

中图法分类号: TN912 文献标识码: A 文章编号: 1006-8961(2024)08-2236-33

论文引用格式: Xu Y X, Li B, Tan S Q and Huang J W. 2024. Research progress on speech deepfake and its detection techniques. Journal of Image and Graphics, 29(08):2236-2268(许裕雄, 李斌, 谭舜泉, 黄继武. 2024. 语音深度伪造及其检测技术研究进展. 中国图象图形学报, 29(08): 2236-2268)[DOI:10.11834/jig.230476]

语音深度伪造及其检测技术研究进展

许裕雄^{1,2,3}, 李斌^{1,2,3*}, 谭舜泉^{1,2,4}, 黄继武^{1,2,3}

1. 广东省智能信息处理重点实验室, 深圳 518060; 2. 深圳市媒体信息安全重点实验室, 深圳 518060;
3. 深圳大学电子与信息工程学院, 深圳 518060; 4. 深圳大学计算机与软件学院, 深圳 518060

摘要: 语音深度伪造技术是利用深度学习方法进行合成或生成语音的技术。人工智能生成内容技术的快速迭代与优化, 推动了语音深度伪造技术在伪造语音的自然度、逼真度和多样性等方面取得显著提升, 同时也使得语音深度伪造检测技术面临着巨大挑战。本文对语音深度伪造及其检测技术的研究进展进行全面梳理回顾。首先, 介绍以语音合成(speech synthesis, SS)和语音转换(voice conversion, VC)为代表的伪造技术。然后, 介绍语音深度伪造检测领域的常用数据集和相关评价指标。在此基础上, 从数据增强、特征提取和优化以及学习机制等处理流程的角度对现有的语音深度伪造检测技术进行分类与深入分析。具体而言, 从语音加噪、掩码增强、信道增强和压缩增强等数据增强的角度来分析不同增强方式对伪造检测技术性能的影响, 从基于手工特征的伪造检测、基于混合特征的伪造检测、基于端到端的伪造检测和基于特征融合的伪造检测等特征提取和优化的角度对比分析各类方法的优缺点, 从自监督学习、对抗训练和多任务学习等学习机制的角度对伪造检测技术的训练方式进行探讨。最后, 总结分析语音深度伪造检测技术存在的挑战性问题, 并对未来研究进行展望。本文汇总的相关数据集和代码可在<https://github.com/media-sec-lab/Audio-Deepfake-Detection>访问。

关键词: 语音深度伪造; 语音深度伪造检测; 语音合成(SS); 语音转换(VC); 人工智能生成内容(AIGC); 自监督学习; 对抗训练

Research progress on speech deepfake and its detection techniques

Xu Yuxiong^{1,2,3}, Li Bin^{1,2,3*}, Tan Shunquan^{1,2,4}, Huang Jiwu^{1,2,3}

1. Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen 518060, China;
2. Shenzhen Key Laboratory of Media Security, Shenzhen 518060, China;
3. College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China;
4. College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

Abstract: Speech deepfake technology, which employs deep learning methods to synthesize or generate speech, has emerged as a critical research hotspot in multimedia information security. The rapid iteration and optimization of artificial intelligence-generated content technologies have significantly advanced speech deepfake techniques. These advancements

收稿日期: 2023-07-10; 修回日期: 2023-11-10; 预印本日期: 2023-11-17

* 通信作者: 李斌 libin@szu.edu.cn

基金项目: 国家自然科学基金项目(U23B2022, U22B2047, 62272314); 广东省基础与应用基础研究基金项目(2019B151502001); 深圳市基础研究重点项目(JCYJ20200109105008228); 亚马逊云科技——2022教育部就业育人项目(20221128)

Supported by: National Natural Science Foundation of China (U23B2022, U22B2047, 62272314); Guangdong Basic and Applied Basic Research Foundation (2019B151502001); Shenzhen R&D Program (JCYJ20200109105008228); Amazon Web Service —— 2022 the Employment and Education Program of China, Ministry of Education (20221128)

have significantly enhanced the naturalness, fidelity, and diversity of synthesized speech. However, they have also presented great challenges for speech deepfake detection technology. To address these challenges, this study comprehensively reviews recent research progress on speech deepfake generation and its detection techniques. Based on an extensive literature survey, this study first introduces the research background of speech forgery and its detection and compares and analyzes previously published reviews in this field. Second, this study provides a concise overview of speech deepfake generation, especially speech synthesis (SS) and voice conversion (VC). SS, which is commonly known as text-to-speech (TTS), analyzes text and generates speech that aligns with the provided input by applying linguistic rules for text description. Various deep models are employed in TTS, including sequence-to-sequence models, flow models, generative adversarial network models, variational auto-encoder models, and diffusion models. VC involves modifying acoustic features, such as emotion, accent, pronunciation, and speaker identity, to produce speech resembling human-like speech. VC algorithms can be categorized as single, multiple, and arbitrary target speech conversion depending on the number of target speakers. Third, this study briefly introduces commonly used datasets in speech deepfake detection and provides relevant access links to open-source datasets. This study briefly introduces two commonly used evaluation metrics in speech deepfake detection: the equal error rate and the tandem detection cost function. This study analyzes and categorizes the existing deep speech forgery detection techniques in detail. The pros and cons of different detection techniques are studied and compared in depth, focusing primarily on data processing, feature extraction and optimization, and learning mechanisms. Notably, this study summarizes the experimental results of existing detection techniques on the ASVspoof 2019 and 2021 datasets in tabular form. Within this context, the primary focus of this study is to investigate the generality of current detection techniques in the field of speech deepfake detection without focusing on specific forgery attack methods. Data augmentation involves a series of transformations on the original speech data. These include speech noise addition, mask enhancement, channel enhancement, and compression enhancement, each aiming to simulate complex real-world acoustic environments more effectively. Among them, one of the most common data processing methods is speech noise addition, which aims to interfere with the speech signal by adding noise to simulate the complex acoustic environment of a real scenario as much as possible. Mask enhancement is the masking operation on the time or frequency domain of speech to achieve noise suppression and enhancement of the speech signal for improving the accuracy and robustness of speech detection techniques. Transmission channel enhancement focuses on solving the problems of signal attenuation, data loss, and noise interference caused by changes in the codec and transmission channel of speech data. Compression enhancement techniques address the problem of degradation of speech quality during data compression. In particular, the main data compression methods are MP3, M4A, and OGG. From the perspective of feature extraction and optimization, speech deepfake detection can be divided into handcrafted feature-, hybrid feature-, deep feature-, and feature fusion-based methods. Handcrafted features refer to speech features extracted with the help of certain prior knowledge, which mainly include constant-Q transform, linear frequency cepstral coefficients, and Mel-spectrogram. By contrast, feature-based hybrid forgery detection methods utilize the domain knowledge provided by handcrafted features to mine richer information about speech representations through deep learning networks. End-to-end forgery detection methods directly learn feature representation and classification models from raw speech signals, which eliminates the need for handcrafted feature extraction. This way allows the model to discover discriminative features from the input data automatically. Moreover, these detection techniques can be trained using a single feature. Alternatively, feature-level fusion forgery detection can be employed to combine multiple features, whether they are identical or different. Techniques such as weighted aggregation and feature concatenation are used for feature-level fusion. The detection techniques can capture richer speech information by fusing these features, which improves performance. For the learning mechanism, this study explores the impact of different training methods on forgery detection techniques, especially self-supervised learning, adversarial training, and multi-task learning. Self-supervised learning plays an important role in forgery detection techniques by automatically generating auxiliary targets or labels from speech data to train models. Fine-tuning the self-supervised-based pretrained model can effectively distinguish between real and forged speech. Then, adversarial training-based forgery detection enhances the robustness and generalization of the model by adding adversarial samples to the training data. In contrast to binary classification tasks, the forgery detection based on multi-task learning captures more comprehensive and useful speech feature information from different speech-

related tasks by sharing the underlying feature representations. This approach improves the detection performance of the model while effectively utilizing speech training data. Although speech deepfake detection techniques have achieved excellent performance in some datasets, their performance is less satisfactory when testing speech data from natural scenarios. Analysis of the existing research shows that the main future research directions are to establish diversified speech deepfake datasets, study adversarial samples or data enhancement methods for enhancing the robustness of speech deepfake detection techniques, establish generalized speech deepfake detection techniques, and explore interpretable speech deepfake detection techniques. The relevant datasets and code mentioned can be accessed from <https://github.com/media-sec-lab/Audio-Deepfake-Detection>.

Key words: speech deepfake; speech deepfake detection; speech synthesis (SS); voice conversion (VC); artificial intelligence-generated content (AIGC); self-supervised learning; adversarial training

0 引言

随着人工智能生成内容 (artificial intelligence-generated content, AIGC) 的兴起, 语音、图像和视频生成技术正在蓬勃发展。微软、Meta 以及百度等科技公司对语音生成技术不断投入大量的人力和物力进行研究与推广, 使得生成式语音在拟人化、真实化和自然化等方面取得长足的进步。特别是自 2022 年以来, 聊天生成预训练转换器 (chat generative pre-trained Transformer, ChatGPT) (OpenAI, 2023) 的发布以及迅速走红, 推动了语音生成领域迈向 GPT (generative pretrained Transformer) 时代。例如, 复旦大学邱锡鹏教授团队 (Zhang 等, 2023a) 提出的大型语言模型 SpeechGPT 具备跨模态对话能力, 能够实现语音与文本之间的无缝对话。Meta 公司发布通用的大规模语音文本引导生成模型 Voicebox (Le 等, 2023), 被誉为开创语音生成新纪元, 可以高效完成多语言语音合成、内容编辑和风格转换等任务。

语音深度伪造 (speech deepfake) 技术是语音生成领域的研究热点之一。它利用大量语音数据训练深度学习模型, 以学习语音的声学特征和说话人特征, 从而生成逼真的任意或特定说话人的语音。然而, 它的快速发展也带来一些安全隐患。人类听觉系统 (human audio system, HAS) 和自动说话人验证系统 (automatic speaker verification, ASV) 成为语音深度伪造技术攻击的主要对象 (Tan 等, 2021)。

听觉系统是人们感知和理解语音信息的途经。当逼真的伪造语音在社交媒体、网络平台等渠道传播时, 人们往往难以通过听觉来准确辨别出它们的真伪, 这给不法分子提供了可乘之机。随着语音深度伪造软件的使用门槛和难度逐渐降低, 伪造语音

的滥用情况变得越来越严重。恶意伪造语音事件屡见不鲜, 给个人和社会带来了严重的威胁和危害, 给新闻真实、社会信任和司法取证等方面带来了严峻的挑战 (李晓龙 等, 2021)。

ASV 系统是一种自动识别和验证说话人身份的技术, 其通过分析输入语音数据中的声学特征和说话人特征, 能够判断该语音是否属于系统内已注册的用户。然而, ASV 系统随时都可能遭受各种攻击和安全威胁。例如, 当不法分子利用逼真的伪造语音绕过 ASV 系统的说话人身份验证时, 可以获取目标用户的访问权限, 并冒充目标用户访问其个人的智能设备、银行账户和社交媒体账户等, 这对基于语音的安全访问控制和个人隐私保护造成了严重的威胁。全球社会各界对语音深度伪造技术的快速迭代高度关注, 人们对“AI (artificial intelligence) 诈骗潮”爆发的担忧日益加剧 (http://gd.news.cn/newscenter/2023-06/09/c_1129680523.htm)。为了应对这一问题, 国内外各界正积极探索“以 AI 制 AI”的反欺骗技术研究, 旨在减少或消除语音深度伪造技术带来的危害。

以语音合成 (speech synthesis, SS) 和语音转换 (voice conversion, VC) 为代表的语音深度伪造技术已经引起了学术界广泛关注和研究。其中, 语音合成技术旨在生成与真实语音高度逼真的语音。而语音转换技术要求在保持语义内容不变的情况下替换原始语音的说话人。为了防范潜在的安全风险, 一些研究学者积极致力于语音深度伪造检测技术的研究, 不断提升伪造语音的检测性能。语音深度伪造检测技术利用深度网络模型学习伪造语音的合成模式和特征, 从而区分语音的真伪。图 1 展示了通过科学引文索引 (Web of Science) 网站检索 2015 年至 2023 年第 2 季度期间以“speech synthesis”、“voice conversion”、“speech deepfake detection”和“fake audio detection”等主题词的科学文献统计结果。显

而易见的是,语音深度伪造检测文献数量远少于语音深度伪造生成文献数量。这一现象表明,与语音深度伪造技术的发展相比,作为“守护者”的语音深度伪造检测技术的研究进展相对滞后。尽管语音深度伪造检测已经取得了一些显著的研究成果,但语音深度伪造检测技术仍然面临许多挑战。如何研究更准确、更有效、更通用的语音深度伪造检测技术,为信息安全保驾护航,已成为学术界和工业界亟待解决的问题。

本文从语音深度伪造及其检测技术的研究进展出发,对主流技术框架和发展历程进行梳理,分析各类技术的优势与不足,以及对本领域未来的研究方向进行讨论和展望。本文与现有综述文献对比如表 1 所示。汇总情况表明,现有的综述文献未从数据增强、特征提取与优化以及机器学习机制方面分析语音深度伪造检测技术的研究现状,本文将从这 3 个方面进行深入探讨。为了方便研究者快

速开展相关研究工作,本文整理并上传了相关数据集的下载链接和伪造检测模型的开源代码,可在<https://github.com/media-sec-lab/Audio-Deepfake-Detection>访问。

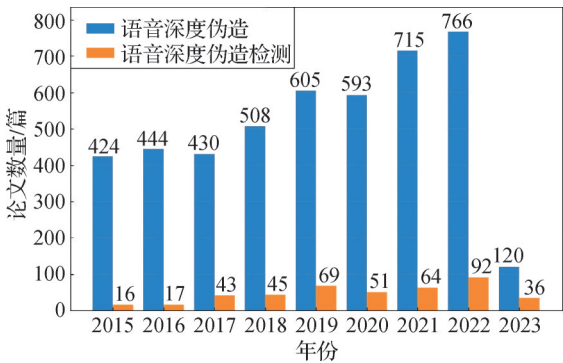


图1 语音深度伪造及其检测技术的学术论文统计分析 (2015年至2023年第2季度,数据来源于Web of Science)
Fig. 1 Statistical analysis of academic papers on speech deepfake and its detection (2015 to the second-quarter of 2023, Web of Science)

表1 与现有综述文献的对比

Table 1 Comparison of the existing survey and review papers

文献	语音深度 伪造技术		语音深度伪造检测技术										
			数据增强				特征提取和优化				学习机制		
	语音 合成	语音 转换	语音 加噪	掩码 增强	信道 增强	压缩 增强	手工 特征	混合 特征	端到 端	特征 融合	自监督 学习	对抗 训练	多任务 学习
Sahidullah 等人(2019)	√	√	×	×	×	×	√	×	√	×	×	×	×
Bevinamarad 和 Shirlondonkar(2020)	√	×	×	×	×	×	√	×	×	×	×	×	×
Kamble 等人(2020)	√	√	×	×	×	×	√	×	√	×	×	×	×
陶建华等人(2020)	√	√	×	×	×	×	√	×	√	×	×	×	×
张雄伟等人(2020)	√	√	×	×	×	×	√	×	√	×	×	×	×
任延珍等人(2021)	√	√	×	×	×	×	√	×	√	×	×	√	×
Tan 等人(2021)	×	×	×	×	×	×	√	√	√	√	×	×	×
Almutairi 和 Elgibreen(2022)	√	×	×	×	×	×	√	×	√	×	×	×	×
Mittal 和 Dua(2022)	√	√	×	×	×	×	√	×	√	×	×	×	×
杨帅等人(2022)	√	√	×	×	×	×	√	×	√	×	×	×	×
Dixit 等人(2023)	√	√	×	×	×	×	√	×	√	×	×	×	×
Ranjan 等人(2023)	√	√	×	×	×	×	√	×	√	×	×	×	×
Khanjani 等人(2023)	√	√	×	×	×	×	√	×	√	×	×	×	×
本文	√	√	√	√	√	√	√	√	√	√	√	√	√

注:“√”表示文献中包含此项,“x”表示文献中无此项。

1 语音深度伪造技术

语音深度伪造技术利用神经网络和机器学习方

法生成逼真度高、自然度高的伪造语音,这给人类听觉系统(HAS)或自动说话人验证系统(ASV)带来了严峻的挑战。目前,语音深度伪造技术主要以语音合成(SS)和语音转换(VC)为主。其中,语音合成,

也称为文本到语音 (text-to-speech, TTS), 是将给定的文本内容合成为任意或特定说话人风格的语音, 从而实现文本到语音的映射。语音转换旨在将源说话人的语音特征转换为目标说话人特征, 实现语音到语音之间的映射。本节主要对语音合成和语音转换的研究现状进行综述。

1.1 语音合成

如图 2 所示, 依据语音合成技术的发展流程可

以分为三步法、两步法和端到端方法。在三步法中, 主要包括文本分析、声学模型和声码器 3 个模块。其中, 文本分析是将输入的文本内容进行规范化、分词以及断句等生成与之相对应的音素序列等相关信息; 声学模型主要是学习语言学表征和相应的声学特征之间的映射关系; 声码器将声学特征转化为语音信号, 从而实现合成语音。表 2 根据发展流程以及生成模型归纳了一些主要的语音合成方法。

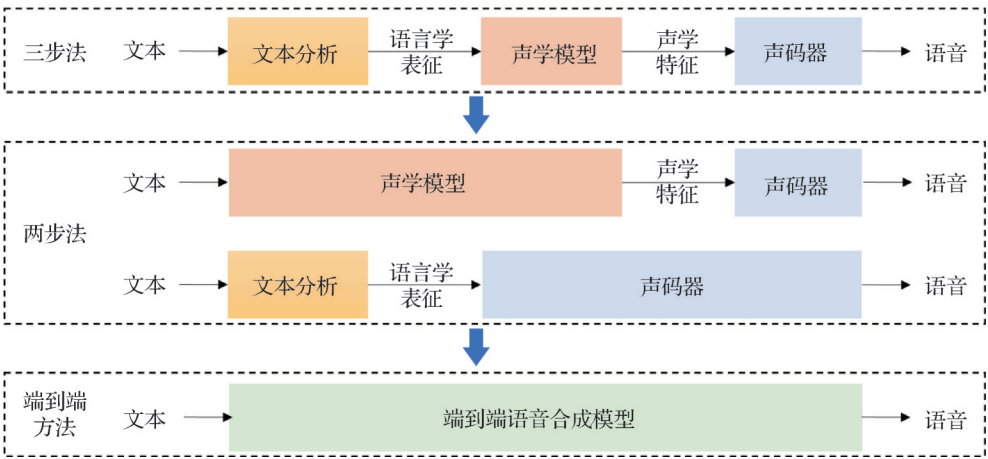


图 2 语音合成技术的发展流程
Fig. 2 The development process of SS

大多数传统的语音合成技术采用的是三步法, 主要包括波形拼接法 (concatenative speech synthesis) 和参数生成法 (statistical parametric speech synthesis, SPSS)。波形拼接法首先需要构建一个语音库, 然后对输入文本进行分析获得音素和韵律等信息, 再根据目标代价函数从语音库中进行匹配寻找候选的语音片段, 最后进行语音拼接 (Hunt 和 Black, 1996; Stylianou, 2001)。波形拼接法可以保留原始语音的高质量和高自然度特性, 但往往会受制于语音库的规模和制作的语音质量。值得注意的是, 参数生成法可以有效解决波形拼接法可控性差的问题, 其通过建立声学模型来获得语音的特征参数。其中, 基于隐马尔科夫模型 (hidden Markov model, HMM) 的 SPSS 方法是最经典的语音合成方法之一, 其利用统计模型和马尔可夫过程建立声学模型生成合成语音 (Tokuda 等, 2000; Zen 等, 2009; Tokuda 等, 2013)。

