小型微型计算机系统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2023-0004
2024年8月 第8期
Vol.45 No.8 2024

基于子带谐波一致性的语音转换反取证框架研究

甘子健[1],叶登攀[1],张 健[2]

[1](武汉大学 国家网络安全学院 空天信息安全与可信计算教育部重点实验室,武汉 430072)
[2](中南大学 计算机学院 湖南省金融货币识别与自助服务平台工程技术研究中心,长沙 410083)
E-mail:2016301610273@whu.edu.cn

摘 要: 语音转换任务指的是在保持语言内容不变的情况下,将一个说话者的声音身份转换为另一个说话者.然而现有工作很少考虑针对音频取证机器分类模型进行抗检测研究,转换音频极容易被取证模型所识别.本文提出了一种具有3个子带频谱鉴别器设计的语音转换反取证框架HADV-GAN,其合成音频在具有高保真度的前提下,对语音欺骗取证模型具有反取证能力.此外,HADV-GAN无需训练额外的声码器,可以直接以原始音频波形作为输入,并以声学特征重建语音,因此可以避免使用声码器所导致的特征不匹配问题.实验结果表明,本文所提出的方法在3种主流的语音欺骗取证模型LFCC-GMM、MCG-Res2Net以及AASIST上,对比基线模型NVC-Net,在合成音频质量相当的条件下,拥有更好的反取证能力.

关 键 词: 语音转换;语音欺骗取证;子带频谱;音频反取证
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2024)08-1960-06

Voice Conversion Anti-forensic Framework Based on Subband Harmonic Consistency

GAN Zijian[1], YE Dengpan[1], ZHANG Jian[2]

[1](Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China)
[2](Hunan Province Financial Currency Identification and Self-service Platform Engineering Technology Research Center, School of Computer Science, Central South University, Changsha 410083, China)

Abstract: The voice conversion task refers to the conversion of one speaker's voice identity to another speaker while keeping the linguistic content unchanged. However, current work rarely considers anti-forensic studies for audio forensic machine classification models, where the converted audio is easily recognized by forensic models. In this paper, we propose a voice conversion anti-forensic framework HADV-GAN with three sub-band spectral discriminator, whose synthesized audio is anti-forensic to voice spoofing forensic models with high fidelity. In addition, HADV-GAN does not need to train additional vocoders, and can directly use the original audio waveform as input and reconstruct the speech with acoustic features, thus avoiding the feature mismatch problem caused by using vocoders. The experimental results show that the proposed method has better anti-forensic capability on the three mainstream voice spoofing forensic models LFCC-GMM, MCG-Res2Net and AASIST than the baseline model NVC-Net with comparable synthesized audio quality.

Keywords: voice conversion; voice spoofing forensic; sub-band spectral; audio anti-forensics

0 引 言

语音转换是音频领域中的一项映射转换任务,指的是在不改变语音中内容的条件下,将源说话人的声音转换成目标说话人的声音.语音转换技术目前被广泛地运用在特定人物语音合成,语音模仿,语音伪装以及电视电影中的语音配音等领域中.然而,目前大部分的语音转换所得到的合成音频在语音欺骗取证模型面前缺乏抗检测性,无法欺骗取证模型从而做到真正的"以假乱真".

在主流的语音转换方法中,语音信号一般通过傅里叶变换转换成梅尔频谱图,再将其作为类似图像风格迁移的任务进行转换,最后通过声码器生成转换音频.梅尔频谱图纵坐标表示频率,横坐标表示时间,颜色深浅表示对应时间对应频率的振幅.因此梅尔频谱图底部反映着音频中的低频信息,而顶部反映着音频中的较高频率信息.由于当前主流的语音转换模型往往是将整段音频的梅尔频谱图作为输入,这些方法往往忽略了局部差异,例如由于女声频率分布通常高于男声频率分布,因此在进行跨性别的语音转换时,容易出现同一子带中的频谱不一致现象,最终导致合成音频波形异常,容易被语音欺骗取证模型所鉴别.

本文针对当前语音转换模型生成语音的特点,提出了一种基于子带谐波一致性的通用语音转换反取证框架HADV-GAN.首先,以多种主流语音转换模型得到的合成音频与原始音频为训练数据,分别按照LFCC与CQCC两种具有代表性的

收稿日期:2023-01-03 收修改稿日期:2023-03-13 基金项目:国家自然科学基金面上项目(62272485)资助. 作者简介:**甘子健**,男,1997年生,硕士研究生,研究方向为语音合成与高质量语音转换、多媒体安全;**叶登攀**,男,1975年生,博士,教授,CCF会员,研究方向为大数据多媒体安全、机器学习与隐私保护;**张 健**,男,1975年生,博士,教授,CCF会员,研究方向为人工智能安全、目标检测与识别等.

