计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(7):12−22 [doi: 10.15888/j.cnki.csa.008641]

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

# 语音合成及伪造、鉴伪技术综述①

杨　帅, 乔　凯, 陈　健, 王林元, 闫　镔

(中国人民解放军战略支援部队信息工程大学, 郑州 450001)
通信作者: 闫　镔, E-mail: ybspace@hotmail.com

**摘　要**: 近年来随着移动智能设备的兴起, 人们越来越频繁的接触和使用语音信息, 语音伪造和鉴伪成为语音处理领域中愈加重要的技术. 本文首先梳理了语音合成系统的一般流程, 并对语音伪造领域中主要的文本到语音 (text-to-speech, TTS) 和语音转换 (voice conversion, VC) 两项技术进行系统归纳; 接着, 对语音鉴伪技术中常见的算法进行介绍和分类; 最后, 针对语音伪造和鉴伪目前存在的问题, 本文从数据、模型、训练方法以及应用场景等多个角度出发提出未来可能的发展方向.

**关键词**: 语音伪造; 神经网络; 频谱转换; 检测技术; 语音合成

引用格式: 杨帅,乔凯,陈健,王林元,闫镔.语音合成及伪造、鉴伪技术综述.计算机系统应用,2022,31(7):12–22. http://www.c-s-a.org.cn/1003-3254/8641.html

## Overview on Speech Synthesis, Forgery and Detection Technology

YANG Shuai, QIAO Kai, CHEN Jian, WANG Lin-Yuan, YAN Bin

(PLA Strategy Support Force Information Engineering University, Zhengzhou 450001, China)

**Abstract**: In recent years, with the rise of mobile intelligent devices, people contact and use voice information more and more frequently. Voice forgery and its detection have become increasingly important technologies in the field of voice processing. Firstly, this study clarifies the general process of a voice generation system and systematically summarizes the two main technologies, text-to-speech (TTS) and voice conversion (VC), in the field of voice forgery. Then, the common algorithms in voice forgery detection technology are introduced and classified. Finally, to tackle the existing problems in voice forgery and its detection, this study puts forward possible development directions from the perspectives of data, models, training methods and application scenarios.

**Key words**: voice forgery; neural network; spectrum conversion; detection technique; speech synthesis

## 1　引言

语音作为人类接受外界信息的重要来源, 在日常交流活动中扮演了不可替代的角色. 特别是近些年来随着电话、电脑、智能手机等信息设备的普及, 人们对于丰富多彩的语音服务例如语音通话、语音助手、短视频配音等需求量越来越大. 随着网络语音资源的爆发式产出和算力水平的显著跃升, 人工智能在语音处理技术方面大放异彩, 有效地满足了社会需求; 但同时, 一些不法分子利用现代语音技术进行电信诈骗或其他违法活动, 引起了人们的担忧和广泛关注.

语音伪造技术一般包含文本到语音 (text-to-speech, TTS) 和语音转换 (voice conversion, VC) 两种形式. 文本到语音是指从文本中生成自然语音[1], 通常不具有欺骗性, 常被用于手机中的语音助理、导航语音以及智能音响等. 语音转换是指将源人物语音的特定信息转换为目标人物语音, 同时保证其他属性不改变[2]. 语音

① 收稿时间: 2021-10-08; 修改时间: 2021-11-08, 2021-12-15; 采用时间: 2021-12-28; csa 在线出版时间: 2022-05-31

转换常涉及频谱和韵律两个方面的转换, 并且依赖大量的目标语音数据. 将 TTS 和 VC 结合, 可以从文字中生成具有某人声音特点的语音, 具有极强的欺骗性.

语音伪造技术的发展满足现实应用需求的同时, 也带来很多潜在的威胁. 个性化语音生成增强了软件对用户的吸引力, 如美团的提示音、高德地图的导航语音; 短视频平台的文字朗读功能方便了用户短视频的制作; 延续风格的影视配音作品可以带给观众怀旧的体验. 另一方面, 伪造语音具有破解微信、支付宝等声纹识别模块的能力, 放大了泄露隐私、损失财产等风险, 给不法分子骗取财物提供可乘之机. 因此, 如何有效的检测伪造语音成为语音处理技术发展道路上不得不直面的难题.

本文组织结构如下: 第 2 节介绍了经典的语音合成系统, 并对文本到语音和语音转换两项技术进行了系统的梳理; 第 3 节对目前流行的语音鉴伪技术进行了分类归纳; 第 4 节分析了目前语音伪造和鉴伪领域的挑战, 并对未来的发展方向进行展望.

## 2 语音合成技术

语音合成技术是利用电子计算机或其他装置模拟人说话的技术, 主要包括文本到语音和语音转换两种技术路线. 语音伪造则是语音合成的一个应用方向, 一方面语音伪造生成结果形式与语音合成一致, 另一方面语音伪造有更明确的应用目标和需求导向. 因此语音合成系统是语音伪造技术的基础, 理解语音合成的基本过程对深入研究语音伪造大有裨益.

本节主要结构如图 1 所示, 首先对语音合成系统的一般划分进行介绍, 进而对文本到语音和语音转换两类语音伪造技术进行梳理.

### 2.1 语音合成系统

如图 2 所示, 经典的语音合成系统一般由 3 个模块构成, 依次为特征分析提取、声学模型和声码器. 将原始语音输入到特征分析提取模块中提取出源特征, 经声学模型处理得到对应的目标语音特征后通过声码器得到音频输出.

具体介绍 3 个模块的功能. 特征分析提取模块可以根据任务需求提取原始输入的特征, 例如短时傅里叶变换幅度谱、基频和梅尔倒谱[3] 等. 声学模型是整个系统中的关键部分, 将原始的声学特征转化为目标的声学特征, 主要由统计学模型和深度学习模型来构建. 早期的统计学方法主要基于矢量量化和频谱映射的模型[4]、联合概率密度的高斯混合模型[5] 和隐马尔科夫模型[6,7]. 近年来声学模型中的深度学习模块使用呈现多样化、普遍化的趋势, 典型的方法有深度神经网络、卷积网络、递归神经网络、长短时记忆网络、注意力机制[8] 等, 并且单个模型中往往会使用多种不同的模块来增强模型的学习和表达能力. 声码器的作用是将声学特征重新恢复成语音信号, 不同的声学特征采用不同的声码器进行处理. 传统的声码器假设语音的生成是信号源对滤波器系统激励产生的结果[9], 近期基于深度学习的声码器[10–12] 突破了传统的规则假设, 在庞大语音数据的驱动下能够学习到更好的语音恢复能力.