基于 HMM 的 SPSS 方法可以根据训练数据的多样性和模型的复杂性进行灵活调整以合成更自然的语音, 但其音质不如波形拼接法合成的语音。为此,

一些研究学者提出基于信号处理的声码器应用于 SPSS 方法中, 以提高合成语音的质量。其中, 基于信号处理的声码器主要有 Griffin-Lim (Griffin 和 Lim, 1984), STRAIGHT (speech transformation and representation using adaptive interpolation of weighted spectrum) (Kawahara, 2006) 和 WORLD (Morise 等, 2016)。同时, 一些研究学者借助深度神经网络 (deep neural network, DNN) 进一步改进声学模型, 提出了基于 DNN 的 SPSS 方法来学习上下文的信息 (Zen 等, 2013; Qian 等, 2014)。

随着深度学习参与到语音合成的全过程中, 语音合成技术也迎来了范式的转变, 即从三步法迈向两步法和端到端方法。在两步法中, 基于声学模型的语音合成方法利用深度网络学习输入文本与其对应声学特征之间的映射, 从而准确地预测声学特征。该方法包括 DeepVoice 3 (Ping 等, 2017)、TransformerTTS (Li 等, 2019) 和 FastSpeech 1 (Ren 等, 2019) 等。其中, 声学特征主要有线性语谱图、梅尔语谱图 (Mel-spectrogram)、基频 (F0) 等。值得注意的是, 目前大多数研究工作主要采用梅尔语谱图作

表 2 语音合成的研究归纳
Table 2 Overview of SS

分类依据	分类类型	语音合成技术
发展流程	三步法	波形拼接法(Hunt 和 Black, 1996; Stylianou, 2001), 基于 HMM 的 SPSS 方法(Tokuda 等, 2000, 2013; Zen 等, 2009), 基于信号处理声码器的 SPSS 方法(Griffin 和 Lim, 1984; Kawahara, 2006; Morise 等, 2016), 基于 DNN 的 SPSS 方法(Zen 等, 2013; Qian 等, 2014)等
	两步法	WaveNet (van den Oord 等, 2016), DeepVoice 1/3 (Arik 等, 2017a; Ping 等, 2017), Tacotron 1/2 (Wang 等, 2017; Shen 等, 2018), Parallel WaveNet (van den Oord 等, 2018), WaveRNN (Kalchbrenner 等, 2018), TransformerTTS (Li 等, 2019), FastSpeech 1 (Ren 等, 2019), HiFi-GAN (Kong 等, 2020), WaveGrad (Chen 等, 2020a), DiffWave (Kong 等, 2021), Parallel Tacotron (Elias 等, 2021), DiffVoice (Liu 等, 2023c)等
	端到端方法	WaveGlow (Prenger 等, 2019), Glow-TTS (Kim 等, 2020), Wave-Tacotron (Weiss 等, 2021), VITS (Kim 等, 2021), EATS (Kim 等, 2021), LightSpeech (Luo 等, 2021), FastDiff (Huang 等, 2022), FastSpeech 2 (Ren 等, 2022)等
生成模型	Seq2Seq	WaveNet (van den Oord 等, 2016), DeepVoice 1/2/3 (Arik 等, 2017a, b; Ping 等, 2017), Tacotron 1/2 (Wang 等, 2017; Shen 等, 2018), WaveRNN (Kalchbrenner 等, 2018), LightSpeech (Luo 等, 2021)等
	Flow	Parallel WaveNet (van den Oord 等, 2018), WaveGlow (Prenger 等, 2019), Glow-TTS (Kim 等, 2020), Wave-Tacotron (Weiss 等, 2021), Glow-WaveGAN 1/2 (Cong 等, 2021; Lei 等, 2022), PaddleSpeech (Zhang 等, 2022a)等
	GAN	WaveGAN (Donahue 等, 2018), GAN-TTS (Bińkowski 等, 2019), VocGAN (Yang 等, 2020), Parallel WaveGAN 1/2/3 (Yamamoto 等, 2020, 2021; Song 等, 2021), HiFi-GAN (Kong 等, 2020), EATS (Kim 等, 2021), BigVGAN (Lee 等, 2023)等
	VAE	GMVAE-Tacotron (Hsu 等, 2018), VAE-TTS (Zhang 等, 2019), WaveVAE (Peng 等, 2020), Parallel Tacotron (Elias 等, 2021), VITS (Kim 等, 2021)等
	Diffusion	WaveGrad (Chen 等, 2020a), DiffWave (Kong 等, 2021), Diff-TTS (Jeong 等, 2021), FastDiff (Huang 等, 2022), DiffVoice (Liu 等, 2023c)等

为声学特征。基于声码器的语音合成方法通过学习语言学表征和声音波形之间的映射关系,从而合成自然流畅的语音。其中,有代表性的声码器包括 WaveNet(van den Oord 等, 2016)、HiFi-GAN(Kong 等, 2020)、DiffWave (Kong 等, 2021)以及 BigVGAN(Lee 等, 2023)等。

基于端到端的语音合成方法采用深度网络以端到端的方式实现了从文本输入到合成语音的全过程。基于端到端的语音合成方法主要有 Glow-TTS (Kim 等, 2020)、VITS (variational inference with adversarial learning for end-to-end text-to-speech) (Kim 等, 2021)和 FastDiff (Huang 等, 2022)等。目前基于端到端的语音合成方法在自然度、拟人化等方面已经接近或达到真实语音的水平。总体而言,通过三步法实现语音合成过程中,可以对每个模块进行独立的优化和调整,但是每个模块的错误会逐步累积,导致合成的语音质量下降;通过两步法合成语音,一定程度上弥补了三步法因传统统计建模造成

语音合成质量不足的问题,但仍然存在模块之间误差累积的问题;基于端到端的语音合成方法有效避免了多阶段建模导致的误差累积,语音合成的质量得到了大幅提升,但是需要大量的训练语音数据和计算资源来训练复杂的深度学习模型。

根据生成模型的不同,基于深度学习的语音合成方法可以分为序列到序列(sequence-to-sequence, Seq2Seq)模型、流(Flow)模型、生成对抗网络(generative adversarial network, GAN)模型、变分自编码器(variational auto-encoder, VAE)模型和扩散(Diffusion)模型。表 2 汇总了上述 5 种生成模型的主要语音合成方法,本节将从上述 5 种生成模型介绍语音合成方法的研究现状。

1. 1. 1 基于 Seq2Seq 模型的语音合成

Seq2Seq 模型通常由编码器和解码器组成。在语音合成任务中,编码器的作用是将输入的文本序列进行编码,从而获得输入文本的上下文信息。而解码器接收编码器输出的上下文信息并解码为相应

的语音波形。基于 Seq2Seq 模型的语音合成方法主要有 DeepVoice(Arik 等, 2017a)以及 Tacotron(Wang 等, 2017)等系列工作。

WaveNet 模型和 Tacotron 模型(Wang 等, 2017)都是由 Google 研究人员提出的基于 Seq2Seq 模型的语音合成方法。其中, WaveNet 模型采用二步法设计, 其通过扩大因果卷积来增加感受野以合成更真实的语音。不同的是, Tacotron 模型采用端到端方法进行设计, 其关注文本和音频之间的对齐和转换, 以合成高质量的语音。在此基础上, Tacotron 2 模型(Shen 等, 2018)引入 WaveNet 网络作为声码器, 能够生成自然度高且质量高的语音。但该模型训练和调优的过程比较复杂, 同时需要大量的数据和计算资源。为了改进合成的性能和效率, 衍生出了一系列基于 Seq2Seq 模型的语音合成方法, 如 WaveRNN(Kalchbrenner 等, 2018)和 LightSpeech(Luo 等, 2021)等。

基于 Seq2Seq 模型的语音合成方法相对简单、易于实现, 同时可以通过端到端的方式进行设计语音合成方法。但是在编码过程中可能会丢失训练语音的部分声调、韵律等细节信息, 导致合成的语音不够清晰和准确。

1. 1. 2 基于 Flow 模型的语音合成

基于 Flow 模型的语音合成方法的目标是学习一个可逆的变换函数, 建立输入文本与说话人声学特征之间的映射关系, 从而实现语音合成。

Glow 模型(Kingma 和 Dhariwal, 2018)是由 OpenAI 公司提出的一种基于 Flow 的生成模型, 其将可逆的 1×1 卷积网络引入到 Flow 网络中, 从而简化 Flow 模型的计算量。基于此, Nvidia 公司(Prenger 等, 2019)将 Glow 网络和 WaveNet 网络相结合, 提出端到端语音合成模型 WaveGlow。该模型采用 Glow 模型的可逆 1×1 卷积网络来混合不同通道间的信息, 以实现快速合成语音。相似地, Cong 等人(2021)提出了一种基于 Glow 和 WaveGAN(Donahue 等, 2018)相结合的端到端语音合成模型 Glow-WaveGAN, 采用 Glow 网络将输入文本映射为语音的潜在表征, 而 WaveGAN 网络则通过波形重构来学习潜在的语音表征。

与基于 Seq2Seq 模型的语音合成方法相比, 基于 Flow 模型的语音合成方法可以学习到数据的复杂分布, 能够合成更加自然流畅的语音, 同时可以调整生成语音的音调、语速等参数。然而, 基于 Flow

模型的语音合成方法在训练过程中仍需要大量的标注数据和计算资源。

1. 1. 3 基于 GAN 模型的语音合成

基于 GAN 模型的语音合成方法通过对生成器网络和鉴别器网络之间的对抗训练来实现语音合成(Bińkowski 等, 2019; Yang 等, 2020; Kong 等, 2020; Kim 等, 2021)。具体而言, 生成器网络以文本或语音特征为输入以合成逼真的语音波形, 鉴别器网络负责判断生成器合成的语音是否真实。

Donahue 等人(2018)首次将 GAN 模型应用于语音合成任务, 提出了 WaveGAN 模型。该模型以语音特征和随机噪声向量作为输入, 通过生成器网络的一系列深度卷积和上采样来生成语音波形。同时, 利用鉴别器区别真实语音和伪造语音来提升生成器网络生成的语音质量。在此基础上, Yamamoto 等人(2020)提出了一种合成速度快、占用内存小的声码器模型 Parallel WaveGAN。该模型采用多重分辨率频谱短时傅里叶变换损失和时域对抗损失函数来训练改进的非自回归生成器网络 WaveNet, 能够有效捕获语音的时频分布。其中, 改进的 WaveNet 网络将原始的因果卷积替换为非因果卷积, 以实现并行合成语音。

尽管基于 GAN 模型的语音合成方法能够合成高质量的语音, 但在训练过程中如何平衡好生成器和鉴别器仍然是一个难点问题。

1. 1. 4 基于 VAE 模型的语音合成

VAE 模型由编码器和解码器组成。具体而言, 基于 VAE 模型的语音合成方法中编码器网络将语音的特征映射为隐变量特征, 而解码器网络则通过隐变量特征生成语音(Hsu 等, 2018)。基于 Seq2Seq 模型的语音合成方法的编码器和解码器通常采用循环神经网络(recurrent neural network, RNN)结构, 其注重序列的建模。与之不同的是, 基于 VAE 模型的语音合成方法中的编码器和解码器通常采用 VAE 网络对语音的潜在空间进行表征。

Zhang 等人(2019)提出 VAE-TTS 模型, 利用高斯先验的 VAE 网络来模拟潜在空间中的方差信息作为正则化项, 从而对合成语音的风格进行建模和控制。此外, 一些研究工作(Peng 等, 2020; Elias 等, 2021; Kim 等, 2021)利用 VAE 网络, 更好地建模方差信息以进行合成语音的风格化。基于 VAE 模型的语音合成方法在合成阶段进行韵律控制或韵律迁移

等操作来生成多样化的语音。但是该方法在编码和解码过程中同样会面临着信息损失的问题,同时需要大量的数据和计算资源。

1. 1. 5 基于 Diffusion 模型的语音合成

基于 Diffusion 模型的语音合成方法是一种新兴且备受关注的技术,通过扩散概率模型(diffusion probabilistic model, DPM)进行多次迭代,将高斯噪声信号转化为目标语音信号。

目前,一些研究工作采用扩散概率模型作为声码器以实现语音合成,如 WaveGrad (Chen 等, 2020a) 和 DiffWave (Kong 等, 2021) 等。这些模型的基本思想是使用扩散过程和反向过程来表达语音数据和潜在分布之间的映射。在扩散过程中,对音素特征或语音特征逐步添加随机噪声,从而获得高斯噪声。在反向过程中,通过对随机高斯噪声逐步迭代使其逐渐逼近目标语音。基于 DPM 的声码器可以产生非常高品质的语音,但需要进行多次迭代计算导致训练和推理过程相对较慢。

基于 Diffusion 模型的语音合成方法在合成语音的质量和可控性方面具有较大的潜力,为语音合成技术的发展开辟了新的前景。

1. 2 语音转换

语音转换技术主要研究的是语音中不同说话人之间的身份转换,其基本思想是在保留源说话人语音的语义信息同时,将源说话人语音的身份特征转换为目标说话人语音的身份特征。其中,语音的说话人身份特征主要包括共振峰、频谱包络、F0 以及音调等(杨帅 等, 2022)。

语音转换技术的基本框架如图 3 所示,主要分为训练阶段和转换阶段。在训练阶段,对源说话人和目标说话人语音进行语音分析获得相应的特征表示。然后,将源说话人和目标说话人语音输入到映射模型中进行训练,从而构建源说话人和目标说话人语音之间的映射关系。转换阶段的主要目标是将源说话人语音转换为目标说话人语音。首先,提取源说话人的语音特征,需要与映射模型进行训练的语音特征类型保持一致。然后,利用映射模型学习到的语音映射关系来生成目标说话人的语音特征,最后对目标说话人的语音特征进行重构生成目标说话人的语音波形。

语音转换的研究归纳如表 3 所示,本文将从训练数据和目标说话人数量两方面对近几年的语音转

换技术进行分类。根据训练过程中源说话人和目标说话人的语音特征是否进行对齐,语音转换技术可以分为基于平行语料的语音转换和基于非平行语料的语音转换。基于平行语料的语音转换技术指的是源说话人和目标说话人语音的语义内容一致,在特征映射前对源说话人和目标说话人的语音特征进行对齐,使得源说话人与目标说话人的语音在相同时间帧的音素之间具有关联性。基于非平行语料的语音转换技术不需要源说话人和目标说话人语音的语义内容保持一致,且不需要进行特征之间的对齐操作。该技术通过训练映射模型直接学习源说话人和目标说话人语音之间的映射关系。

根据目标说话人数量的不同,语音转换技术可以分为单一、多个和任意目标说话人的语音转换。单目标说话人的语音转换专注于特定说话人的语音转换,主要分为一对一、多对一、任意对一的语音转换技术。多目标说话人的语音转换主要为一对多、多对多以及任意对多的语音转换技术,为实现个性化和多样化的语音转换提供了重要的手段。任意目标说话人的语音转换专注于对未知目标说话人的语音转换,主要为任意到任意的语音转换技术。本小节从目标说话人数量的角度介绍现有的语音转换技术的研究进展。

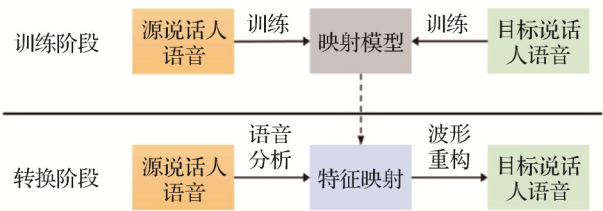


图 3 语音转换技术的基本框架

Fig. 3 The basic framework of VC

1. 2. 1 单目标说话人的语音转换

单目标说话人的语音转换的研究主要以一对一的语音转换为主。一对一的语音转换旨在固定的源说话人和目标说话人语音之间的转换。Kaneko 和 Kameoka (2018) 提出基于循环一致性 GAN 的语音转换模型 CycleGAN-VC。该模型采用循环一致性损失函数学习平行数据中源说话人和目标说话人语音频谱特征之间的正逆映射关系,并通过 GAN 模型学习目标说话人语音的特征分布。针对基于 CycleGAN 网络的语音转换模型训练难度大的问题, Li 等人

表 3 语音转换的研究归纳
Table 3 Overview of VC

分类依据	分类类型	语音转换技术
训练的语音数据	平行语料	基于参数统计的高斯混合模型 (Chen 等, 2003), 最小二乘回归 (Helander 等, 2010), 基于非参数统计的非负矩阵分解 (Aihara 等, 2013), ConvS2S-VC (Kameoka 等, 2020), INVVC (Cai 和 Li, 2022) 等
	非平行语料	CycleGAN-VC 1/2/3 (Kaneko 和 Kameoka, 2018; Kaneko 等, 2019a, 2020), StarGAN-VC (Kameoka 等, 2018), StarGAN-VC2 (Kaneko 等, 2019b), ACVAE-VC (Kameoka 等, 2018), AdaIN-VC (Chou 和 Lee, 2019), AutoVC (Qian 等, 2019), Blow (Serrà 等, 2019), ARVC (Lian 等, 2020), CVC (Li 等, 2021a), FragmentVC (Lin 等, 2021a), TGAVC (Tang 等, 2021), S2VC (Lin 等, 2021b), NVC-Net (Nguyen 和 Cardinaux, 2022), Glow-WaveGAN 2 (Lei 等, 2022), ALGAN-VC (Dhar 等, 2023), SACS-VC (Xu 等, 2023) 等
目标说话人数量	单一	WaveNet (van den Oord 等, 2016), WaveRNN (Kalchbrenner 等, 2018), CycleGAN-VC (Kaneko 和 Kameoka, 2018), ARVC (Lian 等, 2020), Cascade ASR-TTS (Huang 等, 2020), CVC (Li 等, 2021a), INVVC (Cai 和 Li, 2022), ALGAN-VC (Dhar 等, 2023) 等
	多个	EV-GMM (Saito 等, 2011), StarGAN-VC (Kameoka 等, 2018), StarGAN-VC2 (Kaneko 等, 2019), ACVAE-VC (Kameoka 等, 2019), Blow (Serrà 等, 2019), CycleGAN-VC 2/3 (Kaneko 等, 2019a, 2020), Cotatron (Park 等, 2020), ConvS2S-VC (Kameoka 等, 2020), TGAVC (Tang 等, 2021), NVC-Net (Nguyen 和 Cardinaux, 2022), QuickVC (Guo 等, 2023) 等
	任意	AdaIN-VC (Chou 和 Lee, 2019), AutoVC (Qian 等, 2019), FragmentVC (Lin 等, 2021a), S2VC (Lin 等, 2021b), DRVC (Wang 等, 2022c), Glow-WaveGAN 2 (Lei 等, 2022), SACS-VC (Xu 等, 2023) 等

(2021a)提出一种基于对比学习的语音转换模型 CVC,利用对比学习来提高单向 GAN 模型的训练效率,同时建立源说话人和目标说话人之间的映射关系。此外,Dhar 等人(2023)提出了基于自适应学习 GAN 的语音转换框架 ALGAN-VC,来改善语音质量和提高目标说话人的相似度。该模型将密集残差网络架构合并到生成器网络中,并采用自适应学习机制来优化模型的损失函数,能够更有效学习源说话人和目标说话人之间的语音特征。

