] 超向量和信道分析的基础上提出的一种低维定长特征向量. 对于 p 维的输入语音, GMM-UBM 模型采用最大后验概率算法对高斯混合模型中的均值向量参数进行自适应可以得到 GMM 超向量. 其中, GMM-UBM 模型可以表征背景说话人整个声学空间的内部结构, 所有说话人的高斯混合模型具有相同的协方差矩阵和权重参数. 由于说话人的语音中包含了个性差异信息和信道差异信息, 因此全局 GMM 的超向量可以定义为:

$$\boldsymbol{S} = \boldsymbol{m} + T\omega \tag{9}$$

式中, $\boldsymbol{S}$ 表示说话人的超向量, $\boldsymbol{m}$ 表示与特定说话人和信道无关的均值超向量, 即通用背景模型下的超向量, T 是低维的全局差异空间矩阵, 表示背景数据的说话人空间, 包含了说话人信息和信道信息在空间上的统计分布, 也称为全局差异子空间. $\omega = (\omega_1, \omega_2, \cdots, \omega_q)$ 是包含整段语音中的说话人信息和信道信息的全局变化因子, 服从标准正态分布 $\mathrm{N}(0, I)$, 称之为 i 向量, 即身份特征 i 向量.

首先, 将经过预处理的训练语料进行特征提取得到梅尔频率倒谱系数, 将梅尔频率倒谱参数输入高斯混合模型进行训练, 通过期望最大化算法得到基于高斯混合模型的通用背景模型, 根据通用背景模型得到均值超向量 $\boldsymbol{m}$, 通过最大后验概率均值自适应得到说话人的超向量 $\boldsymbol{S}$. 同时, 根据训练所得的通用背景模型提取其鲍姆–韦尔奇统计量, 通过期望最大化算法估计获得全局差异空间矩阵 T. 最终, 通过上述求得的高斯混合模型的超向量 $\boldsymbol{S}$、通用背景模型的均值超向量 $\boldsymbol{m}$、全局差异空间矩阵 T 可以得到 i 向量. 由于上述得到的 i 向量同时含有说话人信息和信道信息, 本文采用线性判别分析和类协方差归一化对 i 向量进行信道补偿, 最终生成鲁棒的低维 i 向量.

2.3 基于 VARSGAN + i-vector 的语音转换方法

基于以上分析, 本文提出 VARSGAN + i-vector 的语音转换模型, 在解码阶段融入表征说话人个性信息的 i 向量, 将 one-hot 标签和 i 向量拼接至语义特征上构成联合特征重构出指定说话人相关的语音. 其中, i 向量含有丰富的说话人个性信息, 能够与传统编码中的 one-hot 标签相互补充, 互为辅助, 前者为语音的合成提供丰富的说话人信息, 后者作为精准的标签能够准确区分不同说话人, 相辅相成有效提升转换后语音的个性相似度, 进一步实现高质量的语音转换. 基于 VARSGAN + i-vector 模型的整体流程如图 1 所示, 分为训练阶段和转换阶段.

2.3.1 训练阶段

获取训练语料, 训练语料由多名说话人的语料组成, 包含源说话人和目标说话人; 将所述的训练语料通过 WORLD[20] 语音分析模型, 提取出各说话人语句的频谱包络、基频和非周期性特征; 利用第 2.2 节的 i 向量提取方法获得表征各个说话人个性信息的 i 向量 $\boldsymbol{i}$; 将频谱包络特征 $\boldsymbol{x}$、说话人标签 $\boldsymbol{y}$、i 向量 $\boldsymbol{i}$ 一同输入 VARSGAN + i-vector 模型进行训练, VARSGAN + i-vector 模型是由 C-VAE 和 RSGAN 结合而成, 将变分自编码器的解码器指定为 RSGAN 的生成器来优化目标函数. 原理如图 2 所示.

该模型完整的目标损失函数为:

$$J_{\text{VARSGAN + i-vector}} = L(x; \phi, \theta) + \alpha J_{\text{RSGAN}} \tag{10}$$

式中, $L(x; \phi, \theta)$ 为 C-VAE 部分的目标函数:

$$L(x; \phi, \theta) = -D_{\text{KL}}(q_\phi(z|x)||p_\theta(z)) + \mathrm{E}_{q_\phi(z|x)}[\ln p_\theta(x|z, y, i)] \tag{11}$$

式中, D_{KL} 表示 KL(Kullback-Leibler) 散度, $q_\phi(z|x)$ 表示编码网络, 该网络将频谱特征 $\boldsymbol{x}$ 编码成潜在变量 $\boldsymbol{z}$. $p_\theta(x|z, y, i)$ 表示解码网络, 将联合特征向量尽可能重构 $\boldsymbol{x}$ 就可以使式 (11) 的期望尽可能大. $p_\theta(z)$ 为潜在变量 $\boldsymbol{z}$ 的先验分布, 该分布为标准多维高斯分布. 使用随机梯度下降法来更新 C-VAE 中的网络模型参数, 其目标是 $\max\{L(x; \phi, \theta)\}$.

