语音伪造与鉴伪的发展与挑战

陶建华[1,2,3], 傅睿博[1,2], 易江燕[1], 王成龙[1], 汪 涛[1,2]

[1]中国科学院自动化研究所模式识别国家重点实验室 北京 中国 100190
[2]中国科学院大学人工智能技术学院 北京 中国 100190
[3]中国科学院自动化研究所中国科学院脑科学与智能技术研究中心 北京 中国 100190

摘要 本文对语音伪造与鉴伪的发展进行了梳理与阐释。针对语音伪造的适用场景与关键技术点, 分别对身份风格伪造、音色与韵律伪造、语音模拟三大核心语音伪造技术的基本概念、发展历程、优势与不足进行梳理与分析。针对语音伪造的应对技术语音鉴伪技术, 首先介绍整理了针对性较强、面向参数式语音伪造、拼接式语音伪造与语音模拟技术框架的应对技术, 在此基础上介绍了具有普适性更强的基于深度鉴别网络语音鉴伪研究进展。在此基础上, 本文针对语音伪造技术所面临口语化、低资源的挑战, 对未来多风格、低成本、鲁棒性发展趋势进行分析。对于语音鉴伪, 本文从语料库、特征挖掘、异常检测三个角度对未来的研究重点进行诠释。

关键词 语音伪造; 语音鉴伪; 发展与挑战
中图法分类号 TP191 DOI 号 10.19363/J.cnki.cn10-1380/tn.2020.02.03

Development and Challenge of Speech Forgery and Detection

TAO Jianhua[1,2,3], FU Ruibo[1,2], YI Jiangyan[1], WANG Chenglong[1], Wang Tao[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Abstract This paper reviews and explains the development ofspeech forgery and forgery detection. According to the applicable scenarios and key technology points of deep speech forgery, the basic concepts, development process, advantages and disadvantages of three core speech forgery technologies, namely identity style forgery, timbre and rhythm forgery, and speech simulation, are analyzed. In view of the response technology of speech forgery, this paper first introduces the countermeasures of speech forgery, which is source limited, and then introduces the research progress of speech forgery based on depth identification network with better generalization. To sum up, this paper analyzes the development trend of multi-style, low cost and robustness in the future, aiming at the challenge of colloquialism and low resource faced by speech forgery technology. As for speech detection, this paper interprets the future research focus from corpus, feature mining and anomaly detection.

Key words speech forgery; speech forgery detection; development and challenge

1 引言

语音伪造与鉴伪作为一组“攻”与“防”的偶合技术, 引起了学术界与工业界的广泛关注。20 世纪早在 60 年代初, 美国国防高级研究计划局(Defense Advanced Research Projects Agency, DARPA)开始介入自主技术的研究, 并意识到人工智能可以满足大量的国家安全需要。在 20 世纪 70 年代初, DARPA 启动了语音识别研究(Speech Understanding Research, SUR)项目, 相继开发了 Sphinx、BYBLOS、DECIPHER 等一系列语音识别系统, 可进行整句连续的语音识别。2000 年之后, DARPA 开始研制通过对话进行人机交互(Human–Computer Interaction, HCI)的系统, 该系统能从与不同人的对话中学习经

通讯作者: 陶建华, 博士, 研究员, Email:jhtao@nlpr.ia.ac.cn。
本课题得到国家重点研发计划(No.2018YFB1005003), 国家自然科学基金(No.61831022, No.61771472, No.61773379, No.61901473), cas-inria 院双边合作项目资助(No.173211KYSB20190049)。
收稿日期: 2019-12-31; 修改日期: 2020-03-10; 定稿日期: 2020-03-11

验, 提供个性化的服务。2005 年, DARPA 启动了全球自动化语言情报利用项目, 寻求能够对标准阿拉伯语和汉语的印刷品、网页、新闻及电视广播进行实时翻译的技术, 目标是使得 95%的文本文档翻译和90%的语音文件翻译均能达到95%的正确率。2008年11月, DARPA 启动了 Machine Reading 项目, 旨在实现人工智能的应用和发展学习系统的过程中对自然文本进行知识插入。2010 年 3 月, DARPA 启动了心灵之眼(Mind's Eye)项目, 目的是为机器建立视觉的智能, 对视频信息进行形象推理。2012 年启动的文本深度发掘和过滤(Deep Exploration and Filtering of Text, DEFT)项目明确地提出要利用深度学习技术发掘大量结构化文本中隐含的、有实际价值的特征信息, 同时还要具备可将处理后的信息进行进一步整合的能力, 可将这些技术应用于作战评估、规划、预测的辅助决策支持中。2014 年, DARPA 的大机制(Big Mechanism)项目, 开发了协助计算机阅读科学和技术文章的技术, 将知识片段综合成更完整的模型, 并提出实现特定目标的干预措施。2016 年 10 月, DARPA 发布可解释的人工智能(Explainable Artificial Intelligence, XAI)项目的广泛机构公告, 其目标是建立一套新的或改进的机器学习技术, 生成可解释的模型, 结合有效的解释技术, 使得最终用户能够理解、一定程度的信任并有效地管理未来的人工智能系统。