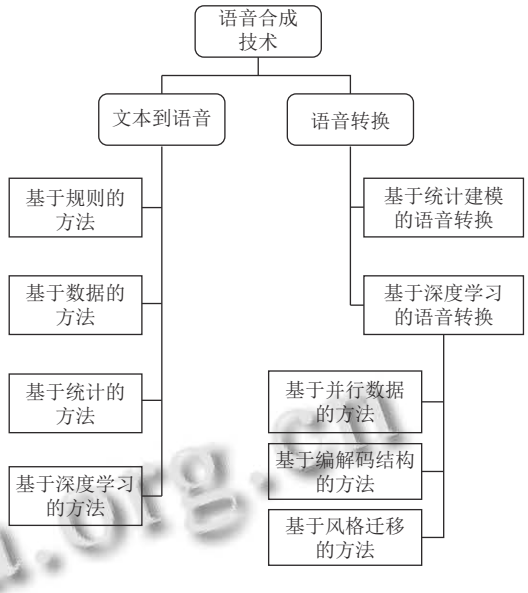


图 1 本节结构梳理


图 2 语音合成流程

### 2.2 文本到语音

文本到语音从语音合成系统的角度来看输入和输出分别对应文本和语音. 首先要对文本进行包括文本规范化、形态分析、句法分析、音素化、韵律生成等多个步骤[13] 的自然语言预处理过程. 其中文本规范化是指纠正文本中的错误, 并将缩写、缩略词等转化为完整文本; 形态分析是指将句子分割为多个单词; 句法分析是指根据分词和词义对语句构造进行拆分; 音素

化是指根据单词到发音的字典进行确定; 韵律生成是进一步确定音素的音调变化、时长、重音、节奏等特征. 进行自然语言处理之后, 需进一步进行数字信息处理. 数字信息处理方法以基于规则的方法、基于统计的方法和基于深度学习的方法为代表, 下面进行具体介绍.

2.2.1 基于规则的方法

基于规则的合成方法主要通过模拟声学物理过程来建立发声模型, 主要包括发音合成、共振峰合成. 发音合成是通过对人类声道发音过程的模拟来实现语音的合成. 为实现声道模型效果, 需要指定发音动作和对应此动作的发声器官变量变化, 例如嘴唇闭合的姿势需要下巴、下唇和上唇的协同配合[14]. 发音合成的难点在于无法对人类声道进行完全模拟, 因此合成的质量欠佳, 但在解释性和灵活性方面具有一定优势. 共振峰合成是基于源和滤波器模型的一种文本到语音声音合成方法, 包含级联型、并联型和混合型 3 种常用模型. 共振峰是指声音中能够反映人类声道特征的能量集中频段, 因此对共振峰进行合成即可实现对人类声音的模拟, 大概 20 多个不同的共振峰即可实现较好的人声恢复[15]. 该技术的优点在于可以在内存和计算速率不高的平台实现, 而缺点在于合成声音的自然性不足.

2.2.2 基于数据的方法

基于数据的方法一般需要在提前构建的语音数据库支持下进行, 可分为拼接合成和单元选择合成两类. 拼接合成通过串联提前准备的音频单元来生成语音, 这些语音单元可分为音素、半音节、单音节、双音节或者三音节. 单元长度越长, 同样的一句话中连接节点越少, 合成语音的效果越自然, 但同时内存的占用越大[16]. 并且在拼接之前需要根据语义对语音单元的韵律进行变化, 从而提高合成语音的真实度. 单元选择合成方法比拼接合成的数据库更加复杂, 因为其对相同语音单元不同韵律的数据也进行了存储, 因此占用内存更大.

基于数据的方法直接对真实语音进行操作, 从而其合成语音比基于规则的方法结果清晰度更高; 但其数据库的构建需要巨大的储存空间, 原始数据的标记也是一项非常繁琐的任务, 因此该方法实用性不足.

2.2.3 基于统计的方法

基于统计的方法主要使用隐马尔可夫模型、高斯混合模型作为基本框架, 从而利用数据中的统计规律生成语音. 典型的基于马尔可夫模型的方法[17] 对上下文相关的频谱和激励参数进行建模, 并使用期望最大化进行最大似然估计, 最后通过激励生成模块和合成滤波器模块产生语音波形. 基于高斯混合模型的方法[18] 在发音、频谱参数向量以及动态参数等的联合概率空间内进行建模, 并结合最小均方误差或者最大似然估计实现音素到语音的映射. 与基于数据的方法相比, 基于统计的方法不用建立复杂庞大的数据库, 并且可通过自适应、插值和特征声音对合成语音的特征进行改变[19]; 但此方法的合成质量距离真实语音还有差距.

2.2.4 基于深度学习的方法

深度学习主要由感知机、卷积神经网络、循环神经网络、长短时记忆网络等深度网络结构构建, 在特定的数据集上通过针对性的训练策略来完成某种学习任务. 早期深度学习与文本到语音技术的结合主要通过将深度学习作为传统语音合成的流程中一部分来体现, 也被称为非端到端的深度学习 TTS. 例如 Zen 等人[20] 利用深度神经网络替代上下文相关隐马尔可夫模型中的决策树聚类模块, 有效改善了原模型无法表达复杂的上下文依赖的问题; Kang 等人[21] 使用深度信念网络对频谱和基频等语音参数直接进行建模, 取得了比传统隐马尔科夫模型保真度更好的效果; Fan 等人[22] 采用带有双向长短时记忆模块的递归神经网络来计算语音的时间相关信息, 从而完成参数化 TTS 合成, 提升了语音合成的质量和稳定性.