音频特征来训练 3 个子带频谱鉴别器,然后通过生成对抗网络的思想对抗训练得到能具有抗检测能力的反取证音频生成器.

本文主要贡献如下:1)提出了一种语音转换模型的反取证框架 HADV-GAN,鉴别网络由 3 个工作在不同频段上的子带鉴别器构成,可以输出不易被语音欺骗取证模型所识别的音频.值得注意的是,根据我们的了解,本框架是首次尝试设计独立于语音转换模型的通用反取证框架,作为后处理模块可以适用于不同迁移方法的语音转换模型;2)融合了原始音频波形作为生成网络输入、输出的声学特征,以及 LFCC 与 CQCC 所提取的音频特征作为鉴别网络的输入.对比过往工作中常见的 MFCC 等特征提取方法,同时兼顾了面对人耳听觉感知系统的不可感知性以及面对语音欺骗取证模型的抗检测性.

1　语音转换与语音欺骗取证技术概述

1.1　语音转换

语音转换的目标是修改源说话人的声音,使其听起来像是由目标说话人发出的声音.换言之,语音转换只修改了与说话人相关的声学特征,例如语音音色、韵律以及声音强度等等,而保留了与说话人无关的语音内容信息.一般来说,无论语音转换模型采用何种方式,都可以用特征提取-特征映射-波形重构 3 个步骤来概括,如图 1 所示.

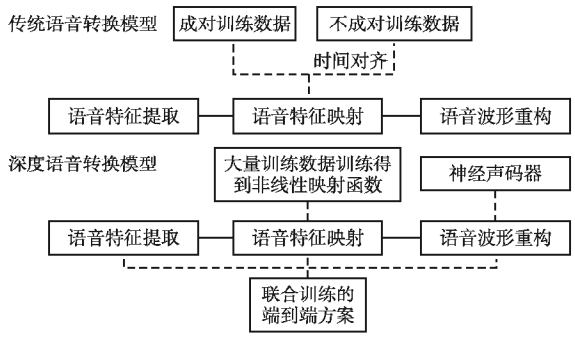

图 1　传统与深度语音转换模型对比

Fig. 1　Comparison of traditional and deep voice conversion models

语音转换有多种分类方法,按照其是否需要并行训练数据可以分为基于并行训练数据的模型,以及基于非并行训练数据的模型[1].最早的传统转换方法一般是基于并行训练数据,即需要数据集中同时包含不同说话人的相同内容音频.此外,基于并行训练数据的语音转换模型还需要对成对的音频数据进行时间对齐,然后进行特征映射,最后通过声码器完成波形重构.然而收集大量并行训练数据以及作时间对齐处理会消耗大量时间,可行性较低.

而对于非并行训练数据的语音转换来说,挑战主要在于如何将源说话人的语音与目标说话人的语音建立映射[2].其中比较具有代表性的工作例如 Erro 等人提出的 INCA 比对技术来解决非并行数据的对齐问题[3],以及基于音素后映射图(PPG)[4]的方法等等.通过时间对齐的方法,基于并行数据的方法就可以转换到基于非并行数据的领域中.

随着深度学习技术的高速发展,基于深度学习方法的语音转换逐渐成为了研究与应用的主流.如图 1 所示,深度学习方法可以在语音转换的 3 个阶段分别改进优化模型.首先,深度学习将语音特征映射模块视作为一个非线性映射函数,允许它从海量的训练数据中学习,从而极大地提升了映射效果且不易拟合.其次,针对语音波形重构环节的神经声码器不同于传统模型中根据信号处理假设设定的传统声码器,神经声码器是数据驱动且可训练的.最后,由于特征提取-特征映射-波形重构环节均可以通过训练进行优化,因此甚至可以推广得到端到端的深度语音转换模型.

常见的深度语音转换模型有 DBLSTM 和 i-vector 联合使用[5]、KL 散度和基于 DNN[6]、平均建模[7]、端到端 Blow 模型[8]等等,它们将合成语音的听觉质量提升到了一个新的高度.近年来,深度语音转换相关研究主要集中在两个方向上,分别是基于自编码器(Auto Encoder)或变分自编码器(VAE)方法和基于生成对抗网络(GAN)的方法.其中,前者以 AutoVC[9]为代表,通过风格编码器和内容解码器将语音中的说话人信息和内容信息解耦合,仅在自重建损失上进行训练,实现了非并行数据的多对多语音转换.而 CycleGAN-VC 及其衍生模型[10-12]、StarGAN-VC 及其衍生模型[13,14]和 StarGANv2-VC[15]等基于生成对抗网络的方法则是在合成自然度上取得了更好的效果.