除了一对一的语音转换,单目标说话人的语音转换还包括多对一和任意对一的语音转换。多对一的语音转换侧重于构建多个源说话人与特定目标说话人之间的映射关系,而任意对一的语音转换侧重于从源说话人语音中学习与所述人无关的通用特征表示。Lian 等人(2020)提出了任意对一的语音转换模型 ARVC(auto-regressive voice conversion),通过自回归模型处理语音的声学特征,采用深度声码器控制频谱形状的输出实现语音转换。在平行语料的语音转换中,Cai 和 Li(2022)提出可逆深度学习的语音转换模型(invertible voice conversion, INVVC),采用可逆 1×1 卷积和仿射耦合层来实现源说话人的身份溯源。该模型利用并行语料的语音数据进行训练,在一对一和多对一的语音转换任务下取得了优异的性能。

目前,单目标说话人的语音转换技术在实际应用场景中局限性较大,且对训练数据的依赖性较强,需要使用大量的语音数据进行训练。

1. 2. 2 多目标说话人的语音转换

随着个性化和多样化语音需求的日益增长,语音转换技术正朝着单目标说话人到多目标说话人的方向迅速发展。多目标说话人的语音转换通过映射模型在训练过程中学习源说话人与多个目标说话人之间的语音特征映射关系,主要包括一对多、多对多以及任意对多的语音转换技术。

在一对多的语音转换中,Saito 等人(2011)提出了一种基于高斯混合模型(Gaussian mixture model, GMM)的特征语音转换模型 EV-GMM,该模型将每个说话人表示构建为矩阵形式并建立基于 GMM 超向量的说话人空间,在转换阶段通过说话人空间可以更灵活地转换目标说话人的语音特征。此外,为了解决转换过程中源说话人和目标说话人语音共振峰的位置差异,Mohammadi(2015)采用动态频率扭曲来均衡源说话人和目标说话人语音帧的共振峰位置,以降低模型的复杂性和提高语音的质量。

在多对多的语音转换中,Kameoka 等人(2018)提出一种基于星型 GAN 的语音转换模型 StarGAN-VC,该模型利用说话人身份标签的独热编码向量以实现多说话人之间的多对多语音转换。Serrà 等人

(2019)将Flow模型引入到语音转换任务中,提出了Blow模型,该模型采用超网络的单尺度归一化Flow模块进行原始音频之间的多对多语音转换。此外,Nguyen和Cardinaux(2022)提出基于对抗训练的端到端语音转换模型NVC-Net。该模型采用GAN网络以语音波形为输入对语音的内容和说话人信息进行解耦,实现了以较少的参数和更快的推理速度进行语音转换。

在任意对多的语音转换中,Park等人(2020)提出了基于转录引导语音编码器Cotatron,用于提取与说话人无关的语音表示。从查找表中检索到的说话人特征来调节语音转换解码器,以实现目标说话人的语音转换。为了提高语音转换的速度,Guo等人(2023)提出一种基于VITS的轻量级语音转换模型QuickVC,在语音的质量和转换速度方面都具有较强的竞争力。该模型采用HuBERT-Soft预训练模型提取语音的语义特征,同时采用短时傅里叶逆变换减少VITS网络中解码器的计算冗余。

与单目标说话人的语音转换相比,多目标说话人的语音转换在输出的目标说话人数量上进行了扩展。然而,该技术在模型设计和训练过程方面变得更加复杂,在一定程度上增加了训练的难度和计算复杂度。同时,多目标说话人语音转换模型需要大量的目标说话人语音数据进行训练,这极大限制了语音转换技术的应用落地。

1.2.3 任意目标说话人的语音转换

与多目标说话人的语音转换不同的是,任意目标说话人的语音转换更专注于对未知目标说话人的语音进行转换,主要以任意到任意的语音转换为主。任意到任意的语音转换建立在语音解耦表征的基础上。它通过将语音解耦为说话人音色和语义内容表示,然后利用说话人的音色表示将源说话人的语音替换为目标说话人的语音,实现对任意目标说话人的语音转换。

Chou和Lee(2019)提出一次性的语音转换模型AdaIN-VC,采用实例归一化将语音解耦为说话人音色和语义内容表示,在推理过程中仅通过源说话人和目标说话人的实例语音进行语音转换。Qian等人(2019)提出基于信息瓶颈的VAE模型AutoVC,采用自重构损失训练以实现非并行的语音转换。该模型使用预训练的说话人编码器并对瓶颈层的长度施加限制,以避免源说话人语音音色或语义内容信息的

泄漏。然而,FragmentVC(Lin等,2021a)直接关注源说话人和目标说话人的语音,采用Wav2Vec2预训练模型提取源说话人语音的特征,从目标说话人语音的对数梅尔语谱图提取频谱特征,然后提取并融合适当的细粒度语音片段,其表现与AdaIN-VC和AutoVC的性能不相上下。

Wang等人(2022c)提出基于解耦表征的语音转换模型(disentangled representation voice conversion, DRVC),以解决语音内容和风格在解缠时的重叠问题。该模型是一个端到端的自监督模型,由内容编码器、音色编码器和生成器所组成。同时,该模型通过循环重构损失和信息相同损失来优化内容和风格信息特征,转换后的语音与AdaIN-VC和AutoVC相比在质量和语音相似度上都有所提高。为了更好地平衡目标说话人的音色和保留语音的语义内容,Xu等人(2023)提出任意到任意语音转换模型SACS-VC,其引入说话人自适应和内容监督模块来解决上述问题。说话人自适应模块采用时间注意力机制根据语音的语义内容分布自适应地重新排列说话人音色表示。同时,采用内容监督模块构建源说话人和目标说话人语音的关联以实现保持语义内容的一致性。实验表明,该模型转换后的语音在质量和相似度上都优于AdaIN-VC和AutoVC模型。

面对应用场景中难以获取大量训练语音数据的问题,一些研究学者针对任意目标说话人的语音转换任务提出了少样本的解决方案。虽然基于少样本或零样本的语音转换在语音质量和相似度上仍然有所不足,但是它们仍然是未来的发展趋势之一。

2 相关数据集及评价指标

2.1 语音伪造数据集

数据集是评估语音深度伪造检测技术的研究基础。在语音伪造技术不断改进的过程中,一些高质量、多样性的数据集也陆续发布。2015—2023年语音伪造检测数据集的信息统计如表4所示。本节对ASVspoof系列挑战赛、ADD系列挑战赛以及Partial-Spoof数据集进行详细介绍。

2.1.1 ASVspoof系列挑战赛

ASVspoof挑战赛由芬兰东芬兰大学、英国爱丁堡大学以及日本国家信息学研究所等知名大学和国际研究机构共同发起,每两年举办一次,是语音伪造

表 4 2015—2023 年语音伪造检测数据集的信息统计表
Table 4 Statistics of the speech forgery detection datasets from 2015 to 2023

伪造类型	数据集	文献	语音数量 (子数据集:真/假)	语言	链接
TTS	Fake-or-Real	Reimao 和 Tzerpos (2019)	111 000/87 000	英	https://bil.eecs.yorku.ca/datasets/
	HAD	Yi 等人(2021)	53 612/107 224	汉	—
	WaveFake	Frank 和 Schönherr (2021)	16 283/117 985	英/日	https://zenodo.org/record/5642694
	ADD 2022	Yi 等人(2022a)	LF: 5 619/46 067	汉	—
			PF: 5 319/46 419		
			FG-D: 5 319/46 206		
	CMFD	Su 等人(2023)	汉: 1 800/1 000 英: 1 800/1 000	英/汉	https://github.com/WuQinfang/CMFD
	FAD	Ma 等人(2022)	115 800/115 800	汉	https://zenodo.org/record/6635521#.Ysjq4nZBw2x
	In-the-Wild	Müller 等人(2022)	19 963/11 816	英	https://deepfake-demo.aisee.fraunhofer.de/in_the_wild
	Yan Dataset	Yan 等人(2023)	22 400/136 209	汉	—
Replay	ASVspoof 2017	Kinnunen 等人 (2017)	3 565/14 465	英	http://dx.doi.org/10.7488/ds/2332
	ReMASC	Gong 等人(2019)	9 240/45 472	英/汉/ 印度	https://github.com/YuanGongND/ReMASC
	Mandalapu Dataset	Mandalapu 等人 (2021)	—	英/挪威/ 印度/法	https://forms.gle/EV5ur4jbw52RgM3v8
TTS 和 VC	AVspoof	Ergünay 等人(2015)	LA: 15 504/120 480 PA: 15 504/14 465	英	https://zenodo.org/record/4081040
	ASVspoof 2015	Wu 等人(2015)	16 651/246 500	英	http://dx.doi.org/10.7488/ds/298
	FMFCC-A	Zhang 等人(2021d)	10 000/40 000	汉	https://github.com/Amforever/FMFCC-A
	EmoFake	Zhao 等人(2022)	35 000/53 200	英/汉	—
	SceneFake	Yi 等人(2022b)	19 838/64 642	英	https://github.com/ADDchallenge/SceneFake
	ADD 2023	Yi 等人(2023)	FG-D: 172 819/ 113 042	汉	—
			RL: 55 468/65 449		
			AR: 14 907/95 383		
	DECRO	Ba 等人(2023)	汉: 21 218/41 880 英: 12 484/42 799	英/汉	https://github.com/petrichorwq/DECRO-dataset
	PartialSpoof	Zhang 等人(2023b)	12 483/108 978	英	https://zenodo.org/record/4817532#.Yld8Yi2l1hF
TTS, VC 和 Replay	ASVspoof 2019	Todisco 等人(2019)	LA: 12 483/108 978 PA: 28 890/189 540	英	https://datashare.ed.ac.uk/handle/10283/3336
	ASVspoof 2021	Delgado 等人(2021)	LA: 18 452/163 114	英	LA: https://zenodo.org/record/4837263
			PA: 126 630/ 816 480		PA: https://zenodo.org/record/4834716
			PF: 14 869/519 059		DF: https://zenodo.org/record/4835108

注:“—”表示原论文中未提及。

检测领域内最热门、最全面且规模最大的挑战赛。ASVspoof挑战赛旨在推动语音深度伪造检测技术的发展。目前,ASVspoof挑战赛已经成功举办了4届(2015年、2017年、2019年和2021年),第5届挑战赛正在处于数据收集阶段。

ASVspoof 2015数据集是第1个用于语音伪造检测研究的数据集,其主要针对逻辑访问(logical access, LA)场景。LA场景指将伪造语音输入到ASV系统中直接进行说话人验证。该数据集包含46位说话人(20男,26女),由7种TTS和3种VC算法所生成。

ASVspoof 2017数据集主要针对在物理访问(physical access, PA)场景下语音的重放(replay)攻击。在PA场景中,伪造语音直接输入到录音或播放设备中,通过将录制或播放的伪造语音输入到ASV系统中以实现欺骗的目的。ASVspoof 2017数据集由177个重放会话和61种重放配置生成。

在此基础上,ASVspoof 2019数据集兼顾LA和PA场景,包括20位说话人(8男,12女)。该数据集的LA场景包含多种TTS攻击、VC攻击以及TTS和VC的混合攻击,更适用于评估语音深度伪造检测技术。

ASVspoof 2021数据集对ASVspoof 2019数据集中LA和PA场景的语音数据进行了扩展,在LA场景中对伪造语音增加信道的变化,在PA场景中采用真实物理环境下录制音频,进一步检验伪造检测技术的鲁棒性。此外,ASVspoof 2021挑战赛引入了语音深度伪造(deepfake, DF)场景,该场景对LA场景的语音数据进行不同格式的压缩,是LA场景的一种拓展。

2.1.2 ADD系列挑战赛

语音深度伪造检测(audio deepfake detection, ADD)挑战赛由中国科学院自动化研究所、清华大学的陶建华教授联合新加坡国立大学、香港中文大学(深圳)的李海洲教授共同发起,每年举办一次,旨在推动和促进语音深度伪造检测技术的研究进展,在2022年和2023年已经成功举办两届。

ADD 2022挑战赛由低质量虚假音频检测(low-quality fake audio detection, LF)、部分虚假音频检测(partially fake audio detection, PF)以及音频攻防博弈(audio fake game, FG)所组成。值得注意的是,FG赛道包括生成任务(generation task, FG-G)和检测任

务(detection task, FG-D)。

ADD 2023挑战赛设置的赛道主要有音频攻防博弈(FG)、篡改区域定位(manipulating regional location, RL)和生成算法溯源(deepfake algorithm detection, AR)。其中,FG赛道包括FG-G和FG-D赛道。与ADD 2022挑战赛相比,ADD 2023挑战赛不仅关注语音的伪造与检测,还关注伪造语音的定位与溯源问题。

2.1.3 PartialSpoof数据集

PartialSpoof数据集(Zhang等, 2023b)是由日本国家信息学研究所、日本SOKENDAI公司和法国EURECOM研究所共同合作创建的,该数据集旨在研究部分语音伪造攻击的问题。

PartialSpoof数据集是基于ASVspoof 2019数据集的LA场景进行构建的,包含12 483段真实语音和108 978段部分伪造语音。部分伪造语音数据是由合成或转换的语音片段被嵌入到真实的语音中所生成的。在生成过程中,该数据集不考虑句子和单词的含义以及语音片段前后的音素。伪造语音数据的时长为0.48 s~21.02 s,伪造语音占整个语音样本的时长比例为0.23%~99.81%。在话语级标签的基础上,采用6种不同时间分辨率构建了分段级标签,分别为20 ms、40 ms、80 ms、160 ms、320 ms和640 ms。在话语级标签中,伪造语音数据的标签占总样本标签约89%,在分段级标签中,这一比例下降至约40%,这有效地解决了标签不均衡的问题。

图4为PartialSpoof数据集的标签处理示意图,其中,1代表真实语音,0代表伪造语音,红色虚线框为伪造语音片段。

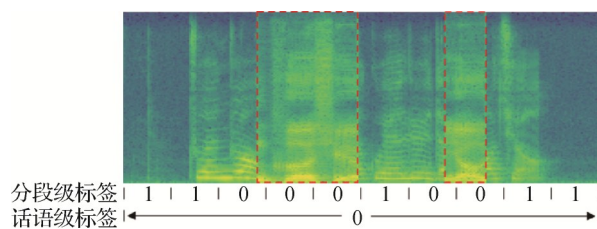


图4 PartialSpoof数据集的标签处理

Fig. 4 Label processing for the PartialSpoof dataset

2.2 评价指标

选择合适的评价指标对评估语音深度伪造检测技术的性能至关重要。通过度量检测技术的有效性、鲁棒性和泛化性,从而评估该模型在实际场景中

的表现。当前常用的评价指标主要包括等错误率(equal error rate, EER)和串联检测代价函数(tandem detection cost function, t-DCF)等。

2.2.1 等错误率

EER是在误识率(false acceptance rate, FAR)和拒识率(false rejection rate, FRR)相等时计算的绩效指标,反映了在相等EER值下检测模型的鲁棒性和平衡性。对于语音深度伪造检测任务而言,FAR表示伪造语音被分类为真实语音的概率,FRR表示真实语音被分类为错误语音的概率。当已知伪造检测模型的检测分数 s 和阈值 θ 时,令 $P_{fa}(\theta)$ 和 $P_{fr}(\theta)$ 分别对应于阈值 θ 下的FAR和FRR,则

$$P_{fa}(\theta) = \frac{N_f(s > \theta)}{N_f} \quad (1)$$

$$P_{fr}(\theta) = \frac{N_r(s \leq \theta)}{N_r} \quad (2)$$

式中, N_f 和 N_r 分别为伪造语音和真实语音的样本总数。 $N_f(s > \theta)$ 为检测分数 s 大于阈值 θ 的伪造语音样本数。同理, $N_r(s \leq \theta)$ 为检测分数 s 小于等于阈值 θ 的真实语音样本数。 $P_{fa}(\theta)$ 是关于阈值 θ 的单调递减函数,而 $P_{fr}(\theta)$ 是关于阈值 θ 的单调递增函数。那么当阈值为 θ 时,该模型的EER为

$$f_{EER} = P_{fa}(\theta) = P_{fr}(\theta) \quad (3)$$

EER值越小代表伪造检测模型的性能越好,但EER在衡量伪造检测模型的性能时并未考虑ASV系统的可靠性。

2.2.2 串联检测代价函数

t-DCF是在ASVspoof 2019挑战赛中引入的新评价指标,反映了伪造语音和语音伪造检测模型对ASV系统性能共同产生的影响(Kinnunen等,2019)。t-DCF借鉴最小风险贝叶斯决策对伪造检测模型的可靠性进行评估。然而,在实际场景中,ASV系统可能遇到真实说话人和冒充的非法说话人等情况,基于此,t-DCF综合考虑了不同情况下的误判代价。t-DCF的计算过程为

$$f_{t-DCF} = C_{fa} \times (1 - P_{ta}) \times P_{fa}(\theta) + C_{fr} \times P_{ta} \times P_{fr}(\theta) \quad (4)$$

式中, C_{fa} 和 C_{fr} 分别表示FAR和FRR的权重值, P_{ta} 和 $(1 - P_{ta})$ 分别表示真实说话人和冒充的非法说话人出现的先验概率。在ASVspoof 2019和ASVspoof 2021挑战赛中, C_{fa} 、 C_{fr} 和 P_{ta} 分别设置为10、1和0.940 5。一般情况下,当阈值为 θ 时,取最小的

t-DCF值作为最终结果。t-DCF值越小说明检测模型的泛化性能越好。

3 语音深度伪造检测技术

随着语音深度伪造技术的不断发展,伪造语音的内容更加丰富且应用场景更加多元化。然而,伪造语音带来的潜在安全风险逐渐增大,及时且有效地鉴别伪造语音是构筑信息安全的关键之一。在此背景下,语音深度伪造检测技术应运而生。

图5给出了语音深度伪造检测的基本框架,其主要由特征提取和分类判决所组成。特征提取是从语音信号中提取具有区分性的特征,以获得语音的频谱信息、能量分布和共振峰分布等信息。特征提取是语音深度伪造检测技术的关键,直接影响着检测技术的性能。分类判决通过分类器来判别输入的语音是否为伪造语音。

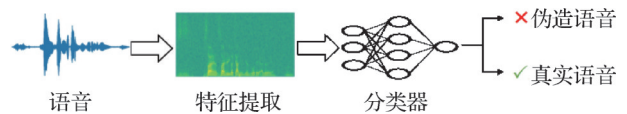


图5 语音深度伪造检测的基本框架

Fig. 5 The basic framework of speech deepfake detection

语音深度伪造检测技术与深度学习的结合,极大推动了语音伪造检测领域的发展。如图6所示,本文从数据增强、特征提取和优化以及学习机制等3个方面对语音深度伪造检测技术进行分类介绍。

本文在介绍语音深度伪造检测技术的过程中更注重于探讨现有伪造检测技术的通用性,而不是针对特定的伪造攻击方式。

3.1 数据增强

数据增强(data augmentation, DA)是深度学习领域中常用的技术之一,其将有限的语音数据进行一系列变换或扩充,从而增加训练语音数据的多样性和数量。数据增强可以缓解模型的过拟合问题,并提高模型的鲁棒性。数据增强方法大致可以分为语音加噪、掩码增强、信道增强以及压缩增强。其中,语音加噪模拟的是真实环境中的背景噪声;掩码增强模拟的是语音信号出现遮挡或部分丢失的情况;信道增强模拟的是不同的录制设备或音频传输通道的变化;压缩增强模拟的是语音在编解码过程中出现的失真或压缩。