式 (10) 中, α是调节 RSGAN 损失的系数, J_{RSGAN} 表示 RSGAN 部分的目标函数, 由生成器和鉴别器的损失函数构成, 其中 RSGAN 的生成器中结合了表征各说话人个性信息的 i 向量 $\boldsymbol{i}$. 由式 (7) 和式 (8) 可知, 生成器网络的损失函数用 L_G 来表示:

$$L_G = -\alpha \mathrm{E}_{(x, z)\sim(p_{data}, q_\phi(z|x))} [\ln(\text{sigmoid}(D_\psi(G_\theta(z, y, i)) - D_\psi(x)))] - \mathrm{E}_{q_\phi(z|x)}[\ln p_\theta(x|z, y, i)] \tag{12}$$

式中, G_θ 表示生成器, D_ψ 表示鉴别器, θ 和 ψ 分别是生成器和鉴别器的相关参数, $G_\theta(z, y, i)$ 表示重构的频谱特征, $D_\psi(G_\theta(z, y, i))$ 表示鉴别器对重构的频谱特征判别真假.

鉴别器网络的损失函数用 L_D 表示:

$$L_D = -\mathrm{E}_{(x, z)\sim(p_{data}, q_\phi(z|x))}[\ln(\text{sigmoid}(D_\psi(x) - D_\psi(G_\theta(z, y, i))))] \tag{13}$$

添加梯度惩罚项后, 鉴别器的损失函数更新为:

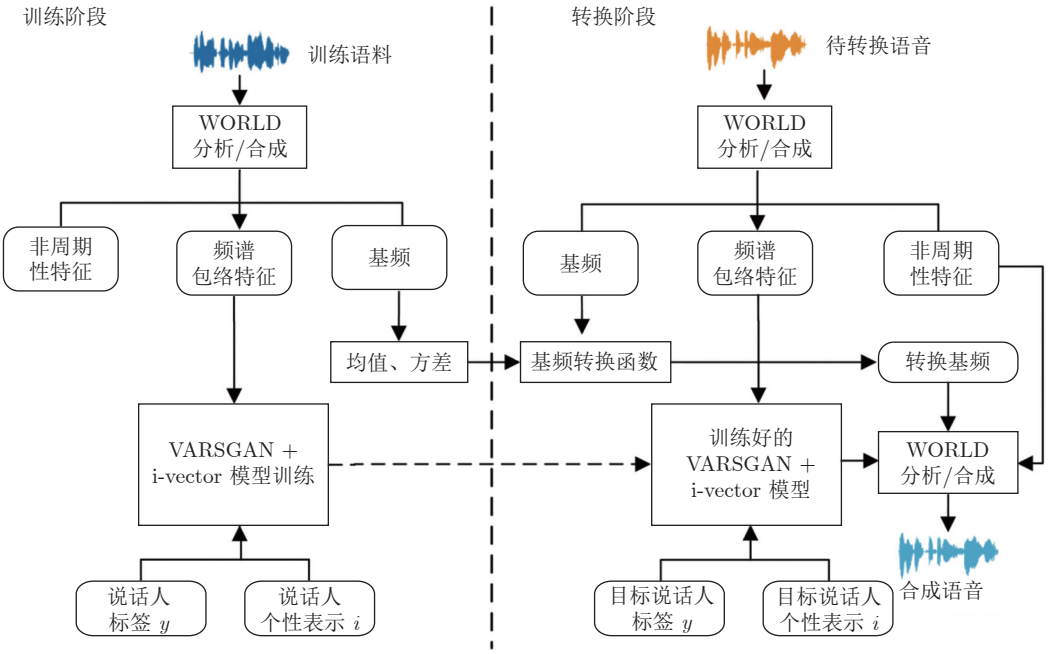


图 1 基于 VARSGAN + i-vector 模型的整体流程图

Fig. 1 Framework of voice conversion based on VARSGAN + i-vector network

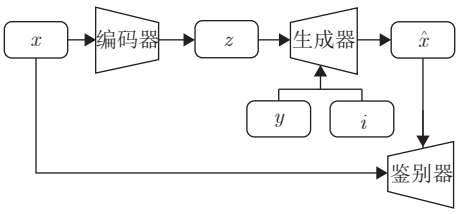


图 2 VARSGAN+i-vector 模型原理示意图

Fig. 2 Schematic diagram of VARSGAN+i-vector network

$$L_D = -\mathrm{E}_{(x,z)\sim(p_{data},q_\phi(z|x))}[\ln(\mathrm{sigmoid}(D_\psi(x)- D_\psi(G_\theta(z,y,i))))]+ \lambda \mathrm{E}_{\hat{x}\sim P_{\hat{x}}}[\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1]^2 \tag{14}$$

式中, $\mathrm{E}_{\hat{x}\sim P_{\hat{x}}}[\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1]^2$ 为梯度惩罚项, 能够加快收敛速度, 使得训练过程更为稳定[13, 21], λ 表示梯度惩罚参数. 训练过程中, 生成器网络的优化目标是 $\min\{L_G\}$, 鉴别器网络的优化目标是 $\min\{L_D\}$, 直至设置的迭代次数, 得到训练好的 VARSGAN + i-vector 网络.

构建从源说话人语音对数基频 $\ln f_0$ 到目标说话人对数基频 $\ln f_0'$ 的转换函数:

$$\ln f_0' = \mu' + \frac{\sigma'}{\sigma}(\ln f_0 - \mu) \tag{15}$$

式中, μ 和 σ 分别表示源说话人的基频在对数域的均值和标准差, μ' 和 σ' 分别表示目标说话人的基频在对数域的均值和标准差.

2.3.2 转换阶段

将待转换语料中源说话人的语音通过 WORLD[20] 语音分析模型提取出不同语句的频谱包络特征 x、基频和非周期性特征; 将频谱包络特征 x、说话人标签 y、i 向量 i 输入训练好的 VARSGAN + i-vector 模型, 从而重构出目标说话人频谱包络特征 $\hat{x}$; 通过式 (15) 表示的基频转换函数, 将源说话人对数基频 $\ln f_0$ 转换为目标说话人的对数基频 $\ln f_0'$; 非周期性特征保持不变. 将重构的目标说话人频谱包络特征 $\hat{x}$、目标说话人的对数基频 $\ln f_0'$ 和源说话人的非周期性特征通过 WORLD 语音合成模型, 合成得到转换后的说话人语音.

3 实验与分析

本实验采用 VCC2018[22] 语料库, 该语料库是由国际行业内挑战赛提供的标准数据库, 为评估不同科研团队的语音转换系统的性能提供一个通用标准. 链接为 http://www.vc-challenge.org/vcc2018/index.html, 其中的非平行文本语料库包括 4 名源说话人 (包括 2 名男性和 2 名女性), 分别是 VCC2SF3、VCC2SF4、VCC2SM3 和 VCC2SM4; 4 名目标说话人 (包括 2 名男性和 2 名女性), 分别是 VCC2TF1、VCC2TF2、VCC2TM1 和 VCC2-

TM2. 每个说话人在训练时均选取 81 句训练语音, 在转换时选取 35 句测试语音进行转换, 一共有 16 种转换情形. 将上述 8 个说话人的训练语料输入 Kaldi 语音识别工具中预训练好的模型来提取 i 向量特征, 分别得到表征上述 8 个人个性信息的各自 100 维的 i 向量.

实验系统在 Python 平台环境下实现. 在 Intel(R) Xeon(R) CPU E5-2660v4@2.00GHz, NVIDIA Tesla V100 (reva1) 的 Linux 服务器上运行, 对语料库中的 8 个说话人的语音基于 5 种模型进行客观和主观评测, 将 VAWGAN[8]作为本文的基准模型与本文提出的改进模型 VARSGAN、VAWGAN + i-vector 和 VARSGAN + i-vector 进行纵向对比, 并进一步与 StarGAN 模型[9] 进行横向对比, 这 5 种模型都是实现非平行文本条件下的多对多转换.

本文使用 WORLD 分析/合成模型提取语音参数, 包括频谱包络特征、非周期性特征和基频, 由于 FFT 长度设置为 1 024, 因此得到的频谱包络和非周期性特征均为 1 024 /2 + 1 = 513 维. 使用 VARSGAN + i-vector 模型转换频谱包络特征, 使用传统的高斯归一化的转换方法转换对数基频, 非周期性特征保持不变. 在 VARSGAN + i-vector 模型中, 所述编码器、生成器、鉴别器均采用二维卷积神经网络, 激活函数采用 LReLU 函数[23]. 图 3 为 VARSGAN + i-vector 模型网络结构图, 其中编码器由 5 个卷积层构成, 生成器由 4 个反卷积层构成, 鉴别器由 3 个卷积层和 1 个全连接层构成.