最近的工作中,NVC-Net[16]作为一种端到端的生成对抗网络模型,直接在原始音频波形上进行语音转换,摆脱了对声码器的依赖,从而避免了特征不匹配、波形重构损失等问题.此外值得一提的是,NVC-Net 首次将反取证能力作为模型的客观评价指标,并在抗检测性上取得了目前最好的效果.

1.2　语音欺骗取证技术

语音欺骗取证是一种将目标语音区分为真实语音和欺骗语音的技术,根据语音欺骗方式可以分为 3 类,分别是语音回放攻击取证、文本到语音转换(TTS)攻击取证以及语音转换(VC)攻击取证,本节主要介绍的是语音转换攻击取证.目前有 ASVspoof 2021[17]以及 ADD 2022[18]等语音欺骗取证的挑战比赛,其中 ASVspoof 2021 挑战中的 DF 任务以及 ADD 2022 挑战都有针对语音转换攻击的取证任务.

目前主流的取证模型可以分为传统取证模型和深度取证模型.其中,传统取证模型是由声学特征提取前端,与基于机器学习的分类器后端构成;而深度取证模型一般是由声学特征提取前端,与基于深度学习的分类网络后端所构成.

语音转换模型中常使用的梅尔倒谱系数(Mel frequency cepstral coefficient,MFCC)声学特征是基于人耳听觉系统对声音频率感知非线性的特点而设计,其特点是对频率轴不均匀划分.然而对于语音欺骗取证模型而言,以梅尔倒谱系数作为语音欺骗取证的声学特征无法准确捕捉欺骗语音在特定频率下所出现的伪影.因此,目前语音欺骗取证模型一般是使用线性频谱倒谱系数(Linear frequency cepstral coefficient,LFCC)、对数常数 Q 倒谱系数(Constant Q cepstral coefficient,CQCC)以及原始波形等等作为提取声学特征的方式[19].

在传统取证模型中,当前主流的后端分类器模型是基于高斯混合模型(GMM)的方法,通过最大期望算法(Expectation maximization,EM)来迭代更新模型参数最终得到能够拟合真实语音与欺骗语音的分类模型[20].而在深度取证模型中,ASVspoof 2021 挑战中的 DF 任务上取得前三名的团队均

采用的是基于轻量卷积神经网络(Light CNN)或者基于残差网络(ResNet)的方法. 其中取得最好效果的团队采用的是 Light CNN 与 ResNet 并行输出到长短期记忆网络(LSTM)中,最后通过经验权重平均进行融合判断的方式,其在结果评估上取得了等错误率(Equal error rate)15.64% 的成绩.

2 子带频谱特征与反取证框架

2.1 子带频谱特征

在声学中,声音一般可以分解为多个正弦波,其中频率最低的正弦波即为基波,其频率为基频(Fundamental frequency, F0),其他频率的正弦波被称为谐波. 对于人耳感知系统来说,基频是区别音高(pitch)的主要因素,而谐波则可以区分不同说话人的声音音色,即语音风格. 一般来说,同一个说话人的同一段语音的基频与谐波不变.

不同频率子带的声学特征明显不同. 以 VCTK 数据集中的 p279 说话人在安静环境下的人声音频窄带线性频谱图为例,如图 2 所示,基波与低次谐波一般分布在 85hz ~ 2khz 的低频频段之间,其子带频谱图的特征为代表着共振峰的纹理清晰可见且平行连续分布. 在中频子带(2khz ~ 5khz)的频谱图中,声纹出现断续现象且较之低频子带颜色偏浅,这是因为只有部分高音处于这个频段,而且能量一般偏小. 而在高于 5khz 的高频子带部分,由于人声中的中高频谐波较少,因此频谱图中纹理也随之偏少.

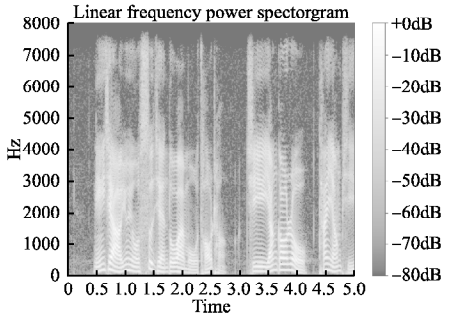

图 2 VCTK 数据集 p279_003 的窄带线性频谱图

Fig. 2 Narrowband linear spectrogram of VCTK dataset p279_003

然而,基于生成对抗网络的语音转换模型的优化目标函数一般如下:

$$\min_{G}\max_{D}V(D,G)=\mathbb{E}_{x\sim P_{data}(x)}[\log D(x)]+\mathbb{E}_{Z\sim P_{noise}(z)}[\log(1-D(G(z)))] \tag{1}$$

主流的语音转换模型都是以整张梅尔频谱图作为转换模型输入,鉴别网络仅通过梅尔谱图的整体进行判别,忽视了局部信息,难以捕捉合成语音子带中的基频或者共振峰出现的异常特征,无法保证合成语音的谐波一致性. 相对应地,这导致了生成器没有从对抗训练中学习生成具有子带特征一致性的音频. 因此最终输出的合成语音容易被语音欺骗取证模型判断为欺骗语音.