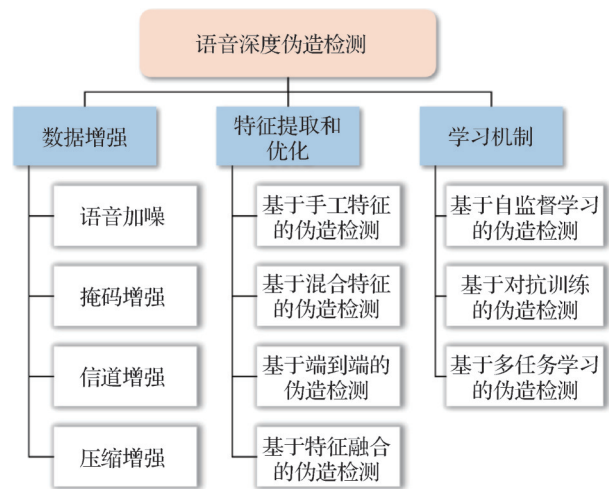


图6 语音深度伪造检测技术的分类

Fig. 6 Taxonomy of speech deepfake detection techniques

3.1.1 语音加噪

在语音深度伪造检测领域中,语音加噪是最常见的数据增强方法之一。具体而言,将噪声与原始干净的语音进行混合获得具有不同噪声特性和强度的语音,从而模拟现实世界中的噪声环境以提高伪造检测模型对噪声的鲁棒性。常见的语音加噪方法主要有白噪声、环境噪声以及非线性失真加噪等。

Yang 等人(2020)采用 0 dB、10 dB 和 20 dB 的咖啡馆噪声、白噪声和街道噪声对 ASVspoof 2015 数据集进行加噪,提取相应的特征比原始语音特征具有更好的表示能力。与之相似的是,Tak 等人(2022a)提出一种针对自动说话人验证反欺骗技术的原始语音数据增强模型 RawBoost。该模型将基于线性和非线性卷积噪声、脉冲信号相关加性噪声和稳态信号无关加性噪声进行组合,直接对原始语音的波形进行处理。ASVspoof 2021 数据集下语音深度伪造检测技术的相关研究汇总如表 5 所示。RawBoost 模型在 ASVspoof 2021 数据集的 LA 场景下获得最小 t-DCF 为 0.31,EER 为 5.31%。

语音加噪的方法可以提升伪造检测模型在噪声环境下伪造语音检测的鲁棒性。然而,它也存在一定的局限性。比较典型的是添加的噪声可能无法匹配测试语音数据的噪声统计特性,导致检测性能不佳。为了进一步提升语音深度伪造检测能力,需要综合考虑其他数据增强方法或更加复杂的信号处理技术,以增强伪造检测模型对噪声和伪造特征的鲁棒性。

3.1.2 掩码增强

掩码增强方法在训练过程中对输入语音的时域或频域进行处理,采用固定值替换需要掩码的语音特征块,有助于伪造检测模型学习部分信息缺失或遮挡的鲁棒性。

Park 等人(2019)提出了一种基于频域的数据增强方法,称为 SpecAugment。该方法将特征扭曲、频率掩码和时间序列掩码直接处理语音的深度特征,有助于将模型的过拟合问题转化为欠拟合问题。这使得研究学者能够探索更大更深的网络模型,以提升检测的性能。与之相似的是,Chen 等人(2020b)采用一种基于频域的数据增强方法 FreqAugment,来提高伪造检测模型的泛化能力。实验表明,该模型在 ASVspoof 2019 数据集的 LA 场景下获得 1.26% 的 EER。

Rostami 等人(2021)采用 SpecAugment 技术将每段训练语音在时域和频域上随机选取一定的子带进行零掩码,从而防止伪造检测模型在特定的时域或频域上出现过拟合现象。同时,作者还进一步探究了注意力掩码在伪造检测的影响。通过实验分析得出,基于注意力掩码的网络更加关注于语音信号的高频部分,且高频信息有利于提高检测模型的性能。为了提高伪造检测模型的泛化性,Cohen 等人(2022)提出了一种变体的 SpecAugment 数据增强方法 SpecAverage,通过计算每个特征的平均值进行在线掩码。作者认为 SpecAverage 能够获得更好性能的原因是平均值相对于输入特征具有统计意义。与之相似的是,Tak 等人(2021a)在训练过程中采用通道掩码对随机选取连续的 Sinc 滤波器通道进行置零,同一批次的训练数据采用相同通道进行掩码。

在对语音的时域或频域进行掩码过程中,掩码区域的选择、数值的替换等操作会增加一定的计算复杂度和时间开销。

3.1.3 信道增强

在 ASVspoof 2021 数据集的 LA 场景中,语音经过公共交换电话网络(public switched telephone network,PSTN)和互联网协议语音网络(voice over internet protocol,VoIP)的各种编码和传输处理。对于上述场景,信道增强方法有助于伪造检测模型提高对信道干扰或设备差异的鲁棒性。

ASVspoof 2021 数据集下语音深度伪造检测技术的相关研究汇总如表 5 所示。为了有效应对

表5 ASVspoof 2021数据集下语音深度伪造检测技术汇总
Table 5 Summary of speech deepfake detection techniques on ASVspoof 2021 dataset

文献	语音深度伪造检测技术				实验结果	
	语音预处理	特征提取	网络结构	损失函数	EER/% ↓	t-DCF ↓
Cáceres 等人(2021)	信道编码和压缩编码	MFCC, logFBE, CQCC, SCMC	TDNN, RawNet2 和 LCNN	CE, Focal Loss	LA: 5.58 (8) PA: 28.41 (3)	LA: 0.275 (4) PA: 0.738 (3)
Chen 等人(2021a)	加噪和编码传输	LFBs	ResNet-L-FM, ResNet-L-LDE, ResNet-S-LDE	SWA	LA: 3.21 (3) DF: 16.05 (7)	LA: 0.261 (2)
Chen 等人(2021b)	信道编码和压缩编码	LFCC	ECAPA-TDNN, ResNet	OC-Softmax, P2SGrad	LA: 5.46 (7) DF: 20.33 (9)	LA: 0.309 (5)
Das(2021)	信道编码和压缩编码	基于 CQT 的 LPS	LCNN	—	LA: 3.62 (5) PA: 29.02 (4) DF: 18.30 (8)	LA: 0.267 (3) PA: 0.767 (5)
Kang 等人(2021)	编码和信道传输	PSCC, LFCC, DCT-DFTspec, LLFB	TDNN 和 SE-ResNet-18	OC-Softmax	DF: 23.55 (11)	LA: 0.589 (10)
Lei 等人(2021)	—	STFT	SE-ResNets	CE	PA: 32.96 (6)	PA: 0.888 (6)
Müller 等人(2021)	—	CQT	ResNet, CNN, LSTM, RawNet2	BCE	LA: 10.38 (12) DF: 6.28 (3)	—
Tomilov 等人(2021)	加噪、Mixup 和 FIR 滤波器	SincConv, CQCC, MSTFT, LEAF, MFCC, LSTFT	LCNN, RawNet2, ResNet18	Center Loss	LA: 1.32 (2) PA: 26.24 (2) DF: 15.64 (6)	—
Wang 等人(2021)	OPUS 编码	对数语谱图, CQCC, LFCC	GMM, VAE	AM-Softmax	LA: 8.23 (11) PA: 24.25 (1)	LA: 0.331 (8) PA: 0.682 (2)
Yoon 和 Yu(2021)	—	LPS, 群延迟 Gram	SE-ResNet, DenseNet, ShuffleNet, MNASNet	—	PA: 29.55 (5)	PA: 0.765 (4)
Fathan 等人(2022)	平滑等47种数据增强方法	梅尔语谱图	WaveletCNN 和 VGG16	AAM Softmax	DF: 21.53 (10)	LA: 0.364 (9)
Martín-Doñas 和 Álvarez(2022)	FIR 滤波器	Wav2Vec2	前向层和线性层	AM-Softmax	LA: 3.54 (4) DF: 4.98 (2)	—
Tak 等人(2022a)	RawBoost	Sinc 滤波器	RawNet2	—	LA: 5.31 (6)	LA: 0.310 (6)
Tak 等人(2022b)	RawBoost	Sinc 滤波器, Wav2Vec2	AASIST	WCE	LA: 0.82 (1) DF: 2.85 (1)	LA: 0.207 (1)
Wang 等人(2022a)	FIR 滤波器	Wav2Vec2	light-DARTS	Contrast Loss	DF: 7.86 (4)	—
Hu 等人(2023)	—	CQT, CQCC, 对数功率幅度频谱图	HFN	BCE, MSE	PA: 24.46 (2)	PA: 0.671 (1)
Ito 和 Horiguchi (2023)	—	Wav2Vec2, HuBERT, WavLM	Conv-TasNet	AAM Softmax	LA: 7.55 (10) DF: 11.08 (5)	—
Zhang 等人(2023c)	设备增强和传输增强	LFCC	LCNN, RawNet2, ResNet18	A-ISO	LA: 6.84 (9)	LA: 0.322 (7)

注：“—”表示原论文中未提及；实验结果中括号内数值为每列在 LA、PA 和 DF 场景的排名；加粗字体表示该场景的最优结果；“↓”表示值越小模型性能越好。

ASVspoof 2021 挑战赛 LA 场景中的未知攻击方式, Cáceres 等人(2021)提出轻量级时延神经网络架构(time delay neural network, TDNN), 并采用焦点损失来处理类不平衡的问题。在 LA 任务上, 作者对原始训练数据进行多媒体传输变换和通话传输变换, 所得训练集的大小是原始训练集的 5 倍。该模型在 ASVspoof 2021 数据集的 LA 场景下获得最小 t-DCF 为 0.275, EER 为 5.58%。此外, Zhang 等人(2023c)构建了基于 ASVspoof 2019 数据集 LA 场景下的设备增强数据集, 并用于执行 3 种通道鲁棒性训练策略, 以应对声学 and 通话信道变化的鲁棒性问题。该模型在 ASVspoof 2021 数据集的 LA 场景下取得 0.322 的最小 t-DCF, 6.84% 的 EER。

3.1.4 压缩增强

在 ASVspoof 2021 数据集的 PA 场景下, 对语音数据进行 MP3、M4A 以及 OGG 等格式的压缩处理。对于上述场景, 通过压缩增强方法对被压缩的音频信号进行增强, 以提高语音的质量和还原度。

Chen 等人(2021a)首先采用 RIR 数据集和 MUSAN 数据集对原始语音数据集进行加噪, 然后采用 MP3 和 M4A 格式进行语音压缩, 最后采用 Opus 编解码数据集对原始的训练语音增加编解码器的传输效果。该系统在 ASVspoof 2021 的 LA 场景上取得了优异的检测性能, 但在 DF 场景上的泛化性仍然有待提升。

为了应对未知的伪造攻击方式, Das 等人(2021)采用基于 A 律和 μ 律的信号压缩方法对原始语音数据进行数据增广, 将训练数据量增加为原本语音数据的 3 倍。该模型在 ASVspoof 2019 数据集的 LA 场景下获得最小 t-DCF 为 0.094, EER 为 3.13%。相似地, Das(2021)采用各种已知和未知的数据增广方法对原始的语音进行处理, 包括各种音

频编解码器、数据压缩方法以及基于噪声的数据增强等。实验结果表明, 该模型在 ASVspoof 2021 数据集的 LA 场景下获得最小 t-DCF 为 0.267, EER 为 3.62%, 在 DF 场景下获得 EER 为 18.3%。

3.2 特征提取和优化

语音特征的提取直接影响着语音深度伪造检测技术的性能。如图 7 所示, 语音特征主要包括手工特征和深度特征。其中, 手工特征是从语音信号中提取的经过人工选择和设计的特征, 其具有对数据要求低、计算效率高和解释性强的特点。深度特征是指由深度网络模型在中间层或最后一层生成深层的特征表示。而分类器则根据输入的语音特征将语音判别为真实语音或伪造语音, 主要包括支持向量机(support vector machine, SVM)、随机森林(random forest)以及 DNN 等。

根据特征提取的类型和处理方式, 可以将语音深度伪造检测技术归纳为 4 类, 分别为基于手工特征的伪造检测、基于混合特征的伪造检测、基于端到端的伪造检测以及基于特征融合的伪造检测。

3.2.1 基于手工特征的伪造检测

基于手工特征的伪造检测技术以精心设计的手工特征作为分类器的输入。频谱特征是手工特征作为分类器的输入。频谱特征是基于手工特征的伪造检测技术最常用的特征之一, 其利用从语音信号中的频率信息特征来进行伪造检测。频谱特征是语音信号在频域上的表示, 反映了语音信号的频率成分、谐波结构和能量分布。频谱特征主要有语谱图、频谱包络、倒谱系数和梅尔频谱等。直接对手工特征进行分类器包括传统的有监督概率分类器和传统的机器学习分类器, 主要有 GMM、SVM、i-vectors 以及逻辑回归等。ASVspoof 2019 数据集下基于手工特征的伪造检测技术的相关研究汇总如表 6 所示。

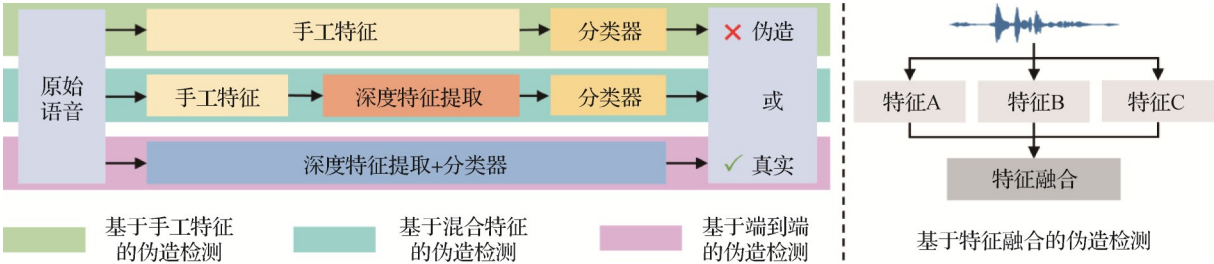


图7 基于特征提取的伪造检测分类

Fig. 7 Taxonomy of speech deepfake detection based on feature extraction

表 6 ASVspoof 2019数据集下基于手工特征的伪造检测技术汇总
Table 6 Summary of recent handcrafted feature-based forgery detection techniques on ASVspoof 2019 dataset

文献	语音深度伪造检测技术				实验结果	
	语音预处理	特征提取	网络结构	损失函数	EER/% ↓	t-DCF ↓
Yamagishi 等人(2019)	–	LFCC	GMM	–	LA: 2.71 (3) PA: 11.96 (3)	LA: 0.066 (3) PA: 0.255 (3)
Yamagishi 等人(2019)	–	CQCC	GMM	–	LA: 0.43 (1) PA: 9.87 (2)	LA: 0.012 (1) PA: 0.195 (2)
Gupta 和 Patil (2022)	–	LFRCC	GMM	BCE	PA: 17.30 (4)	–
Gupta 等人(2022)	–	CFCCIF-QESA	GMM	CE	PA: 25.71 (5)	–
Javed 等人(2022)	–	ATP-GTCC	SVM	汉明距离	LA: 0.75 (2) PA: 1.00 (1)	LA: 0.050 (2) PA: 0.064 (1)

注:“–”表示原论文中未提及;实验结果中括号内数值为每列在 LA 或 PA 场景的排名;加粗字体表示该场景的最优结果。

在 ASVspoof 2019 挑战赛中,赛事组织方提供了将线性频率倒谱系数(linear frequency cepstral coefficients, LFCC)和基于常数 Q 变换的倒谱系数(constant Q cepstral coefficients, CQCC)这两种手工特征与 GMM 分类器相结合的基准模型,分别为 LFCC-GMM 和 CQCC-GMM (Yamagishi 等, 2019)。在 ASVspoof 2019 数据集的 LA 场景中,CQCC-GMM 模型的最小 t-DCF 和 EER 分别为 0.012 和 0.43%。

Gupta 和 Patil (2022)提出了线性频率残差倒谱系数(linear frequency residual cepstral coefficients, LFRCC)的手工特征,通过调节倒谱域中线性预测(linear prediction, LP)的阶数来实现对语音重放攻击的检测。同时,为了捕获语音信号之间的相移信息,Gupta 等人(2022)利用能量分离算法(energy separation algorithm, ESA)在时间分辨率上的优点,将基于耳蜗滤波器倒谱系数的瞬时频率(cochlear filter cepstral coefficients-based instantaneous frequency, CFCCIF)特征与基于正交的能量分离算法(quadrature-based ESA, QESA)相结合,提出了 CFCCIF-QESA 特征。实验表明,在 ASVspoof 2019 数据集的 PA 场景中,CFCCIF-QESA 特征比没有正交相位的 CFCCIF-ESA 特征检测性能更佳。此外,Javed 等人(2022)提出一种声学三元共现模式(acoustic-ternary co-occurrence patterns, ATCoP)与伽马倒谱系数(gammatone cepstral coefficients, GTCC)相结合的手工特征 ATCoP-GTCC,有效地捕捉重放语音中由麦克风导致的失真,从而检测语音的多阶重放和克隆重放攻击。实验表明,该模型采用汉明

距离来优化 SVM 分类器,在高质量的伪造语音数据中取得优异的检测结果。

不同的语音特征关注语音信号中的不同信息,单一的手工特征往往在特定的伪造攻击检测中具有良好的效果。然而,在面对复杂的伪造语音攻击场景时,设计和提取手工特征需要专业的知识和经验。

3.2.2 基于混合特征的伪造检测

深度学习方法在语音深度伪造检测领域取得了显著的进展。基于混合特征的伪造检测技术将手工特征的领域知识与深度学习网络相结合,从手工特征中学习到的更丰富的特征表示。在语音深度伪造检测中,卷积神经网络(convolutional neural network, CNN)是一种常用的深度学习技术,主要有轻量卷积神经网络(light convolutional neural network, LCNN)、残差网络(residual network, ResNet)和挤压—激励网络(squeeze-and-excitation network, SENet)。在 ASVspoof 2019 数据集下基于混合特征的伪造检测技术的相关研究已汇总至表 7 中。

Lei 等人(2020)提出基于高斯概率特征的孪生 CNN 模型 Siamese CNN。该模型以 GMM 分量上语音帧的对数概率作为输入,同时关注 GMM 组件的帧分数和帧与帧之间的局部关系。Wu 等人(2020c)提出了一种基于特征真实化的 LCNN 方法。该模型以基于恒定 Q 变换(constant Q transform, CQT)的对数功率谱(log power spectrum, LPS)特征为输入,采用真实语音来训练基于 CNN 的转换器,从而拟合真实语音的特征分布来检测伪造语音。实验表明,该模型在 ASVspoof 2019 数据集的 LA 场景下获得最小

t-DCF 为 0.102, EER 为 4.07%。Yoon 和 Yu (2021) 采用各种基于 CNN 网络和特征融合方法构建语音深度伪造检测系统。该融合系统将对数功率谱和群延迟 Gram 特征以嵌入和特征图的方式进行特征拼接, 并取所有系统的平均分数为最终分数。实验表明, 该融合系统在 ASVspoof 2021 数据集的 PA 场景下获得最小 t-DCF 为 0.765, EER 为 29.55%。

ResNet 网络是语音深度伪造检测模型中最常见的网络之一。Chen 等人 (2020b) 提出一种基于

ResNet18 的检测模型。该模型将线性滤波器组系数 (linear filter bank, LFB) 特征作为输入, 使用增强边缘余弦损失函数 (large margin cosine loss, LMCL) 和频率掩码增强, 从而学习更加鲁棒的特征。Kwak 等人 (2021) 结合 ResNet 和 LCNN 的最大特征图 (max feature map, MFM) 激活函数提出了 ResMax 模型。该模型以 CQT 特征为输入, 并采用二元交叉熵 (binary cross entropy, BCE) 作为损失函数。在 ASVspoof 2019 数据集集中的 PA 场景下, 该模型获得

表 7 ASVspoof 2019 数据集下基于混合特征的伪造检测技术汇总
Table 7 Summary of recent hybrid feature-based forgery detection techniques on ASVspoof 2019 dataset