图 3 中, h、w、c 分别表示高度、宽度和通道数, k、c、s 分别表示卷积层的内核大小、输出通道数和步长, Input 表示输入, Output 表示输出, Real / Fake 表示鉴别器判定为真或假, Conv 表示卷积, Deconv 表示反卷积 (转置卷积), Fully Connected 表示全连接层, Batch Norm 表示批归一化. 实验中隐变量 z 的维度, 在借鉴基于变分自编码器模型的相关文献基础上结合实验调参, 设置为 128. 实验中 RSGAN 的损失系数 α 设置为 50, 梯度惩罚参数 λ 设置为 10, 训练批次大小设置为 16, 训练周期为 200, 学习率为 0.000 1, 最大迭代次数为 200 000. 本文模型 VARSGAN + i-vector 训练约 120 000 轮损失函数收敛, 能达到稳定的训练效果, 而基准模型耗时相对较长, 并且得到的转换性能不够稳定.

3.1　客观评价

本文选用梅尔倒谱失真距离 (Mel-cepstral distortion, MCD) 作为客观评价标准, 通过 MCD 值来衡量转换后的语音与目标语音的频谱距离[1−2], MCD 计算公式如下:

$$\mathrm{MCD}=\frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}\left(c_d-\hat{c}_d\right)^2} \tag{16}$$

式中, c_d 和 $\hat{c}_d$ 分别是目标说话人语音和转换后语音的第 d 维梅尔倒谱系数, D 是梅尔倒谱系数的维数. 计算 MCD 值时对 16 组转换情形分别选取 35 句转换语音进行统计. 图 4 为 16 种转换情形下 5 种模型的转换语音的 MCD 值对比.

由图 4 可知, 16 种转换情形下 VAWGAN、VARSGAN、VAWGAN + i-vector、VARSGAN + i-vector 和 StarGAN 模型的转换语音的平均 MCD 值分别为 5.690、5.442、5.507、5.417 和 5.583.