随着大型语音数据集不断提出、网络结构的不断优化和计算能力的不断提升, 更多的研究重点集中到了端对端的 TTS 系统上来. 端对端系统没有诸如高斯过程此类的假设, 也没有任何关于音频的先验知识, 因此可以直接看做量化信号的非线性因果滤波器. 这种系统的好处在于模型可以更直接的收敛到数据的本质, 而不会出现不当的假设导致生成语音细节过度损失的情况. 但同时这种模型的设计也是十分困难的, 模型的好坏决定了收敛的难度和输出的质量. WaveNet[10] 是一种直接生成音频的网络模型, 能够基于来自输入文本的语言特征生成对应的语音. 如图 3 所示, WaveNet 首先将输入经过一层因果卷积, 以保证模型不会违反建模数据的顺序; 之后经多层残差模块, 残差模块中的空洞卷积使用几层即可保证指数级的感受野; 每层残差模块的输出和连接到两层 ReLU 函数, 最后通过

Softmax 层计算当前音频的量化值. WaveNet 高效的生成了超越以往模型的自然语音, 但受到感受野大小的影响, 仍存在长期依赖的问题.

同样引人注目的工作是百度提出的 DeepVoice 系列算法[23–25]. DeepVoice 按照传统 TTS 的流程用深度学习的方法构建了分离相邻音素的分割模型、字素到音素的变换模型、音素时间长度估计模型、基频预测模型和音频合成模型 5 个基本模块, 可以实现实时的文字到语音转换. DeepVoice2 是一种文本到语音的增强技术, 可以从不到半个小时的语音数据中学习到针对目标人物的高质量音频合成能力. DeepVoice3 包括编码器、解码器和转换器 3 个主要结构, 其核心在于完全卷积和注意力机制. 其中编码器是全卷积结构, 能够将文本编码成 (key, value) 组合向量; 解码器以完全卷积的结构将文本编码解码为对应于输出音频的梅尔对数幅度谱; 转换器将解码出的声学特征表示为最终的声码器参数. 第 3 代模型避免了端对端模型中的常见错误模式, 取得了更加逼真的语音效果.

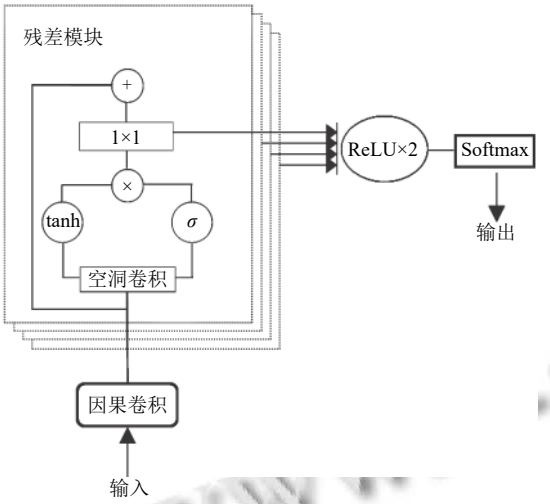


图 3 WaveNet 网络结构

谷歌公司提出的 Tacotron 文本到语音合成系统[26,27]也是该领域的一个重要算法分支. Tacotron 是一种和 DeepVoice3 类似的编解码结构, 以字符为输入生成线性光谱图, 最后转化为语音波形. 该模型不需要音素对齐, 只需给定文本和对应的音频, 因此大容量样本库的构建并不困难. Tacotron2 对 Tacotron 结构进行精简, 并采用 WaveNet 声码器替代 Griffin-Lim 从而提高了语音质量. Tacotron2 由编码器、解码器以及声码器组成, 其中编码器包括字符嵌入模块、3 层卷积和一个双向长短时记忆网络, 能够将文本转为编码特征; 解码器由局部敏感注意力模块、两层长短时网络、两个线性投影模块和一个五层卷积的 Post-Net 组成, 将编码特征转化为梅尔频谱; 最后的声码器改进自 WaveNet, 更加适合将 12.5 毫秒帧跳的梅尔谱图特征转化为时域波形.

### 2.3 语音转换

语音转换是将语音中话者语音特点进行变换的技术, 一方面用于生成具有特定人物语音特征的声音, 另一方面可以解决文本到语音技术中存在的合成效果不够自然的问题. 语音转换涉及多项语音处理技术[28], 其中语音分析是指将原始语音信号分解成某种形式的中间表达形式; 频谱转换是指对频谱中的幅度谱、对数谱、倒谱等基本参数进行映射和转换, 是目前受到广泛关注和重点解决的问题[29]; 韵律转换主要通过对基频包络进行操作, 进而实现话者的节奏、情感和情绪的转换; 语音编码和话者表征是将语音中的某类信息进行编码和压缩, 从而便于表示和减少数据量.

早期的语音转换技术一般通过统计方法建立模型, 近年来深度学习方法在本领域做出很多新的贡献. 本节将从统计建模方法和深度学习方法两个角度进行梳理, 并根据是否使用并行训练数据进一步划分.

#### 2.3.1 基于统计建模的语音转换

在语音转换中, 统计建模的方法主要有码书映射、高斯混合模型、频率扭曲、单元选择算法、INCA 算法和话者建模算法等. 其中高斯混合模型、码书映射、频率扭曲需要并行训练数据的支持, 即训练数据集中要有不同人物说的相同语音; 而单元选择算法、INCA 算法和话者建模算法则可以用非并行数据进行训练.

码书映射的方法将话者的语音个性表示为码本中的码向量, 因此语音转换的问题即可表示为找到两个码本之间的映射函数. Abe 等人[4] 通过实现矢量量化和频谱映射对音频进行变换, 在模型构建阶段得到频谱参数、功率值和音调频率的映射码本, 并在 source 码本和 target 码本之间进行映射. Matsumoto 等人[30]通过对典型频谱的估计说话人向量进行内插, 从而最小化模糊目标函数, 有效降低了矢量量化的量化误差.

基于高斯混合模型的方法[31] 不是对特定的声学特征进行操作, 而是对整个频谱包络进行转换. 此方法首先通过动态时间扭曲对源话者和目标话者语音进行对齐, 之后用高斯混合模型参数进行描述并用最小二乘

优化求解. 此方法结合高斯混合模型作为矢量量化方法的拓展, 起到了改善语音质量的效果, 但存在过渡平滑的问题. Toda 等人[5] 提出了使用动态的特征统计和考虑全局方差特性显著地缓解了过渡平滑效应; Takamichi 等人[32] 提出使用基于调制频谱修正的滤波器来减轻高斯混合模型中的过平滑问题.