文献	语音深度伪造检测技术				实验结果	
	语音预处理	特征提取	网络结构	损失函数	EER/% ↓	t-DCF ↓
Zeinali 等人(2019)	—	CQT, LPS	VGG, SincNet	CE	LA: 8.01 (4) PA: 1.51 (2)	LA: 0.208 (4) PA: 0.037 (5)
Chen 等人(2020b)	RIR 和 MUSAN 数据集加噪、频 率掩码	LFB	ResNet18	LMCL	LA: 1.81 (4)	LA: 0.052 (4)
Lei 等人(2020)	—	LFCC	Siamese CNN	CE	LA: 3.79 (9) PA: 7.98 (5)	LA: 0.093 (9) PA: 0.195 (6)
Yang 和 Das(2020)	按一定信噪比加 噪	ICQC, ICQCC, ICBC, ICLBC	DNN	CE	LA: 7.78 (16)	LA: 0.187 (11)
Wu 等人(2020c)	—	LPS	LCNN	—	LA: 4.07 (11)	LA: 0.102 (10)
Ge 等人(2021a)	频率掩码	LFCC	PC-DARTS	WCE	LA: 4.96 (12)	LA: 0.091 (8)
Kwak 等人(2021)	—	CQT	ResMax	BCE	LA: 2.19 (7) PA: 0.37 (1)	LA: 0.060 (6) PA: 0.009 (1)
Li 等人(2021b)	—	CQT	SE-Res2Net50	BCE	LA: 2.50 (8) PA: 0.46 (2)	LA: 0.074 (7) PA: 0.012 (2)
Ma 等人(2021a)	—	LFCC	LCNN, DFWF	Similarity Loss	LA: 7.74 (15) PA: 8.85 (6)	—
Rostami 等人(2021)	SpecAug、注意力 掩码	LFCC, LPS	EfficientNet-A0, SE-Res2Net50	WCE, Triple Loss	LA: 1.89 (6) PA: 0.86 (4)	LA: 0.507 (12) PA: 0.024 (4)
Zhang 等人(2021b)	—	LFCC	ResNet18	OC-Softmax	LA: 2.19 (7)	LA: 0.059 (5)
Zhang 等人(2021c)	RIR 数据集加 噪、通话编码和 设备编码	LFCC	LCNN, ResNet-OC	OC-Softmax, CE	LA: 3.92 (10)	—
Li 等人(2022)	—	L-VQT	L-DenseNet	NLLLoss	LA: 1.54 (3)	LA: 0.045 (3)
Wang 等人(2022b)	按信噪比 15 dB 或 25 dB 加噪	CQT	SE-Res2Net34- Conformer	CE	LA: 1.85 (5)	LA: 0.060 (6)
Ma 等人(2023b)	—	LFCC, MFCC	TDNN, SENet34	LCML	LA: 0.83 (1)	LA: 0.025 (1)
Xue 等人(2023)	—	LPS, F0	ECANet, SENet	A-Softmax, KL 散 度、L2 损失	LA: 1.00 (2) PA: 0.65 (3)	LA: 0.031 (2) PA: 0.017 (3)

注: “—”表示原论文中未提及; 实验结果中括号内数值为每列在 LA 或 PA 场景的排名; 加粗字体表示该场景的最优结果。

0.37%的EER, 0.009的最小t-DCF。此外,Zhang等人(2021b)提出了基于ResNet的单分类伪造检测模型,其采用单分类的softmax(one-class softmax, OC-softmax)作为损失函数,以学习一个真实语音边界紧凑且远离伪造语音的特征空间。

SENet模型利用全局特征来自适应地调整每个通道的权重,使得网络能够更加关注与伪造检测相关的特征通道。Li等人(2021b)改进ResNet网络,提出了一种新的架构Res2Net,并与SENet网络相结合来提高伪造检测模型的泛化性。

Wang等人(2022b)引入自动语音识别网络Conformer来扩展现有的SE-Res2Net网络,提出伪造检测模型SE-Res2Net-Conformer,以进一步利用声学特征的局部模式。Conformer网络可以更好地处理语音信号的全局时域信息和局部特征模式,以提升伪造语音检测的检测性能。

除了上述研究工作外,Yadav等人(2023)进一步探索伪造检测的可解释性,提出了一种基于解耦语谱图VAE的伪造检测模型DSVAE。该模型的训练分为两个阶段,第1个阶段采用Kullback-Leibler(KL)散度损失函数来优化VAE网络合成语谱图的质量,从而学习语音数据的通用表征;第2个阶段训练解耦表征网络和无监督激活生成网络,来获得可解释的解耦表征和激活图,从而实现区分语音的真伪。

与基于手工特征的伪造检测技术相比,基于混合特征的伪造检测技术利用深度网络模型从手工特征中学习更复杂的特征表示,从而提高伪造检测的性能。

3.2.3 基于端到端的伪造检测

与前两种伪造检测技术需要手工设计和选取特征不同的是,基于端到端的伪造检测技术采用深度网络模型直接从原始语音信号中学习特征表示和训练分类器。基于端到端的伪造检测技术采用先进的信号处理和深度学习技术来检测与语音欺骗攻击相关的异常或不一致之处,如采用Sinc滤波器和预训练模型直接从原始语音信号中提取语音特征,采用CNN网络、ResNet网络和图注意力网络等进行建模和分类,以判断输入语音是否为真实语音或伪造语音。表8汇总了ASVspoof 2019数据集下基于端到端的伪造检测技术的研究工作。

在基于端到端的伪造检测技术中,CNN网络可

以直接处理原始语音信号的时域或频域表示,从中学习与伪造信息相关的特征。Gomez-Alanis等人(2019)将轻量化卷积门控递归神经网络(light convolutional gated recurrent neural network, LC-GRNN)作为深度特征提取器,采用概率线性判别分析(probabilistic linear discriminant analysis, PLDA)作为分类器。LC-GRNN结合LCNN网络和RNN网络的特点,实现语音信号的帧级别特征提取并学习上下文特征。实验表明,该模型在ASVspoof 2019数据集的LA场景下获得EER为6.28%,在PA场景下获得EER为2.23%。RW-ResNet模型(Ma等, 2021b)采用基于残差连接的一维卷积残差块从原始语音波形中学习更好的特征表示,将提取的特征输入到ResNet34网络来判断语音的真伪。实验表明,该模型在ASVspoof 2019数据集的LA场景获得最小t-DCF为0.212和EER为8.09%。

RawNet模型(Jung等, 2020)采用Sinc滤波器对原始语音信号进行处理并提取相应的特征,然后采用ResNet网络进一步挖掘更深层次的语音信息。在此基础上,Tak等人(2021b)首次将RawNet2网络应用于语音深度伪造检测领域中,直接对原始语音波形进行操作并提取语音的深度特征。接着,Wang和Hansen(2022)将RawNet2网络作为编码器提取特征,同时引入一个简单的注意力网络SimAM来计算卷积层中特征图的三维注意力权重,然后优化能量函数以确定每个神经元的重要性。该模型采用加性角边距(additive angular margin, AAM)损失函数对不平衡训练的语音数据中每一类别分配相应的权重,以提高类内的相似性和类间的相异性。同时,该模型采用均方误差(mean square error, MSE)来优化模型的元学习训练模式,以提高伪造检测模型的泛化能力。实验表明,该模型在ASVspoof 2019数据集的LA场景获得最小t-DCF为0.029, EER为0.99%。

Jung等人(2022)将RawNet2网络和图注意力网络相结合,提出一种基于端到端的伪造检测模型AASIST,该模型采用异质注意力机制和堆积节点对跨异质时域和频域的特征进行建模,在ASVspoof 2019数据集的LA场景下获得0.83%的EER和0.0275的最小t-DCF。Liu等人(2023b)在AASIST模型的基础上进行说话人感知的扩展,在语音的帧级别和段级别上引入目标说话人信息。与独立于说话人的基线相比,该模型通过引入注册的说话人信

表 8 ASVspoof 2019 数据集下基于端到端的伪造检测技术汇总
Table 8 Summary of recent end-to-end based forgery detection techniques on ASVspoof 2019 dataset

文献	语音深度伪造检测技术				实验结果	
	语音预处理	特征提取	网络结构	损失函数	EER/% ↓	t-DCF ↓
Gomez-Alanis 等人 (2019)	–	LC-GRNN	PLDA	–	LA: 6.28 (13) PA: 2.23 (1)	LA: 0.152 (10) PA: 0.061 (1)
Ge 等人(2021b)	掩码滤波器	Sinc 滤波器	PC-DARTS	P2SGrad	LA: 1.77 (10)	LA: 0.052 (7)
Hua 等人(2021)	–	DNN	Res-TSSDNet, Inc-TSSDNet	WCE	LA: 1.64 (9)	LA: 0.048 (6)
Ma 等人(2021b)	–	一维卷积残差块	ResNet	CE	LA: 2.98 (11)	LA: 0.082 (9)
Tak 等人(2021b)	–	Sinc 滤波器	RawNet2	CE	LA: 1.12 (5)	LA: 0.033 (3)
Fu 等人(2022)	–	FastAudio filters	X-vector, ECAPA-TDNN	–	LA: 1.54 (7)	LA: 0.045 (5)
Jung 等人(2022)	–	Sinc 滤波器	RawNet2, MGO, HS-GAL	CE	LA: 0.83 (2)	LA: 0.028 (1)
Wang 等人(2022a)	FIR 滤波器	Wav2Vec2	light-DARTS	Contrast loss	LA: 1.08 (4)	–
Wang 和 Hansen(2022)	–	Sinc 滤波器	RawNet2, SimAM	AAM, MSE	LA: 0.99 (3)	LA: 0.029 (2)
Sun 等人(2023)	重采样、加噪	Sinc 滤波器	RawNet2	CE, Softmax	LA: 4.54 (12)	–
Ito 和 Horiguchi(2023)	–	Wav2Vec2, HuBERT, WavLM	Conv-TasNet	AAM Softmax	LA: 0.44 (1)	–
Liu 等人(2023b)	–	Sinc 滤波器	AASIST, M2S 转换器	CE	LA: 1.13 (6)	LA: 0.038 (4)
Wang 等人(2023b)	RawBoost	Sinc 滤波器	RawNet2, TCN	CE, Orthogonal Loss	LA: 1.58 (8)	–

注：“–”表示原论文中未提及；实验结果中括号内数值为每列在 LA 场景的排名；加粗字体表示该场景的最优结果。

息有效提升了检测性能。Wang 等人(2023b)将正交卷积和时间卷积网络(temporal convolutional networks, TCN)相结合应用于 RawNet 网络中,提出了 TO-RawNet 模型。与 EER 为 4.66% 的 RawNet 模型相比,该模型的 EER 为 1.58%,降低了 66.09%,这充分证明了 TO-RawNet 模型在语音伪造检测方面的有效性。

部分研究工作从自动学习网络架构的角度进行探索,以提高语音深度伪造检测技术的性能。Ge 等人(2021b)在部分连接可微架构搜索(partially-connected differentiable architecture search, PC-DARTS)的基础上,提出了一种端到端的伪造检测模型 Raw PC-DARTS。该模型直接对原始波形进行操作,同时对网络架构和网络参数进行联合优化。

值得注意的是,采用基于自监督的预训练模型提取语音的深度特征,能够更有效地应对未知的伪

造攻击。Wang 等人(2022a)提出一种轻量化的可微分架构搜索模型 light-DARTS。该方法采用 Wav2Vec2 预训练模型提取语音特征,同时自动学习和优化由卷积运算和残差块所组成的复杂神经结构来学习深层的语音表征。light-DARTS 模型在 ASVspoof 2019 的 LA 场景中取得了优异的性能, EER 为 1.08%。Martín-Doñas 和 Álvarez(2022)采用基于 Wav2Vec2 的预训练自监督模型提取语音深度特征,并微调分类网络中不同 Transformer 层的上下文表示,以实现伪造语音的检测。Ito 和 Horiguchi(2023)从攻击者的角度训练一个增强伪造语音欺骗能力的模型,使从真实语音和增强后的伪造语音提取的说话人特征嵌入更接近,从而检验伪造检测模型从自监督学习的大规模预训练模型提升的性能是否被抵消。实验表明,攻击者可以从预训练模型中获益,也可以消除防御者从预训练模型中获得的性

能增益。

基于端到端的伪造检测技术借助深度学习模型的强大学习能力,能够取得不错的检测性能。然而,该技术的内部工作机制通常较为复杂,缺乏直观的可解释性。这意味着当伪造语音被错误判定为真实语音时,研究人员很难从模型中了解其具体原因,使得设计和调试模型变得更加困难。

3.2.4 基于特征融合的伪造检测

与单一特征的伪造检测技术不同的是,基于特征融合的伪造检测技术通过特征堆叠和加权求和等方式融合多种相同或不同的特征。表9总结了ASVspoof 2019数据集下基于特征融合的伪造检测技术的研究工作。

Arif等人(2021)提出了一种扩展的局部三元模式(extended local ternary pattern, ELTP)与LFCC相融合的特征来训练深度双向长短时记忆(deep bidirectional long short-term memory, DBiLSTM)网络,以实现真伪语音分类。ELTP可以捕获声道语音动态特征,ELTP-LFCC特征比CQCC和LFCC特征的检测性能更佳。实验结果表明,该系统在ASVspoof 2019数据集的LA场景下获得EER为0.74%,最小t-DCF为0.008。Attorresi等人(2022)采用ECAPA-TDNN网络提取说话人身份的梅尔频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)特征和韵律编码器提取梅尔语谱图中的韵律信息,然后将这两种特征进行拼接,输入到二分类模型SVM中。这两种特征的融合是对语音的物理和行为特征进行建模,可以有效检测出TTS和VC技术生成的语音。此外,

为了减少频率信息的损失,Choi等人(2022)提出以梅尔语谱图为输入,采用重叠频率分布(overlapped frequency-distributed, OFD)模型和非重叠频率分布(non-overlapped frequency-distributed, Non-OFD)模型沿着频率轴对梅尔语谱图特征进行分割,并与频率轴上平均池化和扩张后的特征相加。实验表明,这两种模型能够有效检测出伪造语音在高频和低频范围内的伪影。Wang等人(2023a)提出从声调、发音和Wav2Vec2模型来捕捉更通用的语音特征,以提高深度伪造检测模型的泛化性。该模型将HuBERT预训练模型提取的韵律特征、基于Conformer的音素识别模型提取的发音特征和Wav2Vec2预训练模型提取的特征进行拼接,并采用注意力损失和连续时序分类(connectionist temporal classification, CTC)损失的线性组合作为训练的损失函数。

一些研究学者在图网络模型的框架下进行特征融合。在AASIST模型的基础上,Tak等人(2022b)利用基于自注意力聚合层对时域和频域的特征进行融合,从而实现时域和频域特征进行互补。该模型采用加权交叉熵(weighted cross entropy, WCE)为损失函数,以最小化训练损失。实验结果表明,该模型在ASVspoof 2021数据集的LA场景中具有优异的检测性能。Liu等人(2023a)提出了一种新的伪造检测模型M2S-ADD。该模型将原始单声道音频转换为高质量立体声,其中包括两个声道信号。双分支编码器读取双通道音频并分别提取左通道和右通道的深层特征表示,融合编码器将左、右通道的注意力图

表9 ASVspoof 2019数据集下基于特征融合的伪造检测技术汇总
Table 9 Summary of recent feature fusion-based forgery detection techniques on ASVspoof 2019 dataset

文献	语音深度伪造检测技术			实验结果
	特征提取	网络结构	损失函数	EER/% ↓
Arif等人(2021)	ELTP、LFCC	DBiLSTM	—	LA: 0.74 (2)
Hassan和Javed(2021)	GTCC、MFCC、谱通量和谱中心	Bi-LSTM	—	LA: 3.05 (5)
Attorresi等人(2022)	MFCC、梅尔语谱图	ECAPA-TDNN、韵律编码器	BCE	LA: 5.39 (6)
Choi等人(2022)	梅尔语谱图、CQT	LCNN, ResNet	—	LA: 1.35 (3)、 PA: 0.35 (1)
Liu等人(2023a)	Sinc滤波器	AASIST, M2S转换器	CE	LA: 1.34 (1)
Wang等人(2023a)	音素特征、韵律特征、Wav2Vec2	LCNN, Bi-LSTM	CTC, Attention Loss	LA: 1.58 (4)

注:“—”表示原论文中未提及;实验结果中括号内数值为每列在LA场景的排名;加粗字体表示该场景的最优结果。

进行逐元素相乘来捕获互补信息。实验表明,该模型在 ASVspoof 2019 数据集的 LA 场景中的 EER 值为 1.34%。

基于特征融合的伪造检测技术通过融合多种语音特征以获得表达能力更强的特征,从而提高检测的性能。然而,该技术仍然面临着一些挑战,包括特征之间的冗余性、融合方式和权重分配等。

3.3 学习机制

为了满足实际应用场景的需求,语音深度伪造检测技术需要对各种伪造语音攻击方式具备鲁棒性,并且能够对未知的语音数据具有一定的泛化性。为此,一些研究学者从深度神经网络训练方法的角度展开研究,包括自监督学习、集成学习以及对抗训练等方法。因此,本小节将从基于自监督学习、多任务学习以及对抗训练的伪造检测等方面分析语音深度伪造检测技术的最新进展。

3.3.1 基于自监督学习的伪造检测

自监督学习对大量无标签数据进行处理并赋予相应的代理任务标签信息,实现对网络模型进行监督训练,以学习到对下游任务有价值的数据表征。

Jiang 等人(2020)提出了基于自监督学习的语音伪造检测模型 SSAD。该模型由编码器和两种工作器组成。SSAD 的编码器采用基于 TCN 网络的架构,提取语音信号的深层特征表示,作为回归工作器和二元工作器的输入。回归工作器用于预测 LPS、LFCC 和 CQCC 特征。而二元工作器用于区分真实语音和伪造语音。通过这两种工作器,编码器能够探索更高容量的高级特征。与 SENet12 网络或 LCNN 网络分类器相结合时,SSAD 模型展现出良好的泛化能力和可移植性。此外,Shim 等人(2019)针对 PA 场景的数据集中信道不匹配的问题,提出了一种自监督的重放检测预训练模型。该模型从语音数据集中选取一对语音片段,训练深层神经网络判断这对片段的声学结构是否相同。

基于自监督学习的伪造检测技术利用大量无标签数据进行训练,往往面临训练时间长和计算资源消耗大的问题。

3.3.2 基于对抗训练的伪造检测

基于对抗训练的伪造检测技术将对抗样本添加到训练数据中以增强模型的鲁棒性和泛化性。在语音伪造检测领域中,对抗样本是指对语音数据进行针对性的扰动而生成的语音样本,以欺骗伪造检测

模型并使其产生误判或错误分类。

Wu 等人(2020a)采用对抗训练方法将对抗样本添加到 ASVspoof 2019 训练数据集中,以增强基于 DNN 网络的 ASV 系统。同时,该系统引入了空间平滑技术,以提高深度伪造检测模型的鲁棒性。实验结果表明,基于 SENet 网络的 ASV 系统采用对抗训练后其检测准确率提升了 44.08%。在此基础上,Wu 等人(2020b)使用空间平滑、过滤以及对抗性训练等方法提出一种自监督学习的伪造检测模型 Mockingjay,以保护模型免受对抗样本的攻击。实验表明,自监督学习模型提取的语音特征可以防止对抗样本的可转移性。此外,Zhang 等人(2023c)将通道分类损失的梯度进行反向传播且符号相反,以优化 LFCC 特征的嵌入网络。该模型采用角孤立损失(angular isolate loss, A-ISO)为损失函数,通过最大化嵌入网络和最小化通道分类器的通道分类误差,来构建对抗性训练范式。