语音伪造与鉴伪引起了业界广泛关注, 相关学术研究仍然存在不够深入, 缺乏可解释性问题。为此, 业界对相关产业的研发投入进一步加大。2018 年 3 月 20 日, DARPA 启动一项名为"媒体取证"(简称 MediFor)的项目, 该项目研发了自动评估图像或视频完整性的技术, 并将这些技术集成到端到端, MediFor 平台将自动检测并分析媒体上的伪造, 并推断视觉媒体的完整性, 识别是否经过编辑、带有操控目的的影像。2018 年 9 月 8 日, DARPA 宣布计划投入 20 亿美元开发新的人工智能(AI)技术, 这是该机构"AI Next(下一代人工智能)"计划的一部分。该项目正在开发新一代机器学习技术, 以形成一套基础理论来解释人工智能得出的结论。2019 年 Facebook 计划投入约 1000 万美元的资源来发起 Deepfake 检测挑战赛(Deepfake Detection Challeng)。根据 2020 财年 DARPA 人工智能重点投资项目分析, "可解释的人工智能"项目、"不同来源的主动诠释"项目、"自动知识提取"项目和"确保可 AI 抗欺骗可靠性"项目等, 都涉及到了态势认知的研究, 将数据结合知识、环境等信息转化为认知, 致力于打造具有常识、能感知语境和更高效的系统, 这也是未来认知技术的发展方向。

针对目前日益增长的产业需求, 本文从目前语音伪造与鉴伪的研究进展为出发点, 针对主流技术框架, 对发展历程进行梳理, 总结各类技术的优势与不足, 对研究中的难点进行分析, 对学术界的研究趋势进行预测。本文组织结构如下, 第二章从三大技术框架对语音伪造生成研究进展进行分析。第三章从专用语音鉴伪与普适语音鉴伪两个角度介绍了语音鉴伪的研究现状。第四章是语音伪造与鉴伪的趋势与挑战分析。第五章是总结。

2 语音伪造生成研究进展

语音伪造的功能可分为欺骗人类听者与欺骗机器两个角度。从欺骗机器角度, 语音伪造重在对机器鉴别语音的区分性特征进行针对性精准生成。从欺骗人类听者角度, 语音伪造任务首先要保证伪造音色与目标说话人相一致, 及相似度较高。在此基础上, 一个自然的韵律效果也是语音伪造任务的一大目标, 人类可以轻易地感知韵律过于平均的伪造。此外, 从语音模拟角度直接在语音波形层面直接对语音进行转化与修改也是实现语音伪造的重要渠道。因此本部分将对语音身份风格伪造、语音音色与韵律伪造、语音模拟这三个研究进展分别叙述。

2.1 语音身份风格伪造研究进展

语音身份风格伪造是为了生成目标说话人的说话风格信息。语音身份风格伪造技术起源于 20 世纪末, 其主要的目的是根据给定说话人的语音, 抽取出该说话人的说话风格特征。从直觉上来说, 语音身份风格不像人脸、指纹的个体差异那样明显, 由于每个人先天的发声器官(如舌头、牙齿、口腔、声带、肺、鼻腔)等在尺寸和形态方面存在差异, 再加之年龄、性格、语习惯等各种后天因素的影响, 可以说每个说话人的声纹是独一无二的, 并可以在相对长的时间里保持相对稳定不变。在 20 世纪 40 年代, Bell 实验室的 L.G.Kersta 等人借助肉眼观察语谱图发现不同人的发音在语谱图中具有差异性, 提出通过观察语谱图判断不同说话人的风格。根据语谱图上的共振峰纹路, 首次提出了"声纹"的概念。1966 年, 随着计算机技术的不断进步, 说话人风格抽取逐步由单纯的人耳听讲, 转向基于计算机的自动提取。早期的语音身份风格抽取主要采用有效的声学特征参数和模式匹配的方法, 匹配往往通过特征矢量之间的距离测度来实现, 累计距离为匹配结果。即一个完整的说话人风格抽取系统分为两个阶段: 说话人训练

和说话人风格抽取。在说话人训练阶段, 系统首先对训练语音进行静音剔除和降噪处理, 尽可能得到纯净有效的语音片段, 然后再提取语音对应的声学特征参数, 根据系统建模算法, 得到说话人的特征模型, 每个说话人的训练语音经过训练阶段后得到一个说话人模型。