基于高斯混合模型的方法通常无法保留语音的细节, 从而出现语音模糊效应. 这是因为此方法利用了平均的声学特征但缺少细节的保留, 而直接改变原始频谱的频谱扭曲方法可以较好地解决此类问题. 基于频谱扭曲的语音转换方法主要通过放大或缩小频率区间来调整波峰的位置和频谱宽度, 通过放大或缩减波峰高度来调节能量大小, 最终完成原始语言到目标语音的变换[29]. Valbret 等人[33] 最早提出使用线性多元回归和动态频率扭曲的方法, 系统被分为如图 4 所示的 3 个阶段. 第 1 阶段, 语音波形被分解为平缓的源信号和全局包络信号两个分量; 第 2 阶段, 使用 Time-Domain-PSOLA 算法改变韵律, 使用动态频率扭曲来改变频谱包络; 第 3 阶段将修改后的两个分量转换为最终音频. 此方法较好地保留了频谱的结构信息, 但其转换后的保真度存在明显的差距. 为改善此问题, 很多相关研究提出一些更加复杂的频谱扭曲技术, 例如 Sündermann 等人[34] 提出了单参数扭曲函数和多参数分段线性函数的处理方法, Tian 等人[35] 结合频率扭曲和基于样本的转换方法以保持转换后的细节.

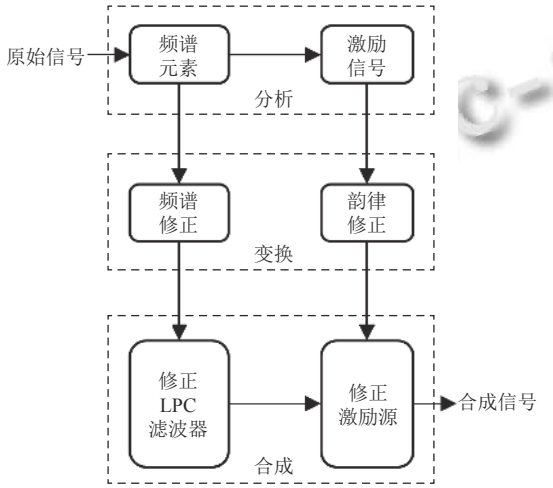


图 4 基于频谱扭曲的语音转换系统

在语音转换的实际应用中一般很难找到大量并行的数据, 因而模型训练的难度也会大大提升. 如果能把语音分解成足够细小的颗粒, 那么在数据量足够的情况下, 即便是两个人的非并行数据, 也能找到对应的细小颗粒. Duxans 等人[36] 采用单元选择技术构建伪并行样本数据库, 进而结合并行转换方法对语音进行转换. 此方法的问题在于数据库的构建较为困难.

INCA 算法[37] 结合了最近邻搜索和转换步骤, 在目标声学空间和源声学空间中分别找到对应的最近邻向量, 进一步迭代构造中间样本和目标样本的对准数据集. Stylianou 等人[38] 在高斯混合模型的基础上实现了 INCA 算法, 实验结果表明此方法与使用相当数据量的并行样本效果基本类似.

基于话者建模的转换方法是一种首先利用文本无关的语音数据建立源人物和目标任务的发音模型, 然后再进行语音转换的技术. Wu 等人[39] 将语音向量分解为语音成分和说话人特征成分, 并对说话人空间采用混合因子分析器[40] 进行因子分析, 从而细化语音转换中 JD-GMM 的协方差. 此方法大大降低了训练样本的需求量, 并且方法性能在主观和客观指标上都优于传统 JD-GMM 方法.

2.3.2 基于深度学习的语音转换

基于统计建模的语音转化方法往往面临着模型泛化能力不足的问题, 从而致使转换后的语音缺失细节、真实度不足. 深度学习的优势在于能够拟合任何复杂的函数, 因此可以更好地学习语音特点这一类的高级语义. 在数据量足够大的情况下, 深度学习的优势更加突出, 因此语音训练数据的准备也是十分重要的问题.

(1) 基于并行数据的方法

早期的深度学习模型大多只能在并行训练数据的支撑下完成语音转换任务. Xie 等人[41] 用神经网络将基音和谱特征直接进行转换, 提高了语音合成的质量. Chen 等人[42] 提出使用深度神经网络对玻尔兹曼机进行生成性训练, 并模拟源话者和目标话者的频谱包络分布, 较好地改善了生成语音中平滑效应带来的问题. 一些基于 LSTM 的工作[43,44] 建模了语音帧之间的时间相关性, 增强了转换语音的连续性和自然性.

(2) 基于编解码器结构的方法

并行数据虽然易于使用, 但制作数据库难度较大, 不利于彻底发挥深度学习强大的学习能力. 并且上一段提到的方法都是一对一的语音转换, 若目标改变还需要重新制作数据集、训练新的模型, 因此灵活性和

操作性明显不足. 借鉴计算机视觉中的思想, 说话人的转换可以看做语音风格的转换, 也就可以借鉴风格迁移中的非监督训练方法. Hsu 等人[45] 提出利用自动编码器分提取与说话人无关的信息, 并串联一个热向量代表目标说话人, 再经解码器实现具有目标话者特征的音频输出. 此方法显式的引入了说话人的身份, 但该模型没有应对未知说话人语音转换的能力. Chou 等人[46] 提出了一种通过实例规范化分离说话人和说话内容的语音转换方法, 仅需一组实例语音就可以执行. 如图 5 所示, 整个模型包括对应目标话者的话者编码器、对应源话者的内容编码器和综合两路信息的解码器. 其中话者编码器用来对说话人的声音特征进行提取, 内容编码器负责将除源说话人身份特征的内容提取出, 解码器综合两路信息并合成转换后的语音. 该模型的优点在于提供了一个真正的多对多模型, 减轻了数据和训练上的要求. 但此方法转换后的效果欠佳, 模型和训练策略都可进一步改进.