值得注意的是,基于对抗训练的伪造检测技术能够有效对抗已知的伪造语音攻击方式。然而,面对未知的对抗攻击方式时,伪造检测模型的鲁棒性较弱。

3.3.3 基于多任务学习的伪造检测

与传统基于单一的二分类伪造检测技术不同的是,基于多任务学习的伪造检测技术可以共享底层的特征表示,从不同的语音相关任务中学习更加全面和有用的语音特征信息,高效利用有限的语音训练数据的同时提升模型的检测性能。与语音相关的任务主要有说话人识别、伪造攻击方式识别等。表 10 汇总了 ASVspoof 2019 数据集下基于多任务学习的伪造检测技术的相关工作。

Ranjan 等人(2022)提出了一个基于深度卷积的多任务学习模型 STATNet。该模型首先通过 RawNet2 网络提取语音的特征图,然后采用时间卷积模块(temporal convolutional module, TCM)和频谱卷积模块(spectral convolutional module, SCM)分别提取语音的时域和频域特征图,最后通过真伪语音的二分类任务和语音伪造方式的分类任务来实现语音真伪检测和伪造攻击的溯源。

Teng 等人(2022)提出了无集成的端到端检测模型 SA-SASV 并设计了一个基于多任务分类器的 SASV 系统。SA-SASV 模型采用 BCE 损失函数优化真伪语音的二分类任务;加性角边际 Softmax(addi-

表 10 ASVspoof 2019数据集下基于多任务学习的伪造检测技术汇总
Table 10 Summary of recent multi-task learning-based forgery detection techniques on ASVspoof 2019 dataset

文献	语音深度伪造检测技术			实验结果	
	特征提取	网络结构	损失函数	EER/% ↓	t-DCF ↓
Zhang 等人(2021a)	LFCC	SELCNN、Bi-LSTM	P2SGrad	LA: 5.90 (5)	—
Ranjan 等人(2022)	Sinc 滤波器	RawNet2、TCM、SCM	CE	LA: 2.45 (3)	LA: 0.062 (2)
Teng 等人(2022)	Fbanks、Sinc 滤波器	ECAPA-TDNN、ARawNet	BCE、AAM Softmax、CE、Triple Loss	LA: 4.86 (4)	—
Zhang 等人(2022b)	梅尔滤波器、Sinc 滤波器	ECAPA-TDNN、AASIST	BCE	LA: 1.53 (2)	—
Ma 等人(2023a)	LFCC、CQT	Dual-Branch Network	分类损失与伪造类型分类损失	LA: 0.80 (1) LA: 0.021 (1)	
Yadav 等人(2023)	语谱图	VAE	KL 散度、BCE	LA: 6.56 (6)	—

注：“—”表示原论文中未提及；实验结果中括号内数值为每列在 LA 场景的排名，加粗字体表示该场景的最优结果。

tive angular margin softmax, AAM softmax) 损失函数优化 ASV 系统的说话人分类任务；交叉熵 (cross entropy, CE) 损失函数优化不同 TTS 和 VC 攻击的分类任务以及三元损失 (triple loss) 函数来优化 TTS、VC 和说话人的分类任务。为了进一步提升伪造检测模型的泛化能力, Ma 等人 (2023a) 提出一种基于端到端的双分支网络模型, 实现语音的真伪检测和伪造攻击的分类。该模型将 LFCC 和 CQT 特征输入到卷积注意力网络中, 并采用多任务学习使网络能够检测不同伪造类型的语音。

基于多任务学习的伪造检测技术通过学习多个与语音鉴伪相关的任务, 来提升伪造检测模型的性能, 然而, 多任务学习意味着模型训练的计算开销增大且增加了实际部署的复杂性。同时, 真实语音和伪造语音数据之间的不平衡, 也可能导致模型具有偏向性。

4 主要挑战和未来展望

随着语音大模型相继推出, 语音深度伪造检测领域得到了更多关注。目前, 语音深度伪造检测技术在特定数据集上都取得了显著性能。根据已有的研究方法和最新的研究思路, 本节对语音深度伪造检测技术的主要挑战与未来展望进行分析。

1) 语音深度伪造数据集的构建。SpeechGPT 和 Voicebox 等语音大模型生成的语音具有高质量、多语种以及风格多样化的特点, 现有主流的伪造检测技术在这些生成语音的评估和测试方面存在一定的

挑战。例如, 目前大多数语音深度伪造检测数据集都是在受控环境下收集的, 与现实条件下记录或生成的真实语音不太匹配。同时, 大多数语音伪造数据集都存在数据不平衡的问题。为了有效应对现实场景下可能的安全威胁, 设计并开发与语音伪造技术同步发展的语音深度伪造检测技术至关重要。然而, 这一目标的关键在于建立符合自然场景、数据相对平衡的基准数据集。语音大模型可以在跨语种、跨说话人、跨情感以及内容的添加与删减等诸多场景下合成伪造语音。同时, 应该考虑在社交媒体平台、互联网或电话渠道等实际环境下收集语音。针对上述这些场景和实际环境下, 如何构造高质量和多样化的语音深度伪造数据集是一个值得深入思考和研究的问题。

2) 伪造检测技术的鲁棒性。语音伪造检测模型在训练过程中容易出现过拟合的问题。在现实世界中, 语音信号往往受到噪声和动态环境的影响。语音伪造检测技术在干净伪造语音的检测方面表现出色, 但是在噪声、混响和信道效应等干扰因素的影响下, 现有语音伪造检测技术的检测性能可能会严重下降。因为这些干扰因素会改变语音信号的特征, 增加了伪造检测的难度。为了提升语音伪造检测技术在复杂的声学环境条件下的鲁棒性, 需要进一步研究和设计更加贴近真实应用场景的数据增强方法。此外, 采用对抗训练的方法可以提升模型的鲁棒性, 但是仍有很大的改进空间。如何更高效地进行对抗训练是值得深入研究的。

3) 伪造检测技术的泛化性。现有的语音深度伪

造检测技术在训练数据类型上具有相对单一、相同分布的特点。当面对不同的数据分布、未知的伪造攻击方式或未知的复杂声学条件时,现有的语音深度伪造检测技术存在检测性能不佳和泛化能力较差的问题。对跨类型的伪造语音进行评估,可以检验伪造检测技术的泛化性,并为更准确、更安全的ASV系统提供额外的信息和方向。虽然一些研究学者在泛化性问题上取得了不错的研究成果,但是该问题仍然是语音伪造检测领域的核心难点之一,也是实际应用部署的关键环节。在未来的研究中,可以进一步探索基于小样本的自监督学习方法,以保证伪造检测性能的前提下降低模型训练的成本;进一步研究持续学习方法,以不断接收新的语音数据进行增量学习和更新伪造检测模型;进一步探索为每个任务定义相应的损失函数,以指导多任务的学习,从而改进伪造检测技术的泛化能力。

4) 伪造检测技术的可解释性。可解释人工智能(explainable artificial intelligence, XAI)是一个新兴领域,旨在提高人工智能系统的可解释性和可理解性。在语音深度伪造检测领域,XAI可以帮助研究者、开发者和决策者设计可解释的伪造检测模型和系统,从而增强用户对于伪造检测结果的理解和信任。在实际应用中提供检测结果的可解释性具有重要意义。例如,语音取证和归因需要对伪造检测结果进行解释,以便于法律和司法机构等相关方能够理解和接受检测结果的可靠性。了解语音为何被判别为伪造语音,以及了解伪造语音的生成工具或深度伪造方法,对进一步的调查和处理至关重要。目前大多数语音深度伪造检测技术都是黑盒,存在可解释性弱、信任度低等问题。伪造检测技术的可解释性仍然是研究的一大难点问题。值得注意的是,基于CNN网络、图神经网络和胶囊网络等的可解释研究在图像分类领域中取得了一定的进展,借鉴这些领域的研究并结合语音深度伪造检测技术的特点,设计基于可解释的语音深度伪造检测技术是未来的核心研究方向之一。

5 结 语

本文对语音深度伪造及其检测技术的研究进展进行梳理回顾。具体而言,首先简要介绍了两种主流的语音深度伪造技术,分别为语音合成和语音转

换。然后,汇总了近年来主要的语音深度伪造检测的数据集以及相关的评价指标。为了方便研究学者开展研究,从数据增强、特征提取和优化以及学习机制等视角对现有的语音深度伪造检测技术进行详细的介绍。最后,从语音深度伪造数据集的构建和伪造检测技术的鲁棒性、泛化性以及可解释性这4个方面,对未来的研究方向进行展望。

参考文献(References)

- Aihara R, Takiguchi T and Ariki Y. 2013. Individuality-preserving voice conversion for articulation disorders using locality-constrained NMF//Proceedings of the 4th Workshop on Speech and Language Processing for Assistive Technologies. Grenoble, France: Association for Computational Linguistics: 3-8
- Almutairi Z and Elgibreen H. 2022. A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms*, 15(5): #155 [DOI: 10.3390/a15050155]
- Arif T, Javed A, Alhameed M, Jeribi F and Tahir A. 2021. Voice spoofing countermeasure for logical access attacks detection. *IEEE Access*, #9: 162857-162868 [DOI: 10.1109/ACCESS.2021.3133134]
- Arik S Ö, Chrzanowski M, Coates A, Diamos G, Gibiansky A, Kang Y G, Li X, Miller J, Ng A, Raiman J, Sengupta S and Shoeybi M. 2017a. Deep voice: real-time neural text-to-speech//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org: 195-204 [DOI: 10.5555/3305381.3305402]
- Arik S Ö, Diamos G, Gibiansky A, Miller J, Peng K N, Ping W, Raiman J and Zhou Y Q. 2017b. Deep voice 2: multi-speaker neural text-to-speech//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 2966-2974 [DOI: 10.5555/3294996.3295056]
- Attorresi L, Salvi D, Borrelli C, Bestagini P and Tubaro S. 2022. Combining automatic speaker verification and prosody analysis for synthetic speech detection//Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges. Montréal, Canada: Springer-Verlag: 247-263 [DOI: 10.1007/978-3-031-37742-6_21]
- Ba Z J, Wen Q, Cheng P, Wang Y W, Lin F, Lu L and Liu Z G. 2023. Transferring audio deepfake detection capability across languages//Proceedings of 2023 ACM Web Conference. Austin, USA: ACM: 2033-2044 [DOI: 10.1145/3543507.3583222]
- Bevinamarad P R and Shirdonkar M S. 2020. Audio forgery detection techniques: present and past review//Proceedings of the 4th International Conference on Trends in Electronics and Informatics (ICOEI) (48184). Tirunelveli, India: IEEE: 613-618 [DOI: 10.1109/ICOEI48184.2020.9143014]
- Bińkowski M, Donahue J, Dieleman S, Clark A, Elsen E, Casagrande

- N, Cobo L C and Simonyan K. 2019. High fidelity speech synthesis with adversarial networks//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia: ICLR
- Cáceres J, Font R, Grau T and Molina J. 2021. The biometric vox system for the ASVspoof 2021 challenge//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [s.l.] : ISCA: 68-74 [DOI: 10.21437/ASVSPPOOF.2021-11]
- Cai Z X and Li M. 2022. Invertible voice conversion [EB/OL]. [2023-06-30]. <http://arxiv.org/pdf/2201.10687.pdf>
- Chen N X, Zhang Y, Zen H G, Weiss R J, Norouzi M and Chan W. 2020a. WaveGrad: estimating gradients for waveform generation//Proceedings of the 9th International Conference on Learning Representations. Virtual Event: ICLR
- Chen T X, Khoury E, Phatak K and Sivaraman G. 2021a. Pindrop Labs' submission to the ASVspoof 2021 challenge//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [s.l.]: ISCA: 89-93 [DOI: 10.21437/ASVSPPOOF.2021-14]
- Chen T X, Kumar A, Nagarsheth P, Sivaraman G and Khoury E. 2020b. Generalization of audio deepfake detection//The Speaker and Language Recognition Workshop (Odyssey 2020). Tokyo, Japan: ISCA: 132-137 [DOI: 10.21437/Odyssey.2020-19]
- Chen X H, Zhang Y, Zhu G and Duan Z Y. 2021b. UR channel-robust synthetic speech detection system for ASVspoof 2021//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [s.l.]: ISCA: 75-82 [DOI: 10.21437/ASVSPPOOF.2021-12]
- Chen Y N, Chu M, Chang E, Liu J and Liu R S. 2003. Voice conversion with smoothed GMM and MAP adaptation//Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003). Geneva, Switzerland: ISCA: 2413-2416 [DOI: 10.21437/Eurospeech.2003-664]
- Choi S, Kwak I Y and Oh S. 2022. Overlapped frequency-distributed network: frequency-aware voice spoofing countermeasure//Proceedings of the 23rd Annual Conference of the International Speech Communication Association. Incheon, Korea (South) : ISCA: 3558-3562 [DOI: 10.21437/Interspeech.2022-657]
- Chou J C and Lee H Y. 2019. One-shot voice conversion by separating speaker and content representations with instance normalization//Proceedings of Interspeech 2019, the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA: 664-668 [DOI: 10.21437/Interspeech.2019-2663]
- Cong J, Yang S, Xie L and Su D. 2021. Glow-WaveGAN: learning speech representations from GAN-based variational auto-encoder for high fidelity flow-based speech synthesis//Proceedings of Interspeech 2021, the 22nd Annual Conference of the International Speech Communication Association. Brno, Czechia: 2182-2186 [DOI: 10.21437/Interspeech.2021-414]
- Cohen A, Rimón I, Aflalo E and Permuter H H. 2022. A study on data augmentation in voice anti-spoofing. *Speech Communication*, 141: 56-67 [DOI: 10.1016/j.specom.2022.04.005]
- Das R K. 2021. Known-unknown data augmentation strategies for detection of logical access, physical access and speech deepfake attacks: ASVspoof 2021//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [s.l.]: ISCA: 29-36 [DOI: 10.21437/ASVSPPOOF.2021-5]
- Das R K, Yang J C and Li H Z. 2021. Data augmentation with signal companding for detection of logical access attacks//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 6349-6353 [DOI: 10.1109/ICASSP39728.2021.9413501]
- Delgado H, Evans N, Kinnunen T, Lee K A, Liu X C, Nautsch A, Patino J, Sahidullah M, Todisco M, Wang X and Yamagishi J. 2021. ASVspoof 2021: automatic speaker verification spoofing and countermeasures challenge evaluation plan [EB/OL]. [2023-06-30]. <https://arxiv.org/pdf/2109.00535.pdf>
- Dhar S, Jana N D and Das S. 2023. An adaptive-learning-based generative adversarial network for one-to-one voice conversion. *IEEE Transactions on Artificial Intelligence*, 4(1): 92-106 [DOI: 10.1109/TAI.2022.3149858]
- Dixit A, Kaur N and Kingra S. 2023. Review of audio deepfake detection techniques: issues and prospects. *Expert Systems*, 40(8): #e13322 [DOI: 10.1111/exsy.13322]
- Donahue C, McAuley J and Puckette M. 2018. Adversarial audio synthesis//Proceedings of the 7th International Conference on Learning Representations. OrleansNew, USA: ICLR
- Elias I, Zen H G, Shen J, Zhang Y, Jia Y, Weiss R J and Wu Y H. 2021. Parallel tacotron: non-autoregressive and controllable TTS//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 5709-5713 [DOI: 10.1109/ICASSP39728.2021.9414718]
- Ergünay S K, Khoury E, Lazaridis A and Marcel S. 2015. On the vulnerability of speaker verification to realistic voice spoofing//Proceedings of the 7th International Conference on Biometrics Theory, Applications and Systems (BTAS). Arlington, USA: IEEE: 1-6 [DOI: 10.1109/BTAS.2015.7358783]
- Fathan A, Alam J and Kang W H. 2022. Mel-spectrogram image-based end-to-end audio deepfake detection under channel-mismatched conditions//Proceedings of 2022 IEEE International Conference on Multimedia and Expo (ICME). Taipei, China: IEEE: 1-6 [DOI: 10.1109/ICME52920.2022.9859621]
- Frank J and Schönherr L. 2021. WaveFake: a data set to facilitate audio deepfake detection//Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021). [s.l.]: [s.n.]
- Fu Q C, Teng Z W, White J, Powell M E and Schmidt D C. 2022.

- FastAudio: a learnable audio front-end for spoof speech detection// Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 3693-3697 [DOI: 10.1109/ICASSP43922.2022.9746722]
- Ge W Y, Panariello M, Patino J, Todisco M and Evans N. 2021a. Partially-connected differentiable architecture search for deepfake and spoofing detection//Proceedings of the 22nd Interspeech Annual Conference of the International Speech Communication Association. Brno, Czechia; ISCA: 4319-4323 [DOI: 10.21437/Interspeech.2021-1187]
- Ge W Y, Patino J, Todisco M and Evans N. 2021b. Raw differentiable architecture search for speech deepfake and spoofing detection [EB/OL]. [2023-06-30]. <http://arxiv.org/pdf/2107.12212.pdf>
- Gomez-Alanis A, Peinado A M, Gonzalez J A and Gomez A M. 2019. A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria; ISCA: 1068-1072 [DOI: 10.21437/Interspeech.2019-2212]
- Gong Y, Yang J, Huber J, MacKnight M and Poellabauer C. 2019. ReMASC: realistic replay attack corpus for voice controlled systems//Proceedings of Interspeech 2019, the 20th Annual Conference of the International Speech Communication Association. Graz, Austria; ISCA: 2355-2359 [DOI: 10.21437/Interspeech.2019-1541]
- Griffin D and Lim J. 1984. Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2): 236-243 [DOI: 10.1109/TASSP.1984.1164317]
- Guo H J, Liu C R, Ishi C T and Ishiguro H. 2023. QuickVC: any-to-many voice conversion using inverse short-time fourier transform for faster conversion [EB/OL]. [2023-06-30]. <https://arxiv.org/pdf/2302.08296v4.pdf>
- Gupta P, Chodingala P K and Patil H A. 2022. Energy separation based instantaneous frequency estimation from quadrature and in-phase components for replay spoof detection//Proceedings of the 30th European Signal Processing Conference (EUSIPCO). Belgrade, Serbia; IEEE: 369-373 [DOI: 10.23919/EUSIPCO55093.2022.9909533]
- Gupta P and Patil H A. 2022. Linear frequency residual cepstral features for replay spoof detection on ASVspoof 2019//Proceedings of the 30th European Signal Processing Conference (EUSIPCO). Belgrade, Serbia; IEEE: 349-353 [DOI: 10.23919/EUSIPCO55093.2022.9909913]
- Hassan F and Javed A. 2021. Voice spoofing countermeasure for synthetic speech detection//Proceedings of 2021 International Conference on Artificial Intelligence (ICAI). Islamabad, Pakistan; IEEE: 209-212 [DOI: 10.1109/ICAI52203.2021.9445238]
- Helander E, Virtanen T, Nurminen J and Gabbouj M. 2010. Voice conversion using partial least squares regression. IEEE Transactions on Audio, Speech, and Language Processing, 18(5): 912-921 [DOI: 10.1109/TASL.2010.2041699]
- Hsu W N, Zhang Y, Weiss R J, Zen H G, Wu Y H, Wang Y X, Cao Y, Jia Y, Chen Z F, Shen J, Nguyen P and Pang R M. 2018. Hierarchical generative modeling for controllable speech synthesis//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA; ICLR
- Hu C L, Zhou R H and Yuan Q S. 2023. Replay speech detection based on dual-input hierarchical fusion network. Applied Sciences, 13(9): #5350 [DOI: 10.3390/app13095350]
- Hua G, Teoh A B J and Zhang H J. 2021. Towards end-to-end synthetic speech detection. IEEE Signal Processing Letters, 28: 1265-1269 [DOI: 10.1109/LSP.2021.3089437]
- Huang W C, Hayashi T, Watanabe S and Toda T. 2020. The sequence-to-sequence baseline for the voice conversion challenge 2020: cascading ASR and TTS//Proceedings Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020. Shanghai, China; ISCA: 160-164 [DOI: 10.21437/VCCBC.2020-24]
- Huang R J, Lam M W Y, Wang J, Su D, Yu D, Ren Y and Zhao Z. 2022. Fastdiff: a fast conditional diffusion model for high-quality speech synthesis//Proceedings of the 31st International Joint Conference on Artificial Intelligence Main Track. Vienna, Austria; IJCAI: 4157-4163 [DOI: 10.24963/ijcai.2022/577]
- Hunt A J and Black A W. 1996. Unit selection in a concatenative speech synthesis system using a large speech database//Proceedings of 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Atlanta, USA; IEEE: 373-376 [DOI: 10.1109/ICASSP.1996.541110]
- Ito A and Horiguchi S. 2023. Spoofing attacker also benefits from self-supervised pretrained model//Proceedings of Interspeech 2023. Dublin, Ireland; ISCA: 5346-5350 [DOI: 10.21437/Interspeech.2023-270]
- Javed A, Malik K M, Malik H and Irtaza A. 2022. Voice spoofing detector: a unified anti-spoofing framework. Expert Systems with Applications, 198: #116770 [DOI: 10.1016/j.eswa.2022.116770]
- Jeong M, Kim H, Cheon S J, Choi B J and Kim N S. 2021. Diff-TTS: a denoising diffusion model for text-to-speech//Proceedings of Interspeech 2021. Brno, Czechia; ISCA: 3605-3609 [DOI: 10.21437/Interspeech.2021-469]
- Jiang Z Y, Zhu H C, Peng L, Ding W B and Ren Y Z. 2020. Self-supervised spoofing audio detection scheme//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China; ISCA: 4223-4227 [DOI: 10.21437/Interspeech.2020-1760]
- Jung J W, Heo H S, Tak H, Shim H J, Chung J S, Lee B J, Yu H J and Evans N. 2022. AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore; IEEE: 6367-6371