为了使生成器在对抗训练中学习到更为具体的局部特征,本文提出了子带频谱谐波一致性鉴别器,通过对反取证框架生成语音中的谐波不一致现象进行惩罚,从而达到输出具有反取证能力的语音转换音频的目的.

2.2 反取证框架

基于子带频率间的不同频谱特征,本文提出的反取证框架 HDAV-GAN 如图 3 所示,系统框架包含一个反取证音频生成器、3 个子带频谱鉴别器. 此外,为了尽可能地保留反取证语音中的语音内容信息,本框架还添加了内容损失的设计,确保输出语音信息不被丢失.

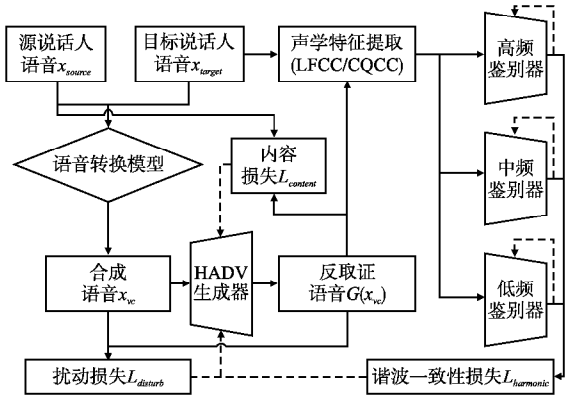

图 3 HADV-GAN 算法流程图

Fig. 3 Algorithm flow of HADV-GAN

反取证音频生成器由编码器网络和解码器网络组成,其网络结构图如图 4 所示. 其中编码器网络由 1 个卷积层、4 个下采样块以及 2 个具有 GELU 激活的卷积层组成,可以将任意长度音频序列作为输入. 其中每个下采样块由 4 个残差块以及 1 个跨步卷积层组成. 而解码器网络由 2 个具有 GELU 激活的卷积层、4 个上采样块以及 1 个卷积层组成. 与编码器网络类似,每个上采样块由 1 个转置卷积层和 4 个残差块组成. 最后通过 1 个带有 tanh 激活的卷积层生成输出音频波形.

具有相同网络架构的 3 个鉴别器被应用在 3 个不同频率的子带频谱上,本架构分别使用 LFCC 以及 CQCC 两种声学特征方式进行提取并输入到 3 个鉴别器中,再依次通过 3 个具有 GELU 激活以及实例归一化(Instance normalization)的二维卷积层,最后经过全连接层得到 0-1 真伪鉴别输出.

2.3 损失函数

HADV-GAN 作为反取证框架,所输出音频首先需要做到人耳感知系统的不可感知性,因此本框架引入了扰动损失 $L_{disturb}$:

$$L_{disturb}(G)=\mathbb{E}_{x_{vc}\sim P_X(x_{vc})}\|x_{vc}-G(x_{vc})\|_2^2 \tag{2}$$

其中 $\boldsymbol{x}_{vc}$ 与 $G(\boldsymbol{x}_{vc})$ 分别表示合成语音与反取证语音,通过 L2 范数作为 $\boldsymbol{x}_{vc}$ 与 $G(\boldsymbol{x}_{vc})$ 的失真度量,从而限制 HADV-GAN 对合成音频的扰动修改幅度,从而保证在人耳听感上不出现明显异常.

同时,为了使得反取证音频能够欺骗取证模型,本文选择在不同频段分别惩罚可能出现的反取证音频谐波不一致现象,即谐波一致性损失 $L_{harmonic}$:

$$L_{harmonic}(G,D_1,D_2,D_3)=\sum_{i=1}^{3}\left\{\begin{array}{l}\mathbb{E}_{x_{target}\sim P_X(x_{target})}[\log D_i(x_{target}^i)]\\+\mathbb{E}_{x_{vc}\sim P_X(x_{vc})}\log[1-D_i(x_{target}^i)]\end{array}\right\} \tag{3}$$

其中 $i\in[1,2,3]$,分别对应判别 85hz ~ 2khz 的低频子带

鉴别器、判别 2khz ~ 5khz 的中频子带鉴别器以及判别 5khz ~ 10khz 的高频子带鉴别器,最后将 3 个鉴别器的损失求和.