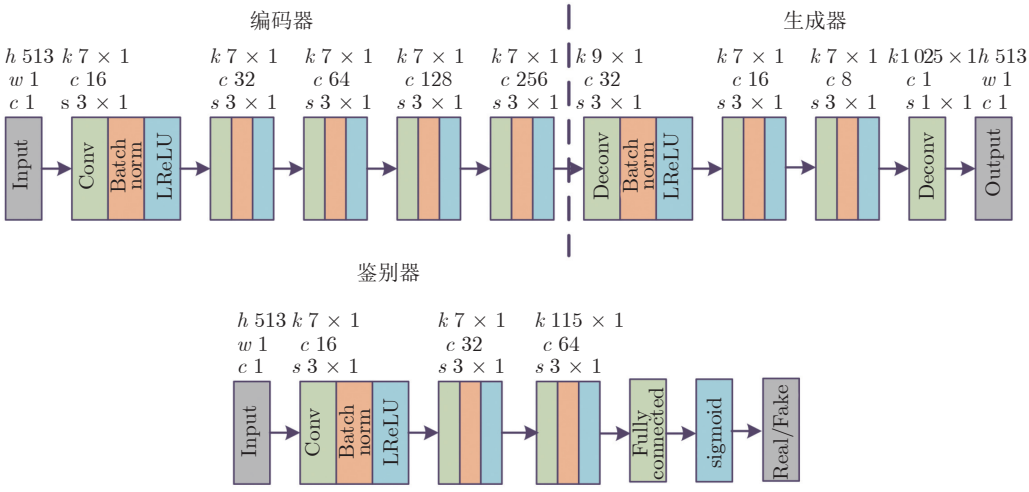


图 3　VARSGAN + i-vector 模型网络结构示意图

Fig. 3　Structure of VARSGAN + i-vector network

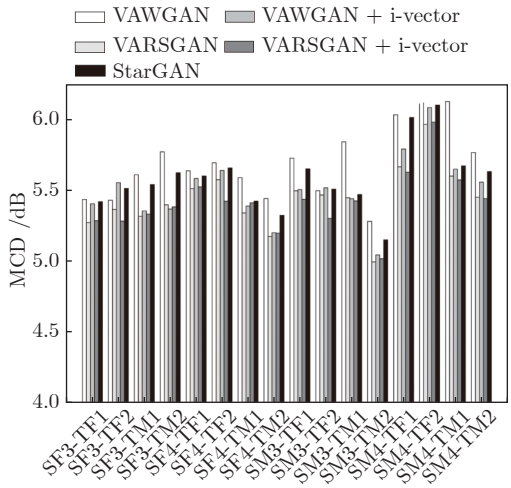


图 4 16 种转换情形下 5 种模型的转换语音的 MCD 值对比

Fig. 4 Average MCD of five models for 16 conversion cases

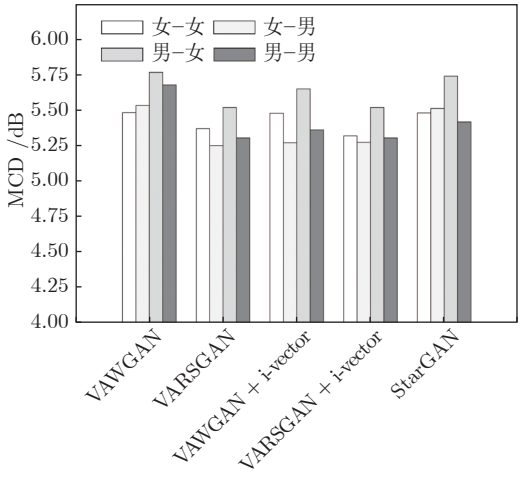


图 5 4 大类转换情形下不同模型的 MCD 值对比

Fig. 5 Comparison of MCD of different models for four conversion cases

本文提出的 3 种模型相比基准模型, 分别相对降低了 4.36%、3.22% 和 4.80%. VARSGAN + i-vector 模型相比 StarGAN 模型相对降低了 2.97%. 表明相对生成对抗网络的结合和 i 向量的引入能够显著改善转换语音的合成自然度, 有助于提升转换语音的质量.

进一步将上述 16 种转换情形按照源−目标说话人性别划分为具有统计性的 4 大类, 即同性别转换女−女、男−男和跨性别转换男−女、女−男. 4 大类转换情形下不同模型的 MCD 值对比如图 5 所示.

进一步分析实验结果可得, 本文提出的方法 VARSGAN + i-vector 在跨性别转换下, 女−男类别下的平均 MCD 值比男−女类别下的平均 MCD 值相对低 4.58%, 表明女性向男性的转换性能稍好于男性向女性的转换. 而这一现象在基准系统 VAWGAN、VARSGAN、VAWGAN + i-vector 和 StarGAN 中也不同程度地存在. 原因主要是, 语音的发音主要由基频和丰富的谐波分量构成, 即使同一语句, 由于不同性别说话人之间的基频和谐波结构存在差异较大[24−25], 会导致不同性别说话人之间的转换存在一定的性能差异.

3.2 主观评价

本文采用反映语音质量的平均意见得分 (Mean opinion score, MOS) 值和反映说话人个性相似度的 ABX 值来评测转换后语音. 主观评测人员为 20 名有语音信号处理研究背景的老师及硕士研究生, 为了避免主观倾向以及减少评测人员的工作量, 从 5 种模型的各 16 种转换情形的 35 句转换语音里面为每个人随机抽取一句, 并将语句顺序进行系统置乱. 其中在 ABX 测试中, 评测人员还需同时测听转换语音相对应的源和目标说话人的语音.

在 MOS 测试中, 评测人员根据听到的转换语音的质量对语音进行打分, 评分分为 5 个等级: 1 分表示完全不能接受, 2 分表示较差, 3 分表示可接受, 4 分表示较好, 5 分表示非常乐意接受. 本文将 16 种转换情形划分为 4 类: 男−男, 男−女, 女−男, 女−女, 4 类转换情形下 5 种模型的转换语音 MOS 值对比如图 6 所示.

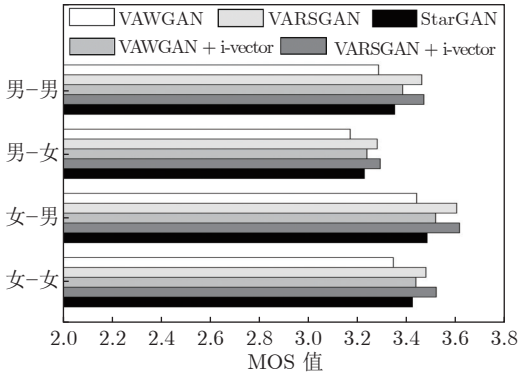


图 6 5 种模型在不同转换类别下的 MOS 值对比

Fig. 6 Comparison of MOS for different conversion categories in five models

通过分析实验结果可得, VAWGAN、VARSGAN、VAWGAN + i-vector、VARSGAN + i-vector 和 StarGAN 的平均 MOS 值分别为 3.382、3.535、3.471、3.555 和 3.446. 相比基准模型, 本文 3 种模型的 MOS 值分别相对提高了 4.52%、2.63% 和 5.12%, VARSGAN + i-vector 相比 StarGAN 提高了

3.16%, 表明本文提出的相对生成对抗网络和 i 向量的引入能够有效地改善合成语音的自然度, 提高听觉质量.