- [DOI: 10.1109/ICASSP43922.2022.9747766]
- Jung J W, Kim S B, Shim H J, Kim J H and Yu H J. 2020. Improved RawNet with feature map scaling for text-independent speaker verification using raw waveforms//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 1496-1500 [DOI: 10.21437/Interspeech.2020-1011]
- Kalchbrenner N, Elsen E, Simonyan K, Noury S, Casagrande N, Lockhart E, Stimberg F, van den Oord A, Dieleman S and Kavukcuoglu K. 2018. Efficient neural audio synthesis//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 2410-2419
- Kamble M R, Sailor H B, Patil H A and Li H Z. 2020. Advances in anti-spoofing: from the perspective of ASVspoof challenges. APSIPA Transactions on Signal and Information Processing, 9 (1) : #21 [DOI: 10.1017/ATSIP.2019.21]
- Kameoka H, Kaneko T, Tanaka K and Hojo N. 2018. StarGAN-VC: non-parallel many-to-many voice conversion using star generative adversarial networks//2018 IEEE Spoken Language Technology Workshop (SLT). Athens, Greece: IEEE: 266-273 [DOI: 10.1109/SLT.2018.8639535]
- Kameoka H, Kaneko T, Tanaka K and Hojo N. 2019. ACVAE-VC: non-parallel voice conversion with auxiliary classifier variational autoencoder. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27 (9) : 1432-1443 [DOI: 10.1109/TASLP.2019.2917232]
- Kameoka H, Tanaka K, Kwaśny D, Kaneko T and Hojo N. 2020. ConvS2S-VC: fully convolutional sequence-to-sequence voice conversion. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 28: 1849-1863 [DOI: 10.1109/TASLP.2020.3001456]
- Kaneko T and Kameoka H. 2018. CycleGAN-VC: non-parallel voice conversion using Cycle-consistent adversarial networks//Proceedings of the 26th European Signal Processing Conference (EUSIPCO). Roma, Italy: IEEE: 2100-2104 [DOI: 10.23919/EUSIPCO.2018.8553236]
- Kaneko T, Kameoka H, Tanaka K and Hojo N. 2019a. CycleGAN-VC2: improved CycleGan-based non-parallel voice conversion//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE: 6820-6824 [DOI: 10.1109/ICASSP.2019.8682897]
- Kaneko T, Kameoka H, Tanaka K and Hojo N. 2019b. StarGAN-VC2: rethinking conditional methods for StarGAN-based voice conversion//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA: 679-683 [DOI: 10.21437/Interspeech.2019-2236]
- Kaneko T, Kameoka H, Tanaka K and Hojo N. 2020. CycleGAN-VC3: examining and improving CycleGan-VCs for mel-spectrogram conversion//Proceedings of the 21st Interspeech Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 2017-2021
- Kang W H, Alam J and Fathan A. 2021. CRIM's system description for the ASVspoof2021 challenge//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [s.l.] : ISCA: 100-106 [DOI: 10.21437/ASVSPOOF.2021-16]
- Kawahara H. 2006. STRAIGHT, exploitation of the other aspect of VOCODER: perceptually isomorphic decomposition of speech sounds. Acoustical Science and Technology, 27 (6) : 349-353 [DOI: 10.1250/ast.27.349]
- Khanjani Z, Watson G and Janeja V P. 2023. Audio deepfakes: a survey. Frontiers in Big Data, 5: #1001063 [DOI: 10.3389/fdata.2022.1001063]
- Kim J, Kim S, Kong J and Yoon S. 2020. Glow-TTS: a generative flow for text-to-speech via monotonic alignment search//Proceedings of the 33rd Advances in Neural Information Processing Systems. 33: 8067-8077
- Kim J, Kong J and Son J. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech//Proceedings of the 38th International Conference on Machine Learning. [s.l.] : PMLR: 5530-5540
- Kingma D P and Dhariwal P. 2018. Glow: generative flow with invertible 1×1 convolutions//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 10236-10245 [DOI: 10.5555/3327546.3327685]
- Kinnunen T, Lee K A, Delgado H, Evans N W D, Todisco M, Sahidullah M, Yamagishi J and Reynolds D A. 2019. t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification//2018 Speaker and Language Recognition Workshop, Odyssey 2018. Les Sables d'Olonne, France: 312-319 [DOI: 10.21437/Odyssey.2018-44]
- Kinnunen T, Sahidullah M, Delgado H, Todisco M, Evans N W D, Yamagishi J and Lee K A. 2017. The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection//Proceedings of the 18th Interspeech Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA: 2-6 [DOI: 10.21437/Interspeech.2017-1111]
- Kong J, Kim J and Bae J. 2020. HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 17022-17033 [DOI: 10.5555/3495724.3497152]
- Kong Z F, Ping W, Huang J J, Zhao K X and Catanzaro B. 2021. Diff-Wave: a versatile diffusion model for audio synthesis//Proceedings of the 9th International Conference on Learning Representations. Virtual Event: ICLR
- Kwak I Y, Kwag S, Lee J, Huh J H, Lee C H, Jeon Y, Hwang J and

- Yoon J W. 2021. ResMax: detecting voice spoofing attacks with residual network and max feature map//Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE: 4837-4844 [DOI: 10.1109/ICPR48806.2021.9412165]
- Le M, Vyas A, Shi B W, Karrer B, Sari L, Moritz R, Williamson M, Manohar V, Adi Y, Mahadeokar J and Hsu W N. 2023. Voicebox: text-guided multilingual universal speech generation at scale [EB/OL]. [2023-08-28]. <http://arxiv.org/pdf/2306.15687.pdf>
- Lee S G, Ping W, Ginsburg B, Catanzaro B and Yoon S. 2023. Big-VGAN: a universal neural vocoder with large-scale training//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: ICLR
- Lei Y, Huo X, Jiao Y Z and Li Y K. 2021. Deep metric learning for replay attack detection//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [s.l.]: ISCA: 42-46 [DOI: 10.21437/ASVSP00F.2021-7]
- Lei Y, Yang S, Cong J, Xie L and Su D. 2022. Glow-WaveGAN 2: high-quality zero-shot text-to-speech synthesis and any-to-any voice conversion//Proceedings of the 23rd Interspeech Annual Conference of the International Speech Communication Association. Incheon, Korea (South): ISCA: 2563-2567 [DOI: 10.21437/Interspeech.2022-684]
- Lei Z C, Yang Y G, Liu C H and Ye J H. 2020. Siamese convolutional neural network using Gaussian probability feature for spoofing speech detection//Proceedings of the 21st Interspeech Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 1116-1120 [DOI: 10.21437/Interspeech.2020-2723]
- Li J L, Wang H X, He P S, Abdullahi S M and Li B. 2022. Long-term variable Q transform: a novel time-frequency transform algorithm for synthetic speech detection. Digital Signal Processing, 120: #103256 [DOI: 10.1016/j.dsp.2021.103256]
- Li N H, Liu S J, Liu Y Q, Zhao S and Liu M. 2019. Neural speech synthesis with Transformer network//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI Press: 6706-6713 [DOI: 10.1609/aaai.v33i01.33016706]
- Li T L, Liu Y C, Hu C X and Zhao H. 2021a. CVC: contrastive learning for non-parallel voice conversion//Proceedings of the 22nd Interspeech Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 1324-1328 [DOI: 10.21437/Interspeech.2021-137]
- Li X, Li N, Weng C, Liu X Y, Su D, Yu D and Meng H L. 2021b. Replay and synthetic speech detection with Res2Net architecture//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 6354-6358 [DOI: 10.1109/ICASSP39728.2021.9413828]
- Li X L, Yu N H, Zhang X P, Zhang W M, Li B, Lu W, Wang W and Liu X L. 2021. Overview of digital media forensics technology. Journal of Image and Graphics, 26(6): 1216-1226 (李晓龙, 俞能海, 张新鹏, 张卫明, 李斌, 卢伟, 王伟, 刘晓龙. 2021. 数字媒体取证技术综述. 中国图象图形学报, 26(6): 1216-1226) [DOI: 10.11834/jig.210081]
- Lian Z, Wen Z Q, Zhou X Y, Pu S B, Zhang S K and Tao J H. 2020. ARVC: an auto-regressive voice conversion system without parallel training data//Proceedings of the 21st Interspeech Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 4706-4710 [DOI: 10.21437/Interspeech.2020-1715]
- Lin J H, Lin Y Y, Chien C M and Lee H Y. 2021b. S2VC: a framework for any-to-any voice conversion with self-supervised pretrained representations//Proceedings of the 22nd Interspeech Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 836-840 [DOI: 10.21437/Interspeech.2021-1356]
- Lin Y Y, Chien C M, Lin J H, Lee H Y and Lee L S. 2021a. FragmentVC: any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 5939-5943 [DOI: 10.1109/ICASSP39728.2021.9413699]
- Liu R, Zhang J H, Gao G L and Li H Z. 2023a. Betray oneself: a novel audio deepfake detection model via mono-to-stereo conversion [EB/OL]. [2023-06-30]. <https://arxiv.org/pdf/2305.16353v1.pdf>
- Liu X C, Sahidullah M, Lee K A and Kinnunen T. 2023b. Speaker-aware anti-spoofing//Proceedings of Interspeech 2023, the Annual Conference of the International Speech Communication Association. Dublin, Ireland: ISCA: 2498-2502 [DOI: 10.21437/Interspeech.2023-1323]
- Liu Z J, Guo Y W and Yu K. 2023c. DiffVoice: text-to-speech with latent diffusion//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSP49357.2023.10095100]
- Luo R Q, Tan X, Wang R, Qin T, Li J Z, Zhao S, Chen E H and Liu T Y. 2021. Lightspeech: lightweight and fast text to speech with neural architecture search//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 5699-5703 [DOI: 10.1109/ICASSP39728.2021.9414403]
- Ma H X, Yi J Y, Tao J H, Bai Y, Tian Z K and Wang C L. 2021a. Continual learning for fake audio detection//Proceedings of the 22nd Interspeech Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 886-890 [DOI: 10.21437/Interspeech.2021-794]
- Ma H X, Yi J Y, Wang C L, Yan X R, Tao J H, Wang T, Wang S M, Xu L and Fu R B. 2022. FAD: a Chinese dataset for fake audio detection//Proceedings of the 36th Conference on Neural Informa-

- tion Processing Systems (NeurIPS 2022). [s.l.] : Zenodo: #6635521 [DOI: 10.5281/zenodo.6635521]
- Ma K J, Feng Y F, Chen B J and Zhao G Y. 2023a. End-to-end dual-branch network towards synthetic speech detection. *IEEE Signal Processing Letters*, 30: 359-363 [DOI: 10.1109/LSP. 2023. 3262419]
- Ma Y X, Ren Z Z and Xu S G. 2021b. RW-ResNet: a novel speech anti-spoofing model using raw waveform//*Proceedings of the 22nd Interspeech Annual Conference of the International Speech Communication Association*. Brno, Czechia: ISCA: 4144-4148 [DOI: 10.21437/Interspeech.2021-438]
- Ma X Y, Zhang S S, Huang S, Gao J, Hu Y and He L. 2023b. How to boost anti-spoofing with X-vectors//*Proceedings of 2022 IEEE Spoken Language Technology Workshop (SLT)*. Doha, Qatar: IEEE: 593-598 [DOI: 10.1109/SLT54892.2023.10022504]
- Mandalapu H, Ramachandra R and Busch C. 2021. Smartphone audio replay attacks dataset//*Proceedings of 2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*. Rome, Italy: IEEE: 1-6 [DOI: 10.1109/IWBF50991.2021.9465096]
- Martín-Doñas J M and Álvarez A. 2022. The vicomtech audio deepfake detection system based on Wav2vec2 for the 2022 ADD Challenge//*Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Singapore, Singapore: IEEE: 9241-9245 [DOI: 10.1109/ICASSP43922.2022.9747768]
- Mittal A and Dua M. 2022. Automatic speaker verification systems and spoof detection techniques: review and analysis. *International Journal of Speech Technology*, 25 (1) : 105-134 [DOI: 10.1007/s10772-021-09876-2]
- Mohammadi S H. 2015. Reducing one-to-many problem in voice conversion by equalizing the formant locations using dynamic frequency warping [EB/OL]. [2023-08-28].
<http://arxiv.org/pdf/1510.04205.pdf>
- Morise M, Yokomori F and Ozawa K. 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, 99 (7) : 1877-1884 [DOI: 10.1587/transinf.2015EDP7457]
- Müller N, Dieckmann F, Czempin P, Canals R, Böttinger K and Williams J. 2021. Speech is silver, silence is golden: what do asvspoof-trained models really learn?//*Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*. [s.l.] : ISCA: 55-60 [DOI: 10.21437/ASVSPOOF. 2021-9]
- Müller N, Czempin P, Dieckmann F, Froghyar A and Böttinger K. 2022. Does audio deepfake detection generalize?//*Proceedings of the 23rd Interspeech Annual Conference of the International Speech Communication Association*. Incheon, Korea (South) : ISCA: 2783-2787 [DOI: 10.21437/Interspeech.2022-108]
- Nguyen B and Cardinaux F. 2022. NVC-Net: end-to-end adversarial voice conversion//*Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Singapore, Singapore: IEEE: 7012-7016 [DOI: 10.1109/ICASSP43922.2022.9747020]
- OpenAI. 2023. GPT-4 technical report [EB/OL]. [2023-08-28].
<http://arxiv.org/pdf/2303.08774.pdf>
- Park D S, Chan W, Zhang Y, Chiu C C, Zoph B, Cubuk E D and Le Q V. 2019. SpecAugment: a simple data augmentation method for automatic speech recognition//*Proceedings of the 20th Annual Conference of the International Speech Communication Association*. Graz, Austria: ISCA: 2613-2617 [DOI: 10.21437/Interspeech. 2019-2680]
- Park S W, Kim D Y and Joe M C. 2020. Cotatron: transcription-guided speech encoder for any-to-many voice conversion without parallel data//*Proceedings of the 21st Interspeech Annual Conference of the International Speech Communication Association*. Shanghai, China: ISCA: 4696-4700 [DOI: 10.21437/Interspeech.2020-1542]
- Peng K N, Ping W, Song Z and Zhao K X. 2020. Non-autoregressive neural text-to-speech//*Proceedings of the 37th International Conference on Machine Learning, ICML 2020*. [s.l.] : PMLR: 7586-7598
- Ping W, Peng K N, Gibiansky A, Arik S Ö, Kannan A, Narang S, Raiman J and Miller J. 2017. Deep Voice 3: scaling text-to-speech with convolutional sequence learning//*Proceedings of the 6th International Conference on Learning Representations*. Vancouver, Canada: ICLR
- Prenger R, Valle R and Catanzaro B. 2019. Waveglow: a flow-based generative network for speech synthesis//*Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, UK: IEEE: 3617-3621 [DOI: 10.1109/ICASSP.2019.8683143]
- Qian K Z, Zhang Y, Chang S Y, Yang X S and Hasegawa-Johnson M. 2019. AutoVC: zero-shot voice style transfer with only autoencoder loss//*Proceedings of the 36th International Conference on Machine Learning*. Long Beach, USA: PMLR: 5210-5219
- Qian Y, Fan Y C, Hu W P and Soong F K. 2014. On the training aspects of deep neural network (DNN) for parametric TTS synthesis//*Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, Italy: IEEE: 3829-3833 [DOI: 10.1109/ICASSP.2014.6854318]
- Ranjan R, Vatsa M and Singh R. 2022. STATNet: spectral and temporal features based multi-task network for audio spoofing detection//*Proceedings of 2022 IEEE International Joint Conference on Biometrics (IJCB)*. Abu Dhabi, United Arab Emirates: IEEE: 1-9 [DOI: 10.1109/IJCB54206.2022.10007949]
- Ranjan R, Vatsa M and Singh R. 2023. Uncovering the deceptions: an analysis on audio spoofing detection and future prospects//*Proceedings of the 32nd International Joint Conference on Artificial Intelligence, IJCAI 2023*. Macao, China: IJCAI: 6750-6758 [DOI: 10.24963/2JCAI.2023/756]