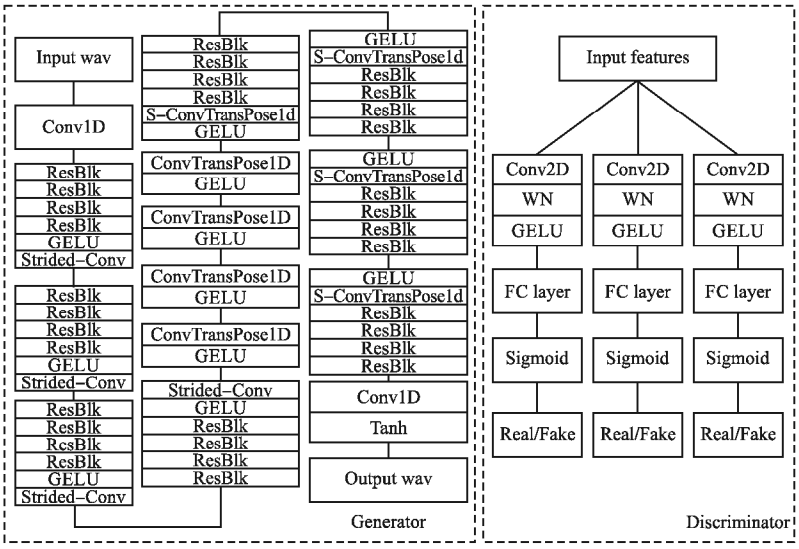

图 4　HADV-GAN 网络结构图

Fig. 4　Network structure of HADV-GAN

此外,为了确保反取证语音的内容信息不变,HADV-GAN 沿用了 AutoVC[9] 中的内容编码器 E_c,对目标说话人语音 $\boldsymbol{x}_{source}$ 与反取证语音 $G(\boldsymbol{x}_{vc})$ 的信息进行提取,并计算二者的内容损失 $L_{content}$.

$$L_{content}(G,E_c)=\mathbb{E}_{x_{source}\sim P_{data}(x_{source}),x_{VC}\sim P_X(x_{VC})} \| E_c(x_{source})-E_c(G(x_{vc}))\|_1 \tag{4}$$

其中 $\boldsymbol{x}_{source}$ 代表输入到语音转换模型中的源说话人语音,通过 L1 范数作为 $\boldsymbol{x}_{source}$ 与 $G(\boldsymbol{x}_{vc})$ 的内容距离度量,从而减少语音中的内容信息损失.

最终,本框架中所提出的总体损失函数 $L_{HADV\text{-}GAN}$ 由扰动损失 $L_{disturb}$、谐波一致性损失 $L_{harmonic}$ 和内容损失 $L_{content}$ 组成:

$$L_{HADV\text{-}GAN}=\lambda_1 L_{disturb}+\lambda_2 L_{harmonic}+\lambda_3 L_{content} \tag{5}$$

3　实验结果与分析

本文所提出的 HADV-GAN 由 Pytorch 实现,并在 NVIDIA GeForce GTX 2080Ti GPU 平台上进行训练. 训练时使用Adam优化器,学习率为10e-4,批大小为8. 超参数λ_1设置为 1,λ_2 为 0.5,λ_3 为 0.1. 具体实验环境如表 1 所示.

表 1　实验环境

Table 1　Experimental environment

实验环境	版本与型号
操作系统	Windows10 LTSC
CPU	Intel Core i7-8700K
GPU	NVIDIA GeForce GTX 2080Ti
深度学习框架	Pytorch 1.7.1
CUDA	CUDA 11.1
集成开发环境	Visual Studio Code

3.1　语音转换音频数据集

本文选取了 AutoVC、StarGANv2-VC 以及 NVC-Net 模型作为基准模型进行语音转换,将转换语音作为本文所提出反取证框架的输入.

对于上述 3 种语音转换模型,统一选取 VCTK 数据集中共 20 位说话人(的语音,将采样率设置为 22050hz,并按照 9:1 的比例分为训练集和测试集进行训练. 3 种语音转换模型均沿用作者开源项目的默认模型与超参数设置,并训练至拟合. 随后每一个语音转换模型均随机选取训练集外的 3 位男说话人和 3 位女说话人,分别作为目标说话人和源说话人,排列组合得到 30 组转换音频,共计 37813 段音频.

3.2　取证模型

本文选取了 3 种语音欺骗取证模型作为对比,如表 2 所示. LFCC-GMM 由 Tak 等人[21] 提出,通过一组简单的分类器,其中每个分类器通过手动调整以检测不同的欺骗攻击,并通过非线性融合评分,最终可以获得比依赖复杂神经网络的模型更好的表现. 2021 年 Li 等人提出的 MCG-Res2Net[22] 通过在 Res2Net 特征组的连接中启用通道门控机制,根据输入动态选择通道特征,以抑制不太相关的通道并增强检测的泛化性. 而 Jung 等人提出的 AASIST[23] 通过一种异构堆叠图注意力层,使用异构注意机制和堆栈节点对时域和频域的伪影进行建模,是目前语音欺骗取证任务中表现最好的模型.