在 ABX 测试中, 评测人员测评 A、B 和 X 共 3 组语音, 其中 A 代表源说话人语音, B 代表目标说话人语音, X 为转换后得到的语音, 评测人员判断转换后的语音更加接近源语音还是目标语音. 一般将 16 种转换情形划分为同性转换和异性转换. 5 种模型在同性转换下的 ABX 测试结果如图 7 所示, 异性转换下的 ABX 测试结果如图 8 所示.

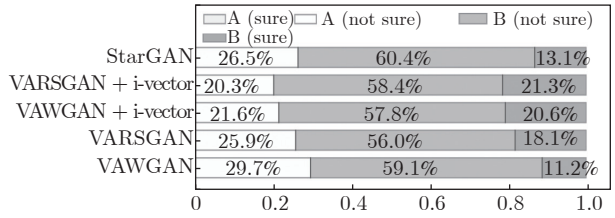


图 7　同性转换情形下 5 种模型转换语音的 ABX 图

Fig. 7　ABX test results of five models for intra-gender

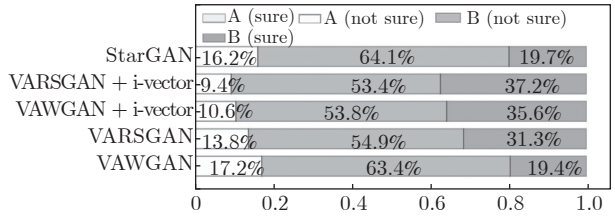


图 8　异性转换情形下 5 种模型转换语音的 ABX 图

Fig. 8　ABX test results of five models for inter-gender

图 8 中, A (sure) 表示转换语音完全确定是源说话人, A (not sure) 表示转换语音像源说话人但不完全确定, B (not sure) 表示转换语音像目标说话人但不完全确定, B (sure) 表示转换语音像目标说话人且完全确定. 在 5 种模型中, 没有评测人员认为转换后的语音确定是源说话人, 因此 A (sure) 没有得分, 即在图中没有比例显示. 在评测结果分析中, 将 B (not sure) 和 B (sure) 的比例之和作为转换语音更像目标说话人的衡量指标.

如图 7 和图 8 所示, 5 种模型在异性转换下的说话人个性相似度均优于同性转换下的说话人个性相似度, 其中在同性转换情形下,VAWGAN、VARSGAN、VAWGAN + i-vector、VARSGAN + i-vector 和 StarGAN 的 ABX 值的比例分别为 70.3%、74.1%、78.4%、79.7% 和 73.5%, 相比基准模型, 本文 3 种模型分别提升了 3.8%、8.1% 和 6.2%, VARSGAN + i-vector 相比 StarGAN 模型提升了 4.4%. 在异性转换情形下 5 种模型的 ABX 值的比例分别为 82.8%、86.2%、89.4%、90.6% 和 83.8%, 相比基准模型, 本文 3 种模型分别提升了 3.4%、6.6% 和 7.8%, VARSGAN + i-vector 相比 StarGAN 提升了 6.8%. 在同性和异性 2 种情形下, 本文提出的 3 种模型相比基准模型,平均 ABX 值分别提升了 3.6%、7.35% 和 8.6%, VARSGAN + i-vector 模型相比 StarGAN 模型提升了 5.6%, 由分析可以看出, 相对生成对抗网络的改进不仅有效地改善了合成语音的自然度, 而且也有助于说话人个性相似度的提高; 结合传统说话人编码 one-hot 实现多对多语音转换的同时, 在解码阶段融入含有丰富说话人个性信息的特征 i 向量, 能够有效增强目标说话人的个性信息, 显著提升说话人的个性相似度. 因此, 本文方法能够显著改善模型的性能.

综上所述, VARSGAN + i-vector 模型相比基准模型 VAWGAN 和 StarGAN, 平均 MOS 值相对提高了 5.12% 和 3.16%, 平均 ABX 值提升了 8.6% 和 5.6%, 表明本文提出的相对生成对抗网络和 i 向量的引入, 能够显著提高合成语音的自然度和个性相似度.