到 20 世纪 70 年代至 80 年代, 动态时间规整(DTW)、矢量量化(VQ)和隐马尔可夫模型技术(HMM)的出现对当时说话人风格抽取有了较大提升。到 90 年代, 高斯混合模型(GMM)以及高斯混合模型-通用背景模型(GMM-HMM)以其简单灵活、鲁棒性强的特点, 将说话人风格抽取研究带入一个新的阶段。进入 21 世纪后, 在传统 GMM-HMM 的方法上, P.Kenny、N. Dehak 等人先后提出了联合因子分析技术(Jiont Factor Analysis, JFA)和扰动属性干扰算法, 使得声纹识别在复杂背景条件下也能取得较好的效果。由 JFA 建模思想得到启示, 提出基于总体变化因子向量(identity vector, i-vector)的说话人建模方法, 这也是该研究领域的经典技术之一。后来研究人员为了解决信道失配问题, 在 i-vector 基础上引入有类内协方差归一化(Within-class Covariance Normalization, WCCN)、概率线性鉴别分析(PLDA)等区分技术。2012 年以来, 基于深度网络的特征学习方法, 利用复杂非线性结构赋予的特征提取能力, 能自动对输入的语音信号进行特征分析, 提取出更高层、更抽象的说话人声纹表征, 如 d-vector、x-vector。d-vector 是纯 DNN 框架下的说话人风格抽取系统, 通过训练说话人标签的 DNN 模型, 提取测试说话人语音的瓶颈特征, 对瓶颈特征进行累加求均值, 得到语音的 d-vector。x-vector 是在 TDNN 网络中提取的嵌入特征, 由于语音经过 TDNN 时延网络, 可以从输出层得到关于输入语音帧的长时特征, 因此 x-vector 在短时说话人特征抽取中能够达到更高的准确率。2016 年, Google 的 Heigold 等人提出了端到端声纹识别系统, 端到端的网络包含两部分: 预先训练好的特征提取网络和用于决策打分的判决网络, 输入为不同说话人的语音信号, 输出即为说话人识别结果, 之后如注意力机制、自适应方法等在端到端系统中的应用进一步提高了系统的性能。

将语音的说话风格与语音模拟框架进行融合即可模拟特定说话人的声音。在传统的多模型级联式语音模拟框架下, 为了实现个性化小数据语音模拟的任务, 在声学模型层面进行自适应, 其中一个角度是将说话人风格特征融合在模型的不同位置上, 具体可以分为三种情况: 输入层, 中间层以及输出层三种位置。通过结合说话人风格特征, 可以很好的学习出说话人的说话风格[1]。此外, 另一个角度是共享声学模型的中间层, 分别在模型的输入端和输出端将不同语音和说话人分解建模, 从而使模型具备产生不同说话风格的能力[2]。

2.2 语音音色伪造研究进展

语音音色伪造是为了伪造目标说话人的音色信息。在拥有较高质量目标伪造人语音的情况下, 波形拼接式语音伪造技术是一种主流技术(Hunt 等, 1996; Chou 等, 2002; Christophe 等, 2002), 其技术思想就是将在大量自然语流中的丰富的语音单元按照一定的规则拼接, 得到高自然度与相似度的语音[3]。可以想象, 由于最终伪造波形几乎原封不动地取自于原始目标伪造说话人, 因此对目标伪造获取的原始语音的获取与整理对伪造结果的自然度至关重要。针对目标伪造语音, 整理的伪造语料库应能采用尽量少的语料覆盖尽量多的语言现象, 包含丰富多变的韵律信息。语料库设计的合理性直接决定合成时是否能够选出所需要的模拟声学单元, 最后能否得到高自然度的语音。在语料库内对每个单元都将保留若干个在不同韵律环境下的变体, 在伪造过程中通过前端对文本的分析获得当前单元的韵律环境属性集, 然后再通过某种基元选取算法选出与当前环境最为匹配的变体, 把这些变体拼接起来得到合成语音。可见语料库中包含的单元变体越多, 越有可能精细地反映出韵律环境的细微变化, 合成结果的表现力就会越强, 越接近真人发声的结果。由上述分析可知, 在实现波形拼接式语音伪造过程中, 除了要设计并收集出一个包含足够多的可用的单元的目标说话人语料库以外, 还要解决模拟声学单元、基元选取、拼接伪造三个环节中存在的问题。波形拼接式语音伪造系统中可使用的模拟声学单元包括: 音素(声韵母)、半音节、音节、双音素、词或短语。在选择系统的声学单元时需要主要考虑四点目标。

模拟声学单元之间的拼接损失要尽可能小, 这一点是拼接式语音伪造的基本原理所要求的。直观的考虑是采用尽量长的单元, 比如词、短语甚至句子, 以这样的单元来合成文本, 需要的拼接点少, 从而也降低了平均拼接损失。但实际上, 拼接处的不连续是任何拼接系统都无法避免的, 少数几处明显的失真就会恶化听者对整个句子的印象, 于是在系统中可以允许轻微的拼接损失, 只要所选择的单元从本质上可以较为平滑的拼接即可。例如, 在听感上元音间谱的不连续要比擦音间的更加明显, 而音节之间的拼接则明显比音节内部声韵母的拼接更平滑。目