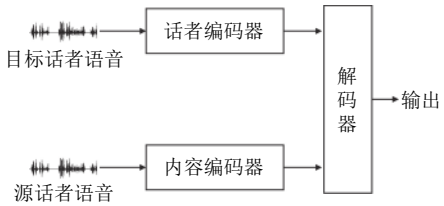


图 5 实例规范化语音转换流程

上述方法遵循逐帧转换的框架, 因此不能实现序列到序列建模持续修改的优点. 文献 [47] 构造了语音序列的识别编码器和基于神经网络的话者编码器, 能够将声音特征转换为解纠缠的语言内容和话者特征表示; 同时建立了序列到序列的解码器, 从编码器的输出中重新获取声学特征, 并进一步通过 WaveNet 声码器进行波形重构. 此方法性能接近最先进的并行训练模型, 并且在 2018 年语音转换挑战竞赛[48] 中取得冠军.

(3) 基于风格迁移的方法

除了编解码器结构, 计算机视觉中的 CycleGAN[49] 也常被用到风格转换的领域中. CycleGAN 能够在训练样本不匹配时实现两类样本之间的风格迁移, 其核心点在于循环一致性损失和对抗损失. 循环一致性损失限制了 $X$ 域的样本变换到 $Y$ 域后再经 $Y$ 域到 $X$ 域的变换也要符合 $X$ 域的分布, 从而使转换后的样本在获得其他域风格的同时仍保留原始的必要特征. 对抗损失则促进判别器的鉴别能力和生成器的生成能力同步提高, 进而提高风格迁移的效果. Kaneko 等人[50] 在原始 CycleGAN 的基础上提出 CycleGAN-VC, 引入门控卷积神经网络和身份映射损失, 在非并行数据下的效果与基于并行数据的传统方法效果相当. 为弥补转换语音在自然度和真实性方面的不足, CycleGAN-VC2[51] 通过引入两步对抗损失、2-1-2 维卷积网络和 PatchGAN, 进一步提升了模型的性能.

针对 CycleGAN-VC 不能实现 many-to-many 转换的缺点, StarGAN-VC[52] 通过扩展 CycleGAN-VC 条件设置变量实现了单一生成器下非并行多域语音转换. 但 StarGAN-VC 生成语音的质量仅仅能达到和 CycleGAN-VC 相当的水平, 与真实语音之间还存在人耳可以分辨的差别. 为了解决这一问题, StarGAN-VC2[53] 从损失函数和网络架构两个角度进行反思, 并进一步提出更先进的方法. 其中在损失函数方面, StarGAN-VC2 提出源-目标条件对抗损失函数, 促使所有转换后的数据在源和目标方面都接近真实数据; 在网络架构方面, 此方法引入一种基于调制的条件方法, 从而以领域相关的方式实现声学调制的转换.

上述风格迁移的方法可以实现说话人身份的改变, 但对于情绪的控制并没有涉及, 这使得转换语音的情感缺失. 为更好地实现语音情绪操纵, Zhou 等人[54] 提出一种采用非配对数据进行训练的语音情感转换方法. 如图 6 所示, 该方法运行时首先使用 WORLD 声码器从源语音中提取频谱特征 Sp、基频 F0 和非周期 Ap; 通过对 F0 进行连续小波变换 (continuous wavelet transform, CWT) 分析得到 10 尺度的 F0 特征; 将 F0 和梅尔谱系数 (Mel-cepstral coefficients, MCEPs) 分别输入到对应训练好的 CycleGAN 模型中进行频谱和韵律转换; 最后利用 CWT 合成逼近法重构 F0, 并由 WORLD 声码器重新构造转换后的语音.

## 3 语音鉴伪

随着语音处理技术不断提升, 伪造语音的身影在社会生活中出现的更加频繁. 一方面语音提醒、语音解锁、短视频配音等自动化语音技术方便了人们的生活, 满足了大众追求美好生活的需要. 另一方面, 语音技术的不当使用甚至滥用影响了人们的正常生活, 更有甚者给社会和国家造成恶劣影响, 成为不得不关注的安全隐患. 例如不法分子暗中收集手机用户的语音

数据, 并伪造声纹破解移动支付的密码进而对钱财进行窃取; 结合深度视觉伪造技术, 对他人声誉形象进行破坏, 一些针对各国重要人物的语音伪造甚至会引起政局和社会的动荡. 在此背景下, 如何实现合成语音的有效识别成为语音技术广泛应用不得不面临的重要问题.

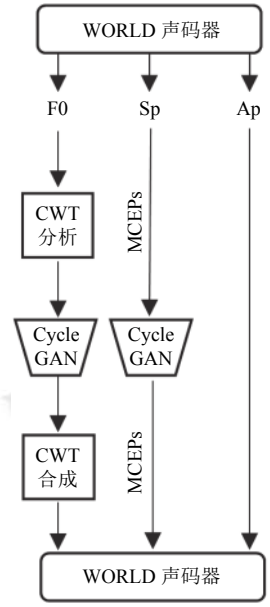


图 6 文献 [54] 工作流程

最原始的语音鉴伪方法是直接让听众判断, 然后计算平均意见分数 (mean opinion score, MOS)[55], 从而对音频的真实度和相似程度进行评判. 此种方法简单有效, 在科研学术中常被用作算法评价的重要依据, 但同时存在花费大量精力、主观评价成分多的不足. 近年来深度学习在语音伪造领域的兴起促进了语音合成往质量高、速度快的方向发展, 我们需要更客观、准确、有效的鉴伪方法来应对该领域的新变化. 目前主流的语音鉴伪技术主要包括基于特征的语音鉴伪和基于数据的语音鉴伪.

### 3.1 基于特征的语音鉴伪

基于特征的语音鉴伪一般分为 2 步, 第 1 步通过人工构建或者神经网络提取的方式获取特征, 第 2 步将特征输入分类器进行下一步的判别. Patel 等人[56] 提出基于耳蜗滤波器倒谱系数和瞬时频率变化构造帧级特征, 再借助高斯混合模型进行判别, 以此捕获跨帧的特征变化. Villalba 等人[57] 使用基于深度神经网络的频谱对数滤波器组和相对相移特征作为分类器的输入, 并使用神经网络进行特征降维后通过支持向量机进一步分类. 上述 2 种方法逐帧的提取特征, 无法在时间维度建立数据之间的联系, 因此无法应对更复杂的伪造情况.