4　结束语

本文提出一种基于 VARSGAN + i-vector 的语音转换模型, 该方法利用 RSGAN 替代基准模型中的 WGAN, 改进了语音转换模型中生成对抗网络的性能, 从而生成语音自然度更好的转换语音. 进一步将 i 向量引入基于 VARSGAN 的语音转换模型, 在模型训练和转换过程中利用 i 向量表征说话人的个性信息, 有效提升转换语音的个性相似度. 充分的客观和主观实验结果表明, 相比于基准模型 VAWGAN 和 StarGAN, 本文提出的方法在有效改善转换语音的合成质量的同时, 也显著提升了说话人个性相似度, 实现了高质量的语音转换. 今后工作将研究序列到序列的语音转换, 进一步考虑韵律特征的建模和转换, 此外, 降低对训练数据量的需求以实现小样本语音转换[26] 也是课题组后续进一步研究的关注点和探索方向, 这也是该技术真正进入工业领域需要接受的挑战之一.

References

1　Godoy E, Rosec O, Chonavel T. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or non-parallel corpora. *IEEE Transactions on Audio, Speech and Language Processing*, 2011, **20**(4): 1313−1323

2　Toda T, Chen L H, Saito D, Villavicencio F, Wester M, Wu Z, et al. The voice conversion challenge 2016. In: Proceedings of the 2016 Interspeech. San Francisco, USA: 2016. 1632−1636

3　Dong M, Yang C, Lu Y, Ehnes J W, Huang D, Ming H, et al. Mapping frames with DNN-HMM recognizer for non-parallel voice conversion. In: Proceedings of the 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and

Conference (APSIPA). Hong Kong, China: IEEE, 2015. 488−494

4 Zhang M, Tao J, Tian J, Wang X. Text-independent voice conversion based on state mapped codebook. In: Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Las Vegas, USA: IEEE, 2008. 4605−4608

5 Nakashika T, Takiguchi T, Minami Y. Non-parallel training in voice conversion using an adaptive restricted boltzmann machine. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, **24**(11): 2032−2045

6 Mouchtaris A, Van der Spiegel J, Mueller P. Nonparallel training for voice conversion based on a parameter adaptation approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, **14**(3): 952−963

7 Hsu C C, Hwang H T, Wu Y C, Tsaoet Y, Wang H M. Voice conversion from non-parallel corpora using variational auto-encoder. In: Proceedings of the 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Jeju, South Korea: IEEE, 2016. 1−6

8 Hsu C C, Hwang H T, Wu Y C, Tsao Y, Wang H M. Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks. In: Proceedings of the 2017 Interspeech. Stockholm, Sweden, 2017. 3364−3368

9 Kameoka H, Kaneko T, Tanaka K, Hojo N. StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. In: Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT). Athens, Greece: IEEE, 2018. 266−273

10 Fang F, Yamagishi J, Echizen I, Lorenzo-Trueba J. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE, 2018. 5279−5283

11 Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning International Conference on Machine Learning. Sydney, Australia: ACM, 2017. 214−223

12 Wang Kun-Feng, Gou Chao, Duan Yan-Jie, Lin Yi-Lun, Zheng Xin-Hu, Wang Fei-Yue. Generative adversarial networks: The state of the art and beyond. *Acta Automatica Sinica*, 2017, **43**(3): 321−332
(王坤峰, 苟超, 段艳杰, 林懿伦, 郑心湖, 王飞跃. 生成式对抗网络GAN的研究进展与展望. 自动化学报, 2017, **43**(3): 321−332)

13 Baby D, Verhulst S. Sergan. Speech enhancement using relativistic generative adversarial networks with gradient penalty. In: Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom: IEEE, 2019. 106−110

14 Dehak N, Kenny P J, Dehak R, Dumouchelet P, Ouellet P. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, **19**(4): 788−798

15 Wang Hai-Bin, Guo Jian-Yi, Mao Cun-Li, Yu Zheng-Tao. Speaker recognition based on universal background-joint estimation (UB-JE). *Acta Automatica Sinica*, 2018, **44**(10): 1888−1895
(汪海彬, 郭剑毅, 毛存礼, 余正涛. 基于通用背景−联合估计 (UB-JE) 的说话人识别方法. 自动化学报, 2018, **44**(10): 1888−1895)

16 Matějka P, Glembek O, Castaldo F, Alam M J, Plchot O, Kenny P, et al. Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification. In: Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic: IEEE, 2011. 4828−4831