前常用的办法是每个单元在音库中都保留多个样本, 伪造时从中挑选出拼接损失最小的。单元应使得伪造系统具备较强的通用性, 语音伪造的最终目的是使得系统能够自然流畅的合成任意输入的文本, 这要求音库中的单元应能覆盖任何可能出现的语言现象。对汉语来说, 如果选择词或者短语作为基本单元, 则要求音库至少应包含所有单字词, 还要包含常用的二字词、多字词、成语甚至常用的短句以提高合成质量。这样固然可以通过无限度地扩大音库规模而提升音质, 但在诸多场合下是不现实的, 特别是当存储空间比较有限的时候。但在某些特定领域内语音伪造, 应用这样的单元可以得到较好的效果, 比如天气预报、报时、金融业务等。单元应能在较低的代价下实现韵律的修改。拼接式语音伪造直接选用收集到的伪造目标音库中的原始波形进行拼接, 实际选出的单元在韵律上可能与期望中的不相符合, 这就需要采用某种韵律修改的算法对基频、音长等韵律信息进行修正。在信号处理的过程中必然会导致音质的损失, 选择单元时应考虑这方面的问题, 使得候选单元与目标单元的韵律模式尽量相近, 比如选用有调音节而不是无调音节。

单元应能满足统计训练的要求。在收集到的伪造语料库规模一定的情况下, 采用的单元越小, 每个单元的样本就会越多, 越有利于采用统计机器学习的方法进行各种模型的训练, 训练的效果也会越好。

语音伪造系统输入的文本通过前端的文本分析与韵律预测模块后, 转化成一串目标单元序列, 每个目标单元都携带着与其对应的特征信息。接着从音库中为每个目标单元挑选出与之相近的样本作为备选, 同时还应保证选出的样本之间过渡得尽量平滑。基于上述目的, Alan W Black 提出的基于目标代价和拼接代价的基元选取算法可以很好的选取出合适的基元。

然而拼接式语音合成需要大量目标伪造人语料, 且容易出现拼接点不连续的问题, 该特点极易被人类或机器所识别, 且拼接式语音伪造极容易出现韵律不自然的问题, 极大地降低了伪造效果。目的是根据给定伪造文本, 生成目标说话人的伪造语音。参数生成式语音伪造[4]在生成过程中首先会对伪造语音的声学参数进行伪造建模, 再将声学参数转换成语音波形。按照功能模块划分, 参数生成式语音伪造系统主要分为两大部分: 文本分析和波形合成。一般地, 文本分析又称为语音伪造系统的前端部分, 波形合成又称为后端部分。波形合成是语音伪造系统的后端, 在文本分析前端给出要伪造语音正确的发音及韵律信息后, 目标伪造语音的声学参数首先通过声学模型被预测, 最终通过声码器技术生成最终伪造语音。

20 世纪末, 语音信号统计建模算法的研究已经趋于成熟[5], 以基于隐马尔可夫(Hidden Markov Model, HMM) 的语音合成最为成功, 其相应的合成系统为基于 HMM 的语音合成系统(HMM based Speech synthesis System, HTS)。HTS 可以在不需人工干预的情况下, 高效自动的搭建合成系统, 由于统计的缘故, 对发音人和发音风格的依赖较小, 合成语音的语音风格和音色容易人为控制, 并且合成系统的规模没有波形拼接的那么大。

进入 21 世纪, 随着深度学习在语音识别领域取得了突破性的进展, 深度神经网络 (Deep Neural Network, DNN)在语音合成中也有了大量的应用, 凌振华等人在传统 HTS 合成框架的基础上, 将深度置信网络 (Deep Belief Network, DBN) 作为语音参数后增强模型应用到语音合成中。利用 DBN 强有力的学习能力, 在语音合成的后端实现在谱参数上的一个更加精细的调整, 使得合成音质有了不少的改善; Heiga Zen 等[6]提出了基于 DNN 的统计参数合成方法, 其核心思想是直接通过深层神经网络来预测声学参数, 避免了基于 HMM 的语音合成方法中由于决策树聚类导致的模型精度降低, 从而提高了合成语音的音质。从 2014 年开始, 长短时记忆模型(Long and Short Times Memory, LSTM)在语音技术中的使用掀起了新的热点。Frank K. Soong 等人[7-8]进一步将双向 LSTM 神经网络应用到语音合成中, 大大提高了合成语音的音质。

而随着机器学习的发展, 基于深度学习的端到端语音合成技术取得了巨大的发展。端到端建模能化繁为简, 降低了系统构建的难度, 也有效避免传统方法多阶段建模导致的误差累积。不仅如此, 端到端语音合成方法还取得了性能上的大幅度提升, 甚至在某些数据集上达到了媲美真实语音的水平。