Gomez-Alanis 等人[58] 提出一种集成轻量级卷积神经网络和递归神经网络的网络架构 LC-GRNN, 从而同时实现提取帧级特征和学习时间相关性. 如图 7 所示, 对于一段语音的频谱图, 该方法按照帧长和帧间隔逐帧提取语音内容, 并输入到轻量级门控循环单元细胞 LC-GRU 中, 在最后一个时间帧之后的最后一层输入到全连接层进行判决. 实验表明, 该方法的检测效果优于未考虑时间相关性的鉴伪模型.

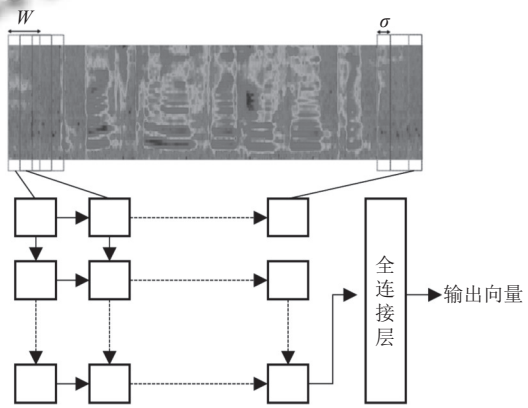


图 7 LC-GRNN 工作流程

### 3.2 基于数据的语音鉴伪

基于特征的语音鉴伪往往利用人工构造的声学特征, 对于某项单一的检测任务能达到要求. 但随着语音伪造技术的不断更迭, 互联网上各种复杂的语音情况此起彼伏, 对语音鉴伪的广泛性和集约性提出了更高的要求. 仅仅靠提取单个或几个语音特征再进行分类的做法不足以彻底挖掘庞大复杂数据的潜力, 更多研究的重点转移到了基于数据的语音鉴伪.

Jung 等人[59] 采用端到端的深度神经网络代替手工提取声学特征的过程, 同时将高分辨率的功率谱密度和频谱图输入到网络中进行处理, 在没有专业知识的辅助的情况下有效完成了检测任务. Zeinali 等人[60] 融合了具有单通道对数频谱图特征的 VGG 网络和两个不同 dropout 概率的 SincNet, 虽然能够在训练集上有很好的表现, 但泛化能力不足, 无法检测训练时看不到的攻击. 考虑基于语音合成和语音转换的麦克风级攻击和再现攻击, Monteiro 等人[61] 引入改进后的轻量

级卷积神经网络和注意力层, 从而应对不同的输入长度和某些重点部分. 该工作表明训练数据有限时使用轻型模型会导致性能的下降, 并观察到语音输入形式对不同类型的攻击检测效果有相当大的影响. Chettri 等人[62] 建立了包含卷积神经网络、卷积循环网络、Wave-U-Net、支持向量机以及高斯混合模型的集成模型, 在训练和验证期间存在攻击类型不同的情况下仍然有着不错的鉴伪效果. 该研究表明, 集成的方法有利于提高语音鉴伪模型的鲁棒性.

## 4 研究展望

尽管深度学习的不断突破和创新给语音伪造和鉴伪领域已经带来了巨大的进步, 但这些领域仍存在诸多亟待克服的困难. 对于语音伪造来说, 现有的模型大多是一对一的模型, 无法方便有效的迁移到其他未知人物的语音合成任务上; 即便是一对一的模型, 若要实现令人满意的效果也需要大量内容上高度一致的配对训练数据, 这对于数据集的构建提出了严格的要求; 同时大多数的模型专注于频率的伪造, 对于韵律的伪造并没有更多的研究. 另外, 如果要落地到现实的应用场景, 还需要考虑转换速率、模型大小以及恶劣噪声环境影响的问题. 针对于这些挑战, 语音伪造下一步应朝以下方向发展.

(1) 多对多模型. 理想的语音伪造框架应该自动的提取目标说话人的风格, 而不限于说话人的具体身份, 因此对于文本到语音任务我们只需输入模型一段文本和一段目标话者的语音, 对于语音转换任务我们只需要分别提供一段源话者和目标话者的语音. 这样的模型需要学习真正将语音的内容和风格完全分离, 因此模型的体量和训练数据集的大小应该都是有一定规模.

(2) 自监督的训练方法. 深度学习是依赖数据的技术, 因此要想提升模型的效果, 数据集必然越大越好、覆盖性越广越好. 自监督的训练方法大大降低的庞大数据集的使用难度, 显著减轻了人工标注的压力, 有利于彻底挖掘模型和数据潜力. 我们可以借鉴目前较为流行的自监督对比学习方法[63], 提出适合语音伪造任务的训练策略.

(3) 考虑韵律转换的模型. 人的语音特征可分为频率特征和韵律特征, 频率特征决定了人的音色, 而韵律特征代表人的说话的节奏、韵脚等. 现有的模型如 Tacotron 都未考虑韵律的转换, 因此合成的语音较为生硬, 下一步的模型应着重实现韵律转换.

(4) 更鲁棒的模型. 实际的语音质量并不一定良好, 很多有背景噪声、音乐等干扰, 如何消除非语音信息的干扰是该领域需要重点关注的方向.

(5) 更快更小的模型. 我们要将模型压缩的技术应用到现有语音伪造模型上, 只有模型的体量降低下来, 移动端设备才可以广泛的使用这些模型, 适用的应用场景才会越来越丰富.

同样的, 语音鉴伪领域也存在不可避免的挑战. 从本质上讲, 伪造语音检测也是一种分类任务, 因此分类中常见的问题在伪造语音检测中也会遇到. 首先, 在模型训练中看不到的伪造样本在测试阶段同样也很难检测出, 这就导致了伪造和鉴伪成了一对猫鼠游戏, 总会有新的伪造方法来躲避既有鉴伪方法的检测, 而现有的鉴伪模型又不得不不断地在训练集中纳入新的伪造样本. 其次, 即便是训练集中存在的伪造样本, 也会存在样本不均衡的问题, 导致某些特定的伪造方法难以被检测出. 针对以上问题, 语音鉴伪未来需要关注以下几点.