表 2　3 种语音欺骗取证模型结构

Table 2　Major voice spoofing forensic model structure

取证模型	特征提取前端	分类后端
LFCC-GMM	LFCC	GMM
MCG-Res2Net	CQT	MCG-Res2Net
AASIST	Raw waveform	AASIST

3.3　反取证能力

本文采用两个客观指标作为反取证能力的评价方式. 首先是在 ASVspoof 挑战中常见的等错误率(EER). 对于语音转换的合成语音来说,等错误率越高说明有取证模型更容易混淆真实语音和合成语音. 因此,等错误率可以作为模型抗检测性的客观指标. 为了测试 HADV-GAN 在不同输入条件下的抗检测性,本节分别对比了 3 种语音转换模型经过 HADV-

GAN 后处理合成的等错误率,同时与原始模型相比较. 其中取证模型的输入为 400 段合成语音与 400 段真实语音. 结果如表 3 中所示,通过对比原始模型与原始模型加上 HADV-GAN 后处理的等错误率,可以看到在 3 种不同的语音转换模型下,经过 HADV-GAN 框架合成的语音出现了不同程度上的等错误率升高现象. 同时,AutoVC 与 StarGANv2-VC 模型在经过 HADV-GAN 后处理后,在等错误率上追平甚至是反超了目前抗检测性最强的 NVC-Net 原始模型. 这说明了 HADV-GAN 合成语音能够欺骗取证模型,具有较强的抗检测性.

表 3 HADV-GAN 抗检测性实验等错误率

Table 3 Equal error rate of anti-detection experiment of HADV-GAN

转换模型	原始模型			原始模型 + HADV-GAN		
	LFCC-GMM	MCG-ResNet	AASIST	LFCC-GMM	MCG-ResNet	AASIST
AutoVC	37.9%	32.2%	24.3%	45.7%	44.7%	41.5%
StarGANv2-VC	33.2%	35.4%	31.5%	47.4%	54.3%	**49.5%**
NVC-Net	47.2%	39.7%	38.2%	**57.1%**	**58.9%**	47.2%

然而,仅以抗检测性作为反取证能力的唯一指标是不准确的,因为抗检测性同时受到假阳率和假阴率的影响. 而对于反取证框架来说,往往主要关注的是生成语音能否欺骗取证模型,即合成语音能否被识别成真实语音. 同时,这也可以解释为何在等错误率指标上经过 HADV-GAN 处理的音频提升不够明显. 所以本文提出了一个用于衡量合成语音反取证能力的客观指标——平均欺骗得分(average spoof score). 具体操作是将 3 个取证模型的输出得分归一化至[0,1]之间,然后计算 800 段合成语音的平均得分,平均欺骗得分越高说明合成语音越容易被识别为真实语音. 最终结果如表 4 中所示,相比等错误率指标,平均欺骗得分的提升更为明显,3 种模型在加入 HADV-GAN 框架后在平均欺骗得分上均取得了显著提高,均超过了原始模型上的最好表现. 此项指标更加说明了本框架生成的语音在反取证能力上表现优异,对于取证模型来说具有极强的欺骗性.

表 4 HADV-GAN 抗检测性实验平均欺骗得分

Table 4 Average spoof score of anti-detection experiment of HADV-GAN

转换模型	原始模型			原始模型 + HADV-GAN		
	LFCC-GMM	MCG-ResNet	AASIST	LFCC-GMM	MCG-ResNet	AASIST
AutoVC	0.131	0.156	0.139	0.449	0.468	0.424
StarGANv2-VC	0.156	0.154	0.104	0.467	0.471	0.462
NVC-Net	0.240	0.136	0.158	**0.492**	**0.516**	**0.508**

同时,为了探究本框架中所提出框架各部分的有效性,本文采用消融实验以分别验证 HADV-GAN 框架中 3 种鉴别器、内容损失对合成音频反取证能力的影响,具体的实验设置如表 5 中所示.

从不同设置的网络结构在 3 种取证模型上的表现可以看出,只使用单个鉴别器的网络在平均欺骗得分上分别下降了 38.8%、56.0% 以及 51.2%,无内容损失的网络则是分别下降了 29.3%、28.3% 以及 8.9%. 这说明了多鉴别器和内容损失的设计都能提升 HADV-GAN 框架的反取证能力,而多鉴别器的提升效果优于内容损失的提升效果. 我们分析这是因为受益于 3 个子带鉴别器的设计,相比于单鉴别器网络,HADV-GAN 中的鉴别网络可以在多频段下协同训练生成网络,从而合成更具有反取证能力的音频. 此外,添加内容损失也能带来一定的性能提升,这是由于内容损失可以有效限制合成反取证音频过程中的内容信息丢失. 从上述结果可以推断,在引入多子带鉴别器、内容损失等模块后,HADV-GAN 在 3 种取证模型下的平均欺骗得分得到了显著提升,这说明了这两种模块的有效性.