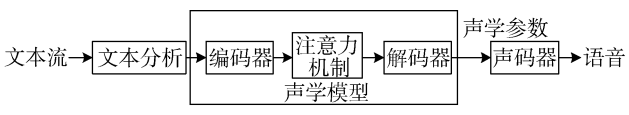

图 1 端到端语音伪造系统框架

Figure 1 End-to-end speech forgery framework

近年来, 一些学者致力于端到端的语音合成模型的建模, 并取得了性能上的巨大提升。2016 年, 谷歌 deepmind 研究团队提出了基于深度学习的

WavetNet 语音生成模型[9]。该模型可以直接对原始语音数据进行建模, 避免了声码器对语音进行参数化时导致的音质损失, 在语音合成和语音生成任务中效果非常好。然而由于该模型是样本级自回归采样的本质(sample-level autoregressive nature), 速度较慢。同时, 它还需要对来自现有语音合成文本分析前端的语言特征进行调节, 因此不是端到端的。另一个最近开发的神经模型是百度提出的 Deep Voice[10]和支持多说话人的 Deep Voice2[11], 它通过相应的神经网络代替传统参数语音合成流程中的每一个组件。但其中的每个组件都是独立训练出来的, 因此也不是端到端的。2017 年 1 月, Bengio 等人提出了一种端到端的用于语音合成的模型 Char2Wav, 其有两个组成部分: 一个读取器(reader)和一个神经声码器(nerual vocoder)。读取器用于构建文本(音素)到声码器声学特征之间的映射; 神经声码器则根据声码器声学特征生成原始的声波样本[12]。本质上讲, Char2Wav 是真正的意义上的端到端的语音合成系统。2017 年 3 月, 谷歌科学家王雨轩等人提出了一种新的端到端语音合成系统 Tacotron, 该模型可接收字符的输入, 输出相应的原始频谱图, 然后将其提供给 Griffin-Lim 重建算法直接生成语音。在美式英语测试里的平均主观意见评分达到了 3.82 分。此外, 由于 Tacotron 是在帧(frame)层面上生成语音, 所以它比样本级自回归(sample-level autoregressive)方式快得多[13]。谷歌科学家王雨轩等人还进一步将 Tacotron 和 WaveNet 进行结合, 在某些数据集上能够达到媲美人类说话的水平[14]。

2.3 语音韵律伪造研究进展[15]

语音韵律伪造是为了模拟目标说话人在日常生活语言交流中的态度, 设想以及注意力等个性化信息。它包括内涵意义和外延意义。内涵意义指的是说话者所表达的情感或者听者从中猜测的情感等等信息, 而外延意义指的是口头或者书面信息的语义内容。韵律在指导听者理解外延意义上有很重要的支持作用, 而且在提示内涵意义或者说话者对话语、听者以及整个交流的态度都起主要的作用。一个好的韵律预测可以让伪造语音与真实语音相比难以判断, 从而进一步提升语音伪造水平。韵律部分的研究是一个复杂的系统工程, 涉及语言学、语音学、心理学、语用学等学科的综合知识。一个语音单元除了由元音和辅音按时间顺序排列的音段成分之外, 还必须包括一定的超音段成分, 否则这个音节就不可能成为有区别意义的有声语言。韵律的声学参数一般包括基频、时长、能量。目前对韵律研究的重点是音高、音长、音强三个超音段参数在连续语流中的分布规律及其相互的作用, 而基本的研究方法仍是基于对生理特征的分析及对大语料库的统计分析。

对于一个语音伪造系统而言, 韵律预测是十分重要的, 它承启文本分析模块的分析信息, 生成对合成系统具有指导意义的声学参数, 是语音伪造系统中一个必不可少的模块。韵律预测的方法分为基于规则的方法和基于数据驱动的方法两类。早期的韵律预测采用基于规则的方法。研究人员需要大量的语音学试验的数据, 试图从中发现语音在不同的上下文环境就能得到针对不同语言、不同风格的韵律模型。

基频一直是语音合成研究的焦点。研究表明, 基频曲线对于不同的音节或音节组合, 有其基本的规律, 有相对稳定的变化模式, 这些为进一步的连续语流的音高曲线(语调)的研究奠定了基础。连续语音的音高曲线融入了发音人的生理特征、感情、语义、语境以及很多的个人特征信息。赵元任先生的“大波浪小波浪”学说以及“橡皮带”理论[16]是语调研究的奠基学说, 初步说明了语调的本质规律。沈炯则进一步扩充了这种思想, 提出了语调调节的“双线模型”[17]。Fujisaki[18]、Kochansaki[19]等结合发音生理机制及表面现象, 提出了控制语调的具体模型。这些理论及相应的模型都能够反映连续语流音高曲线的基本规律, 从而提高了对语音模拟的可操控性。

2.4 语音模拟研究进展

语音模拟是为了通过改变语音信号中发音人的属性而保持语音内容以及背景信息不变, 使得伪造语音听起来像是目标说话人发出的。相比于上述参数生成式语音伪造与波形拼接式语音伪造, 语音模拟系统的输入也是语音而不是文本, 主要实现的是对发音人的风格进行伪造与隐藏[15]。

语音模拟的基本框图可以用图表示, 实现一个完整的语音模拟系统一般包括训练和转换两个模块。在训练模块, 语音模拟系统提供了一组来自源说话人和目标伪造说话人的语音数据, 语音特征分析和映射特征这两个计算步骤将语音波形信号编码成可允许修改语音属性的表示形式, 然后将特征进行模型训练以得到映射或转换函数。在转换模块中, 利用训练后的转换函数对新的源说话人话语的特征进行转换, 然后利用转换后的语音特征合成转换后的语音。