(1) 实际使用中关注最新的语音伪造方法, 不断更新训练集, 尽可能多的包含不同种类的样本.

(2) 采用重采样、人工产生数据样本等方法改善训练样本中数据不均衡的问题.

(3) 在集成模型方面进行更多的尝试. 现有工作表明集成模型对未知攻击有一定的检测效果, 未来构建更好的集成模型也是重点需要关注的方向.

此外, 语音伪造与视觉伪造的结合也是建立未来虚拟世界的支柱, 需要研究视觉和听觉协同时将面临的新挑战. 面对语音伪造的威胁, 一些个人账户平台可以采用多种手段进行验证, 以提升抵御未知风险的能力.

## 5 结束语

新的技术带来新的发展, 新的发展迎来新的挑战. 语音技术是现代人工智能发展的一个缩影, 给人们的生活、社会的进步带来新的活力. 同时技术被一些不法分子掌握之后, 又给社会带来了不稳定的因素. 我们要看清楚技术本身并没有好坏之分, 无论怎样都不能抵制技术的发展, 而要引导技术往好的方向去应用. 这就要求我们技术的研发者要多方面的考虑问题, 既要做推动技术腾飞的发动机, 也要守好基本底线, 做遏制

技术脱离正轨的防护栏.

## 参考文献


1 Taylor P. Text-to-speech synthesis. Cambridge: Cambridge University Press, 2009.

2 Nakashika T, Takashima R, Takiguchi T, *et al*. Voice conversion in high-order Eigen space using deep belief nets. Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon: ISCA, 2013. 369–372.

3 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, E99-D(7): 1877–1884. [doi: 10.1587/transinf.2015EDP7457]

4 Abe M, Nakamura S, Shikano K, *et al*. Voice conversion through vector quantization. Journal of the Acoustical Society of Japan (E), 1990, 11(2): 71–76. [doi: 10.1250/ast.11.71]

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

6 Morizane K, Nakamura K, Toda T, *et al*. Emphasized speech synthesis based on hidden Markov models. Proceedings of 2009 Oriental COCOSDA International Conference on Speech Database and Assessments. Urumqi: IEEE, 2009. 76–81.

7 Tokuda K, Nankaku Y, Toda T, *et al*. Speech synthesis based on hidden Markov models. Proceedings of the IEEE, 2013, 101(5): 1234–1252. [doi: 10.1109/JPROC.2013.2251852]

8 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

9 Kawahara H, Masuda-Katsuse I, de Cheveigné A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Communication, 1999, 27(3–4): 187–207. [doi: 10.1016/S0167-6393(98)00085-5]

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

11 van den Oord A, Li YZ, Babuschkin I, *et al*. Parallel wavenet: Fast high-fidelity speech synthesis. Proceedings of the 35th International Conference on Machine Learning. Stockholm: JMLR, 2018. 3918–3926.

12 Prenger R, Valle R, Catanzaro B. Waveglow: A flow-based generative network for speech synthesis. Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 3617–3621.

13 Yin ZG. A simplified overview of TTS techniques. Proceedings of the 2nd International Conference on Artificial Intelligence and Engineering Applications (AIEA 2017). 2017. 2017. 165–171.

14 Browman CP, Goldstein L. Tiers in articulatory phonology, with some implications for casual speech. Kingston J, Beckman ME. Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech. New York: Cambridge University Press, 1990. 341–376.

15 Tabet Y, Boughazi M. Speech synthesis techniques. A survey. Proceedings of the International Workshop on Systems, Signal Processing and Their Applications, WOSSPA. Tipaza: IEEE, 2011. 67–70.

16 Dutoit T. High quality text-to-speech synthesis: An overview. Journal of Electrical and Electronics Engineering, 1997, 17(1): 25–37.

17 Yoshimura T, Tokuda K, Masuko T, *et al*. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Proceedings of the 6th European Conference on Speech Communication and Technology. Budapest: ISCA, 1999.

18 Toda T, Black AW, Tokuda K. Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model. Speech Communication, 2008, 50(3): 215–227. [doi: 10.1016/j.specom.2007.09.001]

19 Kayte S, Mundada M, Gujrathi J. Hidden Markov model based speech synthesis: A review. International Journal of Computer Applications, 2015, 130(3): 35–39. [doi: 10.5120/ijca2015906965]

20 Zen HG, Senior A, Schuster M. Statistical parametric speech synthesis using deep neural networks. Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013. 7962–7966.

21 Kang SY, Qian XJ, Meng H. Multi-distribution deep belief network for speech synthesis. Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013. 8012–8016.

22 Fan YC, Qian Y, Xie FL, *et al*. TTS synthesis with

bidirectional LSTM based recurrent neural networks. Proceedings of the INTERSPEECH 2014. 2014. 1964–1968.

23 Arik SÖ, Chrzanowski M, Coates A, *et al*. Deep voice: Real-time neural text-to-speech. Proceedings of the 34th International Conference on Machine Learning. Sydney: JMLR, 2017. 195–204.

24 Gibiansky A, Arik SÖ, Diamos GF, *et al*. Deep voice 2: Multi-speaker neural text-to-speech. Proceedings of the Annual Conference on Neural Information Processing Systems 2017. Long Beach, 2017. 2962–2970.

25 Ping W, Peng KN, Gibiansky A, *et al*. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. Proceedings of the 6th International Conference on Learning Representations. Vancouver, 2018.

26 Wang YX, Skerry-Ryan RJ, Stanton D, *et al*. Tacotron: Towards end-to-end speech synthesis. Proceedings of the INTERSPEECH 2017. 2017. 4006–4010.

27 Shen J, Pang RM, Weiss RJ, *et al*. Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions. Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary: IEEE, 2018. 4779–4783.

28 Sisman B, Yamagishi J, King S, *et al*. An overview of voice conversion and its challenges: From statistical modeling to deep learning. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2021, 29: 132–157. [doi: 10.1109/TASLP.2020.3038524]

29 张雄伟, 苗晓孔, 曾歆, 等. 语音转换技术研究现状及展望. 数据采集与处理, 2019, 34(5): 753–770.

30 Matsumoto H, Yamashita Y. Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function. Journal of the Acoustical Society of Japan (E), 1993, 14(5): 353–361. [doi: 10.1250/ast.14.353]

31 赵玲丽. 基于高斯混合模型的语音转换技术研究 [ 硕士学位论文 ]. 南京: 南京邮电大学, 2011.