表 5 HADV-GAN 消融实验的平均欺骗得分

Table 5 Average spoof score of HADV-GAN ablation experiment

网络结构	LFCC-GMM	MCG-ResNet	AASIST
NVC-Net	0.240	0.136	0.158
NVC-Net + HADV-GAN(无内容损失)	0.348	0.370	0.463
NVC-Net + HADV-GAN(单个鉴别器)	0.301	0.227	0.197
NVC-Net + HADV-GAN(无内容损失 + 单个鉴别器)	0.279	0.193	0.153
NVC-Net + HADV-GAN	**0.492**	**0.516**	**0.508**

3.4 语音质量

对于语音转换反取证框架的评价,除了对于生成语音的抗检测性评估,还需要评估合成语音质量. 本文采用合成语音的相似度平均意见值(Similarity Mean Opinion Score)以及自然度平均意见值(Natural Mean Opinion Score)作为语音质量的主观指标.

为了检验不同条件下的语音质量情况,我们将转换音频分为 4 组,分别是男性源说话人转换到男性目标说话人(M2M)、男性源说话人转换到女性目标说话人(M2F)、女性源说话人转换到男性目标说话人(F2M)以及女性源说话人转换到女性目标说话人(F2F). 实验分别对比 AutoVC、StarGANv2-VC、NVC-Net 原始模型以及采用了 HADV-GAN 作为反取证后处理的相似度以及自然度.

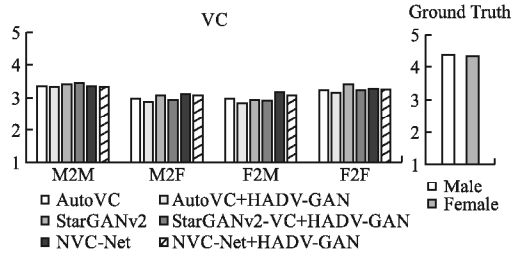

图 5 语音质量相似度对比

Fig. 5 Voice quality similarity comparison

图 5 绘制了 3 种语音转换原始模型与添加 HADV-GAN 框架的相似度平均意见值结果. 从图中结果可以看到,相对于原始模型,添加了 HADV-GAN 的模型在相似度上出现了约 1.57% 的轻微质量下降,特别是在跨性别转换场景(即 M2F 和 F2M)下较为明显. 我们认为这是由于 HADV-GAN 仅以语

音转换模型的合成音频作为输入,缺少目标说话人语音作为风格参考,导致了输出音频的相似度略有下降.

自然度平均意见值的结果如图 6 所示,在这部分指标上 HADV-GAN 取得了更好的效果. 尤其是在跨性别语音转换的条件下,在加入子带鉴别器的设计后,HADV-GAN 生成的语音克服了传统语音转换模型中存在的音调、音色异常等问题,因此在自然度上明显优于原始模型.

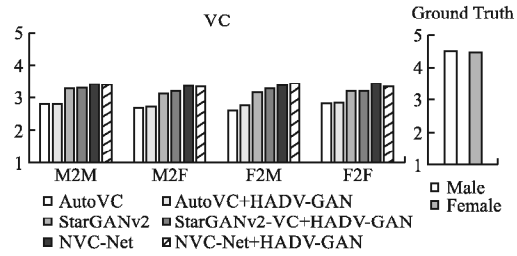

图 6　语音质量自然度对比

Fig. 6　Voice quality natural comparison

从上述的实验结果可以看到,在 4 种转换情况下,无论是在相似度还是自然度的平均意见值,采用 HADV-GAN 作为后处理的音频质量基本保持不变,没有出现语音质量明显下降的情况,甚至在自然度上还略有提升. 这说明本文所提出的反取证框架不会损害语音转换模型的语音质量.

4　总　结

本文提出了一种对转换音频谐波不一致性现象进行惩罚并生成具有抗检测性、高保真性音频的反取证框架 HADV-GAN,它通过 3 个鉴别器分别对音频中的低频、中频以及高频中可能出现的波形异常进行分类判别,然后按生成对抗网络的思想生成反取证音频. HADV-GAN 作为后处理模块,可以独立添加在当前学术界主流的语音转换模型上,具有独立性与通用性.

此外,不同于其他基于生成对抗网络的音频合成方法,本框架的反取证音频生成器直接在原始波形上进行编码、解码处理,从而避免了将输入音频转换为梅尔频谱图和声码器根据梅尔频谱图重构输出音频这两个过程中所带来的音频异常问题.