- Reimao R and Tzerpos V. 2019. FoR: a dataset for synthetic speech detection//Proceedings of 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). Timisoara, Romania: IEEE: 1-10 [DOI: 10.1109/SPED.2019.8906599]
- Ren Y, Hu C X, Tan X, Qin T, Zhao S, Zhao Z and Liu T Y. 2022. FastSpeech 2: fast and high-quality end-to-end text to speech//Proceedings of the 9th International Conference on Learning Representations. Virtual Event: ICLR
- Ren Y, Ruan Y J, Tan X, Qin T, Zhao S, Zhao Z and Liu T Y. 2019. FastSpeech: fast, robust and controllable text to speech//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 3171-3180 [DOI: 10.5555/3454287.3454572]
- Ren Y Z, Liu C Y, Liu W Y and Wang L N. 2021. A survey on speech forgery and detection. *Journal of Signal Processing*, 37(12): 2412-2439 (任延珍, 刘晨雨, 刘武洋, 王丽娜. 2021. 语音伪造及检测技术研究综述. *信号处理*, 37(12): 2412-2439) [DOI: 10.16798/j.issn.1003-0530.2021.12.011]
- Rostami A M, Homayounpour M M and Nickabadi A. 2021. Efficient attention branch network with combined loss function for automatic speaker verification spoof detection. *Circuits, Systems, and Signal Processing*, 42 (7) : 4252-4270 [DOI: 10.1007/s00034-023-02314-5]
- Sahidullah M, Delgado H, Todisco M, Kinnunen T, Evans N, Yamagishi J and Lee K A. 2019. Introduction to voice presentation attack detection and recent advances//Marcel S, Nixon M S, Fierrez J and Evans N, eds. *Handbook of Biometric Anti-Spoofing*. Cham, Germany: Springer: 321-361 [DOI: 10.1007/978-3-319-92627-8_15]
- Saito D, Yamamoto K, Minematsu N and Hirose K. 2011. One-to-many voice conversion based on tensor representation of speaker space//Proceedings of the 12th Annual Conference of the International Speech Communication Association. Florence, Italy: ISCA: 653-656 [DOI: 10.21437/Interspeech.2011-268]
- Serrà J, Pascual S and Segura C. 2019. Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion//Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver, Canada: NIPS: 6790-6800
- Shen J, Pang R M, Weiss R J, Schuster M, Jaitly N, Yang Z H, Chen Z F, Zhang Y, Wang Y X, Skerrv-Ryan R, Saurous R A, Agiomyriannakis Y and Wu Y H. 2018. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE: 4779-4783 [DOI: 10.1109/ICASSP.2018.8461368]
- Shim H J, Heo H S, Jung J W and Yu H J. 2019. Self-supervised pre-training with acoustic configurations for replay spoofing detection//Proceedings of the 21st Interspeech Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 1091-1095 [DOI: 10.21437/Interspeech.2020-1345]
- Song E, Yamamoto R, Hwang M J, Kim J S, Kwon O and Kim J M. 2021. Improved parallel WaveGAN vocoder with perceptually weighted spectrogram loss//Proceedings of 2021 IEEE Spoken Language Technology Workshop (SLT). Shenzhen, China: IEEE: 470-476 [DOI: 10.1109/SLT48900.2021.9383549]
- Stylianou Y. 2001. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(1): 21-29 [DOI: 10.1109/89.890068]
- Su Z P, Li M K, Zhang G F, Wu Q F, Li M Q, Zhang W M and Yao X. 2023. Robust audio copy-move forgery detection using constant Q spectral sketches and GA-SVM. *IEEE Transactions on Dependable and Secure Computing*, 20(5): 4016-4031 [DOI: 10.1109/TDSC.2022.3215280]
- Sun C Z, Jia S, Hou S W and Lü S W. 2023. AI-synthesized voice detection using neural vocoder artifacts//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, Canada: IEEE: 904-912 [DOI: 10.1109/CVPRW59228.2023.00097]
- Tak H, Jung J W, Patino J, Kamble M, Todisco M and Evans N. 2021a. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [s. l.] : ISCA: 1-8 [DOI: 10.21437/ASVSPOOF.2021-1]
- Tak H, Kamble M, Patino J, Todisco M and Evans N. 2022a. Rawboost: a raw data boosting and augmentation method applied to automatic speaker verification anti-spoofing//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 6382-6386 [DOI: 10.1109/ICASSP43922.2022.9746213]
- Tak H, Patino J, Todisco M, Nautsch A, Evans N and Larcher A. 2021b. End-to-end anti-spoofing with RawNet2//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 6369-6373 [DOI: 10.1109/ICASSP39728.2021.9414234]
- Tak H, Todisco M, Wang X, Jung J W, Yamagishi J and Evans N. 2022b. Automatic speaker verification spoofing and deepfake detection using Wav2vec 2.0 and data augmentation//Odyssey 2022, The Speaker and Language Recognition Workshop (Odyssey 2022). Beijing, China: ISCA: 112-119 [DOI: 10.21437/Odyssey.2022-16]
- Tan C B, Hijazi M H A, Khamis N, Nohuddin P N E B, Zainol Z, Coenen F and Gani A. 2021. A survey on presentation attack detection for automatic speaker verification systems: state-of-the-art, taxonomy, issues and future direction. *Multimedia Tools and Applications*, 80 (21) : 32725-32762 [DOI: 10.1007/s11042-021-11235-x]
- Tang H Z, Zhang X L, Wang J Z, Cheng N, Zeng Z, Xiao E and Xiao J. 2021. TGAVC: improving autoencoder voice conversion with

- text-guided and adversarial training//Proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Cartagena, Colombia: IEEE: 938-945 [DOI: 10.1109/ASRU51503.2021.9688088]
- Tao J H, Fu R B, Yi J Y, Wang C L and Wang T. 2020. Development and challenge of speech forgery and detection. *Journal of Cyber Security*, 5(2): 28-38 (陶建华, 傅睿博, 易江燕, 王成龙, 汪涛, 2020. 语音伪造与鉴伪的发展与挑战. *信息安全学报*, 5(2): 28-38) [DOI: 10.19363/J.cnki.cn10-1380/tn.2020.02.03]
- Teng Z W, Fu Q C, White J, Powell M E and Schmidt D C. 2022. SASV: an end-to-end spoof-aggregated spoofing-aware speaker verification system//Proceedings of Interspeech 2022, the 23rd Annual Conference of the International Speech Communication Association. Incheon, Korea (South): ISCA: 4391-4395 [DOI:10.21437/Interspeech.2022-11029]
- Todisco M, Wang X, Vestman V, Sahidullah M, Delgado H, Nautsch A, Yamagishi J, Evans N, Kinnunen T H and Lee K A. 2019. ASVspoof 2019: future horizons in spoofed and fake audio detection//Proceedings of the 20th Interspeech Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA: 1008-1012 [DOI: 10.21437/Interspeech.2019-2249]
- Tokuda K, Nankaku Y, Toda T, Zen H G, Yamagishi J and Oura K. 2013. Speech synthesis based on hidden Markov models. *Proceedings of the IEEE*, 101(5): 1234-1252 [DOI: 10.1109/JPROC.2013.2251852]
- Tokuda K, Yoshimura T, Masuko T, Kobayashi T and Kitamura T. 2000. Speech parameter generation algorithms for HMM-based speech synthesis//Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Istanbul, Turkey: IEEE: 1315-1318 [DOI: 10.1109/ICASSP.2000.861820]
- Tomilov A, Svishchev A, Volkova M, Chirkovskiy A, Kondratev A and Lavrentyeva G. 2021. STC antispoofing systems for the ASVspoof2021 challenge//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [s.l.]: ISCA: 61-67 [DOI: 10.21437/ASVSPPOOF.2021-10]
- van den Oord A, Dieleman S, Zen H G, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A W and Kavukcuoglu K. 2016. WaveNet: a generative model for raw audio//The 9th ISCA Speech Synthesis Workshop. Sunnyvale, USA: ISCA: #125
- van den Oord A, Li Y Z, Babuschkin I, Simonyan K, Vinyals O, Kavukcuoglu K, van den Driessche G, Lockhart E, Cobo L C, Stimberg F, Casagrande N, Grewe D, Noury S, Dieleman S, Elsen E, Kalchbrenner N, Zen H G, Graves A, King H L, Walters T, Belov D and Hassabis D. 2018. Parallel WaveNet: fast high-fidelity speech synthesis//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 3918-3926
- Wang C L, Yi J Y, Tao J H, Sun H Y, Chen X, Tian Z K, Ma H X, Fan C H and Fu R B. 2022a. Fully automated end-to-end fake audio detection//Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia. Lisboa, Portugal: Association for Computing Machinery: 27-33 [DOI: 10.1145/3552466.3556530]
- Wang C L, Yi J Y, Tao J H, Zhang C Y, Zhang S and Chen X. 2023a. Detection of cross-dataset fake audio based on prosodic and pronunciation features [EB/OL]. [2023-06-30]. <http://arxiv.org/pdf/2305.13700.pdf>
- Wang C L, Yi J Y, Tao J H, Zhang C Y, Zhang S, Fu R B and Chen X. 2023b. TO-Rawnet: improving Rawnet with TCN and orthogonal regularization for fake audio detection [EB/OL]. [2023-06-30]. <http://arxiv.org/pdf/2305.13701.pdf>
- Wang L, Yeoh B and Ng J W. 2022b. Synthetic voice detection and audio splicing detection using SE-Res2Net-Conformer architecture//The 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). Singapore, Singapore: IEEE: 115-119 [DOI: 10.1109/ISCSLP57327.2022.10037999]
- Wang Q Q, Zhang X L, Wang J Z, Cheng N and Xiao J. 2022c. DRVC: a framework of any-to-any voice conversion with self-supervised learning//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 3184-3188 [DOI: 10.1109/ICASSP43922.2022.9747434]
- Wang X M, Qin X Y, Zhu T L, Wang C, Zhang S L and Li M. 2021. The DKU-CMRI system for the ASVspoof 2021 challenge: vocoder based replay channel response estimation//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [s.l.]: ISCA: 16-21 [DOI: 10.21437/ASVSPPOOF.2021-3]
- Wang Y X, Skerry-Ryan R J, Stanton D, Wu Y H, Weiss R J, Jaitly N, Yang Z H, Xiao Y, Chen Z F, Bengio S, Le Q V, Agiomyrgiannakis Y, Clark R and Saurous R A. 2017. Tacotron: towards end-to-end speech synthesis//Proceedings of the 18th Interspeech Annual Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA: 4006-4010 [DOI: 10.21437/Interspeech.2017-1452]
- Wang Z Y and Hansen J H L. 2022. Audio anti-spoofing using simple attention module and joint optimization based on additive angular margin loss and meta-learning//Proceedings of the 23rd Interspeech Annual Conference of the International Speech Communication Association. Incheon, Korea (South): ISCA: 376-380 [DOI: 10.21437/Interspeech.2022-904]
- Weiss R J, Skerry-Ryan R J, Battenberg E, Mariooryad S and Kingma D P. 2021. Wave-Tacotron: spectrogram-free end-to-end text-to-speech synthesis//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 5679-5683 [DOI: 10.1109/ICASSP39728.2021.9413851]

- Wu H B, Liu A T and Lee H Y. 2020a. Defense for black-box attacks on anti-spoofing models by self-supervised learning//Proceedings of the 21st Interspeech Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 3780-3784 [DOI:10.21437/Interspeech.2020-2026]
- Wu H B, Liu S X, Meng H L and Lee H Y. 2020b. Defense against adversarial attacks on spoofing countermeasures of ASV//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 6564-6568 [DOI: 10.1109/ICASSP40776.2020.9053643]
- Wu Z Z, Das R K, Yang J C and Li H Z. 2020c. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks//Proceedings of the 21st Interspeech Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 1101-1105 [DOI:10.21437/Interspeech.2020-1810]
- Wu Z Z, Kinnunen T, Evans N, Yamagishi J, Hanilci C, Sahidullah M and Sizov A. 2015. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge//Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany: ISCA: 2037-2041 [DOI: 10.21437/Interspeech.2015-462]
- Xu X X, Shi L, Chen X Q, Lin P Y, Lian J, Chen J H, Zhang Z H and Hancock E R. 2023. Any-to-any voice conversion with multi-layer speaker adaptation and content supervision. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31: 3431-3445 [DOI: 10.1109/TASLP.2023.3306716]
- Xue J, Fan C H, Yi J Y, Wang C L, Wen Z Q, Zhang D and Lü Z. 2023. Learning from yourself: a self-distillation method for fake speech detection//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece: IEEE: 1-5 [DOI: 10.1109/ICASSP49357.2023.10096837]
- Yadav A K S, Bhagtani K, Xiang Z Y, Bestagini P, Tubaro S and Delp E J. 2023. DSVAE: interpretable disentangled representation for synthetic speech detection [EB/OL]. [2023-06-30]. <http://arxiv.org/pdf/2304.03323.pdf>
- Yamagishi J, Todisco M, Sahidullah M, Delgado H, Wang X, Evans N, Kinnunen T, Lee K A, Vestman V and Nautsch A. 2019. ASVspoof 2019: automatic speaker verification spoofing and countermeasures challenge evaluation plan [EB/OL]. [2023-10-20]. https://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf
- Yamamoto R, Song E, Hwang M J and Kim J M. 2021. Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE: 6039-6043 [DOI: 10.1109/ICASSP39728.2021.9413369]
- Yamamoto R, Song E and Kim J M. 2020. Parallel WaveGAN: a fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE: 6199-6203 [DOI: 10.1109/ICASSP40776.2020.9053795]
- Yan X R, Yi J Y, Tao J H, Wang C L, Zhang C Y and Fu R B. 2023. System fingerprint recognition for deepfake audio: an initial dataset and investigation [EB/OL]. [2023-08-28]. <http://arxiv.org/pdf/2208.10489.pdf>
- Yang J, Lee J, Kim Y I, Cho H Y and Kim I. 2020. VocGAN: a high-fidelity real-time vocoder with a hierarchically-nested adversarial network//Proceedings of the 21st Interspeech Annual Conference of the International Speech Communication Association. Shanghai, China: ISCA: 200-204 [DOI:10.21437/Interspeech.2020-1238]
- Yang J C and Das R K. 2020. Long-term high frequency features for synthetic speech detection. Digital Signal Processing, 97: #102622 [DOI: 10.1016/j.dsp.2019.102622]
- Yang S, Qiao K, Chen J, Wang L Y and Yan B. 2022. Overview on speech synthesis, forgery and detection technology. Computer Systems and Applications, 31(7): 12-22 (杨帅, 乔凯, 陈健, 王林元, 闫斌. 2022. 语音合成及伪造、鉴伪技术综述. 计算机系统应用, 31(7): 12-22) [DOI: 10.15888/j.cnki.csa.008641]
- Yi J Y, Bai Y, Tao J H, Ma H X, Tian Z K, Wang C L, Wang T and Fu R B. 2021. Half-Truth: a partially fake audio detection dataset//Proceedings of the 22nd Interspeech Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 1654-1658 [DOI:10.21437/Interspeech.2021-930]
- Yi J Y, Fu R B, Tao J H, Nie S, Ma H X, Wang C L, Wang T, Tian Z K, Bai Y, Fan C H, Liang S, Wang S M, Zhang S, Yan X R, Xu L, Wen Z Q and Li H Z. 2022a. ADD 2022: the first audio deep synthesis detection challenge//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 9216-9220 [DOI: 10.1109/ICASSP43922.2022.9746939]
- Yi J Y, Tao J H, Fu R B, Yan X R, Wang C L, Wang T, Zhang C Y, Zhang X H, Zhao Y, Ren Y, Xu L, Zhou J Z, Gu H, Wen Z Q, Liang S, Lian Z, Nie S and Li H Z. 2023. ADD 2023: the second audio deepfake detection challenge [EB/OL]. [2023-08-28]. <http://arxiv.org/pdf/2305.13774.pdf>
- Yi J Y, Wang C L, Tao J H, Tian Z K, Fan C H, Ma H X and Fu R B. 2022b. SceneFake: an initial dataset and benchmarks for scene fake audio detection [EB/OL]. [2023-06-30]. <https://arxiv.org/pdf/2211.06073v1.pdf>
- Yoon S and Yu H J. 2021. Multiple-point input and time-inverted speech signal for the ASVspoof 2021 Challenge//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [s. l.]: ISCA: 37-41 [DOI: 10.21437/ASVSPPOOF.2021-6]

- Zeinali H, Stafylakis T, Athanasopoulou G, Rohdin J, Gkinis I, Burget L and Cernocký J H. 2019. Detecting spoofing attacks using VGG and SincNet: BUT-omilia submission to ASVspoof 2019 challenge//Proceedings of Interspeech the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA: 1073-1077 [DOI:10.21437/Interspeech.2019-2892]
- Zen H G, Senior A and Schuster M. 2013. Statistical parametric speech synthesis using deep neural networks//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE: 7962-7966 [DOI: 10.1109/ICASSP.2013.6639215]
- Zen H G, Tokuda K and Black A W. 2009. Statistical parametric speech synthesis. *Speech Communication*, 51(11): 1039-1064 [DOI: 10.1016/j.specom.2009.04.004]
- Zhang D, Li S M, Zhang X, Zhan J, Wang P Y, Zhou Y Q and Qiu X P. 2023a. SpeechGPT: empowering large language models with intrinsic cross-modal conversational abilities//Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore, Singapore: Association for Computational Linguistics: 15757-15773 [DOI: 10.18653/v1/2023.findings-emnlp.1055]
- Zhang H, Yuan T, Chen J K, Li X T, Zheng R J, Huang Y X, Chen X J, Gong E L, Chen Z Y, Hu X G, Yu D H, Ma Y J and Huang L. 2022a. PaddleSpeech: an easy-to-use all-in-one speech toolkit//Proceedings of 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations. Washington, USA: Association for Computational Linguistics: 114-123 [DOI: 10.18653/v1/2022.naacl-demo.12]
- Zhang L, Wang X, Cooper E, Evans N and Yamagishi J. 2023b. The PartialSpoof database and countermeasures for the detection of short fake speech segments embedded in an utterance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 813-825 [DOI: 10.1109/TASLP.2022.3233236]
- Zhang L, Wang X, Cooper E and Yamagishi J. 2021a. Multi-task learning in utterance-level and segmental-level spoof detection//2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. [s.l.]: NII Yamagishi Laboratory
- Zhang X W, Li J K, Sun M and Zheng L L. 2020. Speech anti-spoofing: the state of the art and prospects. *Journal of Data Acquisition and Processing*, 35(5): 807-823 (张雄伟, 李嘉康, 孙蒙, 郑琳琳, 2020. 语音欺骗检测方法的研究现状及展望. 数据采集与处理, 35(5): 807-823) [DOI: 10.16337/j.1004-9037.2020.05.002]
- Zhang Y, Jiang F and Duan Z Y. 2021b. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 28: 937-941 [DOI: 10.1109/LSP.2021.3076358]
- Zhang Y, Jiang F, Zhu G, Chen X H and Duan Z Y. 2023c. Generalizing voice presentation attack detection to unseen synthetic attacks and channel variation//Marcel S, Fierrez J and Evans N, eds. *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*. Singapore, Singapore: Springer Nature: 421-443 [DOI: 10.1007/978-981-19-5288-3_15]
- Zhang Y, Zhu G and Duan Z Y. 2022b. A probabilistic fusion framework for spoofing aware speaker verification//Odyssey 2022: The Speaker and Language Recognition Workshop. Beijing, China: ISCA: 77-84 [DOI: 10.21437/Odyssey.2022-11]
- Zhang Y, Zhu G, Jiang F and Duan Z Y. 2021c. An empirical study on channel effects for synthetic voice spoofing countermeasure systems//Proceedings of the 22nd Interspeech Annual Conference of the International Speech Communication Association. Brno, Czechia: ISCA: 4309-4313 [DOI: 10.21437/Interspeech. 2021-1820]
- Zhang Y J, Pan S F, He L and Ling Z H. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis//Proceedings of 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE: 6945-6949 [DOI: 10.1109/ICASSP.2019.8683623]
- Zhang Z Y, Gu Y W, Yi X W and Zhao X F. 2021d. FMFCC-A: a challenging mandarin dataset for synthetic speech detection//Digital Forensics and Watermarking-20th International Workshop, IWDW 2021. Beijing, China: Springer-Verlag: 117-131 [DOI: 10.1007/978-3-030-95398-0_9]
- Zhao Y, Yi J Y, Tao J H, Wang C L, Zhang C Y, Wang T and Dong Y F. 2022. EmoFake: an initial dataset for emotion fake audio detection [EB/OL]. [2023-06-30]. <http://arxiv.org/pdf/2211.05363.pdf>

作者简介

许裕雄,男,博士研究生,主要研究方向为伪造语音检测。

E-mail:xuyuxiong2022@email.szu.edu.cn

李斌,通信作者,男,教授,博士生导师,主要研究方向为多媒体信息安全和智能信息处理。E-mail:libin@szu.edu.cn

谭舜泉,男,教授,硕士生导师,主要研究方向为多媒体信息安全、隐写及隐写分析和深度学习。E-mail:tansq@szu.edu.cn

黄继武,男,特聘教授,博士生导师,主要研究方向为多媒体信息安全和信息隐藏。E-mail:jwhuang@szu.edu.cn