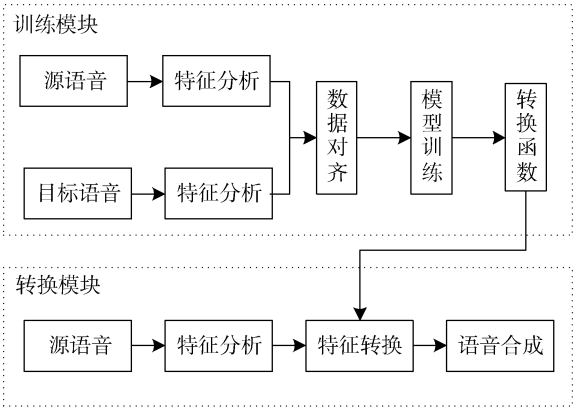

图 2　语音模拟基本框图

Figure 2　Speech simulation framework

语音模拟研究的重点主要在训练模块, 包括特征分析和对齐、转换模型的训练。数据对齐模块用于建立两说话人语音特征间的映射规则。有些语音转换要求源和目标数据提供相同文本的训练数据。在建立转换规则之前, 由于说话人发音速度的不一致, 需要对提取的两说话人语音特征进行时间对齐, 常用的方法有动态时间规整(Dynamic Time Warping, DTW)和基于语音识别中的隐马尔可夫模型切分的方法, 而有些语音模拟不要求平行的训练语料, 甚至可以任意训练语料, 这种情况还需要其他的特殊处理。针对所需训练数据形式的不同和数据对齐任务的不同, 语音模拟可以分为两种: 基于平行语料的语音模拟和使用非平行语料的语音模拟。对于基于平行语料的语音模拟, 需要源和目标伪造说话人提供平行语料, 即, 多个说话人基于同一文本的语音数据, 这样做的目的是最大程度地消除文本内容对说话人个性信息转换的影响, 因为在同样的录音环境及录音要求下使用相同的录音文本, 我们可以认为不同录音人的语音数据之间的差异主要是由说话人的不同导致的, 数据对齐的工作就是在时域上调整数据的对齐关系。而基于非平行语料的语音模拟由于训练语音的任意性, 只能在整个数据匹配的过程中尽量剔除语音内容信息造成的干扰, 难度要比平行语料的训练大很多。另外, 基于非平行语料的语音模拟还有一种特殊情况, 即跨语言语音模拟, 是指源说话人和目标伪造说话人提供的训练语料不属于同一种语言的情况。相比同语言语音转换, 这种情况有一个新的问题, 即不同语言之间的音素数据“失配”问题, 给数据对齐进一步造成麻烦。

语音模拟技术已有数十年的发展历程, 国内外关于语音转换技术的研究起源与 70 年代, 语音转换的研究最早可以追溯到 1971 年, Atal 使用 LPC 声码器来改变声音特性的工作, Abe 等人提出使用矢量量化码本映射来实现语音转换; 国内由初敏等人提出了基于时域基频同步叠加技术(Time Domain Pitch Synchronous Overlap Add, TD-PSOLA)的方法来实现男声与女声转换; 吴宪, 刘民航等采取矢量量化(VectorQuantization, VQ)和基音周期变换结合的方法; YiningChen 和 Min Chu 等使用平滑高斯混合模型(Gaussian Mixture Model, GMM)来克服 VQ 算法中的不连续现象; Seneff 等通过频谱包络估计的方法, 对语音信号解卷积, 实现语音转换; Desai 等提出使用 BP 神经网络的方法实现语音模拟。Tokuda 等提出基于隐马尔可夫模型的统计参数语音生成方法, 结合说话人自适应变换技术(Speaker Adaptation, SA)实现了不同说话人的语音模拟。Liang 等提出同享决策树聚类时 HMM 的参数, 用以提高转换语音的质量; 王民等使用 STRAIGHT 工具并构建深度信念网络(Deep Belief Nets, DBN)进行源说话人和目标说话人特征参数的训练, 再利用人工神经网(Artificial Neural Network, ANN)实现源说话人以及目标说话人的特征参数映射, 实现语音模拟。Kawahara Het 等人提出了一种动态编程算法, 在使用匹配最小化算法匹配源和目标说话人时, 同时估计最佳频率弯曲和权重变换; 还有使用神经网络建模实现基于频谱的语音转换方法。这些模型通常必须使用大量的目标说话人和源说话人的音频数据对进行训练。语音模拟中, 基于 HMM 或 GMM 模型是常用的方法。但是, 不管是基于 GMM 或者是 HMM 的方法, 都会有过度平滑的问题存在。而目前为止传统的所有语音模拟方法, 普遍存在语音模拟音质不高的问题。

随着计算设备的快速发展, GPU 集群的部署以及大数据技术的逐步应用, Graves 等人提出的深度学习 (Deep Learning, DL)方法发挥出惊人的建模能力。深度学习作为一种新的特征参数学习方法, 可以凭借构建多层非线性模型来实现对高维深度数据的特征表征。Zhang 等人使用的受限制玻尔兹曼机(Restricted BoltzmannMachine, RBM)模型可以直接自动地实现语音信号特征的提取、编码、降噪和升降维等; Hinton 等使用深度神经网络用于声学模型的构建, 并用以完成语音识别的研究工作; Latorre 等使用了多个国家的语料库进行 HMM 建模, 并且设计了相关的问题集, 实现了多语言的语音合成。另外, 一些学者使用了新的方法对原来经典的方法进行补充, 语音转换技术往往和语音合成技术结合起来研究。由于语音转换技术发展时间短, 上述方法都还没有研究到语言差异和说话人的特征差异细节部分,