17 Kanagasundaram A, Vogt R, Dean D, Sridharan S, Mason M. I-vector based speaker recognition on short utterances. In: Proceedings of the 12th Annual Conference of the International Speech Communication Association. International Speech Communication Association (ISCA). Florence, Italy, 2011. 2341−2344

18 Zhang Yi-Ke, Zhang Peng-Yuan, Yan Yong-Hong. Data augmentation for language models via adversarial training. *Acta Automatica Sinica*, 2018, **44**(5): 891−900
(张一珂, 张鹏远, 颜永红. 基于对抗训练策略的语言模型数据增强技术. 自动化学报, 2018, **44**(5): 891−900)

19 Mao X, Li Q, Xie H, Lau R Y K, Wang Z, Smolley S P. Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2794−2802

20 Morise M, Yokomori F, Ozawa K. World: A vocoder-based high-quality speech synthesis system for real-time applications. *Ieice Transactions on Information and Systems*, 2016, **99**(7): 1877−1884

21 Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A C. Improved training of wasserstein gans. In: Proceedings of the Advances in Neural Information Processing Systems. Leicester, United Kingdom: IEEE, 2017. 5767−5777

22 Lorenzo-Trueba J, Yamagishi J, Toda T, Satio D, Villavicencio F, Kinnunen T, et al. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. In: Proceedings of the Odyssey 2018 The Speaker and Language Recognition Workshop. Les Sables d′Olonne, France: ISCA Speaker and Language Characterization Special Interest Group, 2018. 195−202

23 Maas A L, Hannun A Y, Ng A Y. Rectifier nonlinearities improve neural network acoustic models. *Computer Science*, 2013, **30**(1): 1152−1160

24 Liang Rui-Qiu, Zhao Li, Wang Qing-Yun. *Speech Signal Preprocessing* (*C++*). Beijing: China Machine Press, 2018.
(梁瑞秋, 赵力, 王青云. 语音信号处理(C++版). 北京: 机械工业出版社, 2018.)

25 Zhang Xiong-Wei, Chen Liang, Yang Ji-Bin. *Modern Speech Processing Technology and Application*. Beijing: China Machine Press, 2003.
(张雄伟, 陈亮, 杨吉斌. 现代语音处理技术及应用. 北京: 机械工业出版社, 2003.)

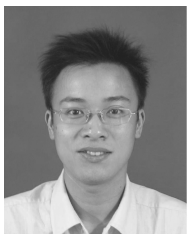
26 Chou J C, Lee H Y. One-shot voice conversion by separating speaker and content representations with instance normalization. In: Proceedings of the 2019 Interspeech. Graz, Austria, 2019. 664−668

李燕萍 南京邮电大学通信与信息工程学院副教授. 2009 年获南京理工大学博士学位. 主要研究方向为语音转换和说话人识别. 本文通信作者.
E-mail: liyp@njupt.edu.cn
(**LI Yan-Ping** Associate professor at the School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications. She received her Ph.D. degree from Nanjing University of Science and Technology in 2009. Her interest research covers voice conversion and speaker recognition. Corresponding author of this paper.)

曹 盼 南京邮电大学通信与信息工程学院硕士研究生. 2017 年获淮阴师范学院学士学位. 主要研究方向为语音转换和深度学习.
E-mail: abreastpc@163.com
(**CAO Pan** Master student at the School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications. She received her bachelor degree from Huaiyin Normal University in 2017. Her research interest covers voice conversion and deep learning.)

左宇涛 南京邮电大学通信与信息工程学院硕士研究生. 主要研究方向为语音转换.
E-mail: zuoyt@chinatelecom.cn
(**ZUO Yu-Tao** Master student at the School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications. His main research interest is voice conversion.)

张 燕 金陵科技学院软件工程学院教授. 2017 年获南京理工大学博士学位. 主要研究方向为模式识别和领域软件工程. E-mail: zy@jit.edu.cn
(**ZHANG Yan** Professor at the School of Software Engineering, Jinling Institute of Technology. She received her Ph.D. degree from Nanjing University of Science and Technology in 2017. Her research interest covers pattern recognition and domain software engineering.)



钱 博 南京电子技术研究所高级工程师. 2007 年获南京理工大学博士学位. 主要研究方向为模式识别和人工智能. E-mail: sandson6@163.com
(**QIAN Bo** Senior engineer at Nanjing Institute of Electronic Technology. He received his Ph.D. degree from Nanjing University of Science and Technology in 2007. His research interest covers pattern recognition and artificial intelligence.)