有关音频的反取证能力实验结果表明,本文所提出的方法相对于当前语音转换基线模型的有效性和优越性. 此外,对于合成音频的主观评测指标说明了本框架所生成音频在自然度上相对于基线模型略有提高,而在相似度上则有轻微的下降,这也是我们下一步研究的优化目标. 总的来说,HADV-GAN 合成音频在人耳感知基本不变,甚至略有提升的情况下,在语音欺骗取证模型上取得了显著的反取证效果.

References:

[1] Mohammadi S H, Kain A. An overview of voice conversion systems[J]. Speech Communication, 2017, 100(88): 65-82.
[2] Sisman B, Yamagishi J, King S, et al. An overview of voice conversion and its challenges: from statistical modeling to deep learning[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29(1): 132-157.
[3] Erro D, Moreno A, Bonafonte A. INCA algorithm for training voice conversion systems from nonparallel corpora[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2009, 18(5): 944-953.
[4] Sun L, Li K, Wang H, et al. Phonetic posteriorgrams for many-to-one voice conversion without parallel data training[C]//IEEE International Conference on Multimedia and Expo (ICME), 2016: 1-6.
[5] Wu J, Wu Z, Xie L. On the use of i-vectors and average voice model for voice conversion without parallel data[C]//Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), IEEE, 2016: 1-6.
[6] Xie F L, Soong F K, Li H. A KL divergence and DNN-based approach to voice conversion without parallel training sentences[C]//Interspeech, 2016: 287-291.
[7] Tian X, Wang J, Xu H, et al. Average modeling approach to voice conversion with non-parallel data[C]//Odyssey, 2018: 227-232.
[8] Serrà J, Pascual S, Segura C. Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion[C]//33rd International Conference on Neural Information Processing Systems, 2019: 6793-6803.
[9] Qian K, Zhang Y, Chang S, et al. Autovc: zero-shot voice style transfer with only autoencoder loss[C]//International Conference on Machine Learning, PMLR, 2019: 5210-5219.
[10] Kaneko T, Kameoka H. Cyclegan-vc: non-parallel voice conversion using cycle-consistent adversarial networks[C]//26th European Signal Processing Conference (EUSIPCO), IEEE, 2018: 2100-2104.
[11] Kaneko T, Kameoka H, Tanaka K, et al. Cyclegan-vc2: improved cyclegan-based non-parallel voice conversion[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: 6820-6824.
[12] Kaneko T, Kameoka H, Tanaka K, et al. CycleGAN-VC3: examining and improving CycleGAN-VCs for mel-spectrogram conversion[C]//Interspeech, 2020: 2017-2021.
[13] Kameoka H, Kaneko T, Tanaka K, et al. Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks[C]//IEEE Spoken Language Technology Workshop (SLT), 2018: 266-273.
[14] Kaneko T, Kameoka H, Tanaka K, et al. StarGAN-VC2: rethinking conditional methods for starGAN-based voice conversion[C]//Interspeech, 2019: 679-683.
[15] Li Y A, Zare A, Mesgarani N. Starganv2-vc: a diverse, unsupervised, non-parallel framework for natural-sounding voice conversion[C]//Interspeech, 2021: 1349-1353.
[16] Nguyen B, Cardinaux F. Nvc-net: end-to-end adversarial voice conversion[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 7012-7016.
[17] Yamagishi J, Wang X, Todisco M, et al. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection[J]. arXiv preprint arXiv: 2109.00537, 2021.
[18] Yi J, Fu R, Tao J, et al. Add 2022: the first audio deep synthesis detection challenge[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 9216-9220.
[19] REN Y Z, LIU C Y, LIU W Y, et al. A review of research on speech forgery and detection techniques[J]. Journal of Signal Processing, 2021, 37(12): 2412-2439.
[20] GAN H L, LEI Z C, YANG Y G. Twin Bi-LSTM model for speech spoofing detection[J]. Journal of Chinese Computer Systems, 2022, 43(6): 1265-1271.
[21] Tak H, Patino J, Nautsch A, et al. Spoofing attack detection using the non-linear fusion of sub-band classifiers[J]. arXiv preprint arXiv: 2005.10393, 2020.
[22] Li X, Wu X, Lu H, et al. Channel-wise gated res2net: towards robust detection of synthetic speech attacks[C]//Interspeech, 2021: 4314-4318.
[23] Jung J, Heo H S, Tak H, et al. Aasist: audio anti-spoofing using integrated spectro-temporal graph attention networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: 6367-6371.

附中文参考文献:

[19] 任延珍, 刘晨雨, 刘武洋, 等. 语音伪造及检测技术研究综述[J]. 信号处理, 2021, 37(12): 2412-2439.
[20] 甘海林, 雷震春, 杨印根. 孪生 Bi-LSTM 模型在语音欺骗检测中的研究[J]. 小型微型计算机系统, 2022, 43(6): 1265-1271.