也没有研究同一个人表达不同语言以及不同时期的发音差异, 因此模拟出的语音音质不够理想而且缺乏个性化的差异。

得益于深度学习的发展, 人们对以往的模型进行了新的改造。考虑到 GMM 模型和 HMM 模型都存在过平滑和过拟合等问题, 许多研究者选用深度的神经网络模型来对声学特征建模, 如 Sun, Lifa 等人提出了一种深度双向长短时记忆递归神经网络(Deep Bidirectional Long Short-Term Memory Recurrent Neural Network, DBLSTM-RNN)体系结构实现语音转换, 使用音素后验概率(PosteriorGrams, PPGs)[20]作为源和目标说话人之间的桥梁, 很好的解决了源说话人和目标说话人因语音内容不相同而引起的帧不对齐的问题。Saito 等使用高速路网络来进行语音模拟, 在转换的自然度和可懂度标准上都有所提高, 但是依旧存在过度平滑带来的语音自然度上不足的问题。为了进一步提高语音转换的质量, Chris Donahue 等提出使用基于深度卷积生成对抗网络的 Wave GAN 来实现语音模拟, 但由于直接简单的将语音信号处理为语谱图, 所以实验效果不理想, 表明不能简单的用 DCGAN 的思想来对语谱图做细致的建模。后来 Takuhiro Kaneko 等人提出用循环一致性的生成对抗网络(Cycle-Consistent Adversarial Networks, CycleGAN)实现一对一的语音模拟。在 2019 年, Zhang 等人[21]提出了一种带注意力机制的 WaveNet 编码器架构实现语音模拟。吴志勇等人提出了使用全局说话人嵌入(global speaker embeddings, GSEs)来控制语音模拟系统的转换目标, 利用与说话人无关的音素后验概率(PPGs)作为条件 WaveNet 声码器的局部条件输入, 同时从给定的话语中提取频谱特征, 并将其输入到参考编码器中, 并做一个注意力机制生成 GSEs, 并将其作为全局条件输入到 WaveNet 声码器中来控制合成语音波形, 可以仅用一句话就能实现语音模拟的方法, 并且语音音质也较好。

3 语音鉴伪研究进展

3.1 面向参数生成式语音鉴伪研究进展

检测参数生产式语音伪造的大多数方法依赖于特定参数生成算法的伪造。并且参数生成语音的语音参数的动态变化往往小于天然语音的动态变化, Satoh 等人[22]研究了使用帧内差异作为鉴别特征。该方法在无全局方差补偿的情况下, 可以很好地检测基于 HMM 的参数生成语音。Chen 等人[23]使用高阶梅尔倒谱系数(MCEPs)检测基于 HMM 的参数生成系统产生的语音。其中, 反映谱包络细节的高阶倒谱系数在 HMM 模型参数训练和生成过程中趋于平滑。因此, 参数生产语音的高阶倒谱分量比自然语音的变化小。虽然这种差异的估计提供了一种区分真实语音和参数生产语音的方法, 但这种方法是基于一个特定的 hmm 的语音参数生成系统的全部知识。因此, 同样的对策可能不适用于其他使用不同声学参数的生成器。

最近, 有一些工作尝试关注声码器和自然语音之间的声学差异。由于人类的听觉系统被认为对相位相对不敏感, 声码器通常不会重建语音类的相位信息。这导致了人耳和参数生成语音之间的相位谱的差异, 这些差异可被用于识别。这些方法在结合声码器的先验知识方面卓有成效。

在基于单元选择和统计参数化语音参数生成中可靠韵律建模的问题上, 其他参数生成语音检测方法使用 F0 统计量。统计参数语音生成方法生成的 F0 模式往往被过度平滑, 而单元选择方法在语音单元的连接点上经常出现“F0 跳变”, 二者得以区分。

3.2 面向波形拼接式语音鉴伪研究进展

由于波形拼接式语音的操作简单性, 该方法被大量应用于语音伪造上。工业界和学术界(Shang 和 Stevenson[24]; Villalba 和 Lleida[25];Wang 等人[26])对研究防止波形拼接式语音伪造展现出了极大的兴趣。

第一种波形检测方法是在 Shang 和 Stevenson[24] 的文章中报道的, 在一个依赖文本的自动说话人识别系统中, 使用固定的密码短语。检测器对新访问样本与存储的过去访问尝试实例进行比较。如果新访问产生的相似度得分高于预先定义的阈值, 则将其识别为波形拼接攻击。通过三个不同的通信渠道收集到的真实数据库和波形拼接访问的数据库, 并使用三个不同的波形拼接设备, 对检测性能进行评估。大量实验证实, 在大多数波形拼接检测实验中, 探测器都成功地降低了 EER。