32 Takamichi S, Toda T, Black AW, *et al*. Modulation spectrum-based post-filter for GMM-based voice conversion. Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific. Siem Reap: IEEE, 2014. 1–4.

33 Valbret H, Moulines E, Tubach JP. Voice transformation using PSOLA technique. Speech Communication, 1992, 11(2–3): 175–187. [doi: 10.1016/0167-6393(92)90012-V]

34 Sündermann D, Strecha G, Bonafonte A, *et al*. Evaluation of VTLN-based voice conversion for embedded speech synthesis. Proceedings of the 9th European Conference on Speech Communication and Technology. Lisbon: ISCA, 2005. 2593–2596.

35 Tian XH, Lee SW, Wu ZZ, *et al*. An exemplar-based approach to frequency warping for voice conversion. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(10): 1863–1876. [doi: 10.1109/TASLP.2017.2723721]

36 Duxans H, Erro D, Pérez J, *et al*. Voice conversion of non-aligned data using unit selection. Proceedings of the TC-STAR Workshop. Barcelona, 2006.

37 Erro D, Moreno A, Bonafonte A. INCA algorithm for training voice conversion systems from nonparallel corpora. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 18(5): 944–953. [doi: 10.1109/TASL.2009.2038669]

38 Stylianou Y, Cappé O, Moulines E. Continuous probabilistic transform for voice conversion. IEEE Transactions on Speech and Audio Processing, 1998, 6(2): 131–142. [doi: 10.1109/89.661472]

39 Wu ZZ, Kinnunen T, Chng ES, *et al*. Mixture of factor analyzers using priors from non-parallel speech for voice conversion. IEEE Signal Processing Letters, 2012, 19(12): 914–917. [doi: 10.1109/LSP.2012.2225615]

40 Ghahramani Z, Hinton GE. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Toronto: University of Toronto, 1996.

41 Xie FL, Qian Y, Soong FK, *et al*. Pitch transformation in neural network based voice conversion. Proceedings of the 9th International Symposium on Chinese Spoken Language Processing. Singapore: IEEE, 2014. 197–200.

42 Chen LH, Ling ZH, Liu LJ, *et al*. Voice conversion using deep neural networks with layer-wise generative training. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1859–1872. [doi: 10.1109/TASLP.2014.2353991]

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

44 Ming HP, Huang DY, Xie L, *et al*. Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion. Proceedings of the 17th Annual Conference of the International Speech Communication Association. San Francisco: ISCA, 2016. 2453–2457.

45 Hsu CC, Hwang HT, Wu YC, *et al*. Voice conversion from non-parallel corpora using variational auto-encoder. Proceedings of 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Jeju: IEEE, 2016. 1–6.

46 Chou JC, Lee HY. One-shot voice conversion by separating speaker and content representations with instance normalization. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz: ISCA, 2019. 664–668.

47 Zhang JX, Ling ZH, Dai LR. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 28: 540–552.

48 Lorenzo-Trueba J, Yamagishi J, Toda T, *et al*. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. Proceedings of the Speaker and Language Recognition Workshop. Les Sables d'Olonne: ISCA, 2018. 195–202.

49 Zhu JY, Park T, Isola P, *et al*. Unpaired image-to-image translation using cycle-consistent adversarial networks. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2242–2251.

50 Kaneko T, Kameoka H. CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks. Proceedings of the 26th European Signal Processing Conference (EUSIPCO). Rome: IEEE, 2018. 2100–2104.

51 Kaneko T, Kameoka H, Tanaka K, *et al*. Cyclegan-VC2: Improved cyclegan-based non-parallel voice conversion. Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 6820–6824.

52 Kameoka H, Kaneko T, Tanaka K, *et al*. StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. Proceedings of 2018 IEEE Spoken Language Technology Workshop (SLT). Athens: IEEE, 2018. 266–273.

53 Kaneko T, Kameoka H, Tanaka K, *et al*. StarGAN-VC2: Rethinking conditional methods for StarGAN-based voice conversion. Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz: ISCA, 2019. 679–683.

54 Zhou K, Sisman B, Li HZ. Transforming spectrum and prosody for emotional voice conversion with non-parallel training data. Proceedings of the Speaker and Language Recognition Workshop. Tokyo: ISCA, 2020. 230–237.

55 Kumar K, Kumar R, de Boissiere T, *et al*. MelGAN: Generative adversarial networks for conditional waveform synthesis. Proceedings of the Annual Conference on Neural Information Processing Systems 2019. Vancouver, 2019. 14881–14892.

56 Patel TB, Patil HA. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. Proceedings of the INTERSPEECH 2015. Dresden, 2015. 2062–2066.

57 Villalba J, Miguel A, Ortega A, *et al*. Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge. Proceedings of the INTERSPEECH 2015. Dresden, 2015. 2067–2071.

58 Gomez-Alanis A, Peinado AM, Gonzalez JA, *et al*. A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection. Proceedings of the INTERSPEECH 2019. Graz, 2019. 1068–1072.

59 Jung JW, Shim HJ, Heo HS, *et al*. Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVspoof 2019 Challenge. Proceedings of the INTERSPEECH 2019. Graz, 2019. 1083–1087.

60 Zeinali H, Stafylakis T, Athanasopoulou G, *et al*. Detecting spoofing attacks using VGG and sincNet: BUT-omilia submission to ASVspoof 2019 challenge. Proceedings of the INTERSPEECH 2019. Graz, 2019. 1073–1077.

61 Monteiro J, Alam J, Falk TH. End-to-end detection of attacks to automatic speaker recognizers with time-attentive light convolutional neural networks. Proceedings of 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP). Pittsburgh: IEEE, 2019. 1–6.

62 Chettri B, Stoller D, Morfi V, *et al*. Ensemble models for spoofing detection in automatic speaker verification. Proceedings of the INTERSPEECH 2019. Graz, 2019. 1018–1022.

63 He KM, Fan HQ, Wu YX, *et al*. Momentum contrast for unsupervised visual representation learning. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9726–9735.

(校对责编: 牛欣悦)