Villalba 和 Lleida[25]提出了一种基于语谱比(spectral ratio)和调制指数(modulation indexes)的替代对策。这样做的动机来自于波形拼接时产生的噪音和混响。结果使频谱变平, 从而降低了调制指数。使用支持向量机对通过固定电话和 GSM 电话收集的真实录音和波形拼接语音的语谱和调制指数进行建模。研究结果表明, 对与文本无关的联合因素分析(JFA) 自动说话人确认系统, 固定电话的 FAR 将从 68%降低到 0%, GSM 电话的 FAR 将从 68%降到 17%。

Wang 等人[26]提出了一种基于信道噪声检测的

防波形拼接攻击对策。合法记录只包含来自自动说话人确认系统的记录设备的信道噪声, 而波形拼接攻击会招致由记录设备和用于波形拼接的说话人所引入的额外信道噪声。因此, 除了自动说话人确认系统的记录设备所引入的信道效应之外, 对信道效应的检测可以作为波形拼接攻击的指示器。实验表明, 欺骗干扰下, GMM-UBM 基准系统的 EER 从 40.17%下降到 10.26%。未来, 需要进行进一步的工作来制订更有效的对策。

3.3 面向语音模拟鉴伪研究进展

语音模拟与参数生成式语音伪造有一定的相似性, 因为一些波形转换算法使用了类似于统计参数语音生成的语音编码技术(Zen 等人[27])。因此, 第一批检测波形转换语音的工作借鉴了参数生成语音检测的相关工作(De Leon 等人[28])。

Wu 等人[29]将声码器引入了伪造语音, 将伪造的语音与自然语音区别开来。余弦归一化相位(cos-phase)和改进群延迟相位(MGD-phase)功能被证明是有效的。在 2006 年 NIST SRE 数据集进行的实验显示, 使用余弦相位可以使 EER 降低 5.95%, 使用 MGD 相位对策能使 EER 降低 2.35%。这项工作在 Wu 等人[30]中得到了扩展, 研究声音防伪性能对自动说话人确认的影响。通过使用对策, PLDA 自动说话人确认系统的 FAR 从 41.25%降低到 1.71%。

Alegre 等人[31]和 Alegre 等人[32]评估了一种方法, 既可以检测保留真实语音相位的语音转换攻击, 也可以检测人工信号攻击。Alegre 等人[31]的结果表明, 基于超向量的 SVM 分类器对人工信号攻击具有天然的鲁棒性, 而 Alegre 等人[32]表明, 语音转换攻击可以使用话语级别, 动态语音变换的估计值进行有效检测。转换后的语音比自然语音表现出更少的动态变化。

3.4 基于深度鉴别网络语音鉴伪研究进展

随着深度学习的快速发展, 基于深度学习的声音防伪检测系统开始进入人们的视线。深度学习的神经网络系统是一种特殊的机器学习系统, 最近几年成为了一种广泛使用的模型, 且应用于各种生物识别等任务中几乎均取得了当时最优的成果。在 2017 年的 ASVspoof 比赛中, 第一名运用了轻量型卷积神经网络(Light Convolutional Neural Network, LCNN), 获得了最优结果。每一个卷积层都用到了最大输出的方法, 这个方法构成了 Max-Feature-Map(MFM)层。

此外还有其他的分类器模型例如 Resnet、百度的 Deep Speaker 等都可以应用于防伪检测。在 2019 年的 ASVspoof 比赛中, 比赛的第四名就应用了 Deep Speaker 的框架。Deep Speaker 是百度提出的一种新的声纹识别方法, 他主要应用了残差网络 Resnet, 将网络层数变深, 深层网络可以比浅层网络更好地对语音建模。

4 趋势与挑战

4.1 语音伪造的趋势与挑战

现有声音伪造技术主要存在两方面的挑战: 一是自然口语声音的伪造很难接近真人; 二是资源受限条件下伪造声音的自然度和可懂度下降明显。

虽然目前端到端的声音伪造技术在特定数据集和限定条件下能伪造出逼近真人的声音, 但是仍然存在一些问题, 比如虽然发音和真人类似, 但往往是朗读风格, 并且需要单个发音人录制十多个小时的高音质语音作为训练数据。如果当发音人说话比较随意, 口语化, 并且只能获取很少量的训练语音(比如 1 分钟), 则伪造系统很难接近真人水平。因此, 进一步提高自然口语声音的伪造自然度和提升资源受限条件下伪造声音的音质是声音伪造的未来发展趋势。下面针对语音伪造的发展趋势分项阐述。

4.1.1 多风格语音伪造

互协同的多风格声音伪造旨在探索口语化语音模仿, 研究口语化风格在声音生成中的关键技术, 实现多样性差异化韵律风格建模, 提出不确定性语音模仿机制, 在符合言语习惯的情况下对语气词、语调变化及副语言等模仿建模。目前声学建模在语音模仿领域内测评中能够达到媲美人类录音的水平。然而这些研究只是单纯地将文本信息转化为中性风格的语音, 人长时间听后产生单调乏味的感觉, 而且没有携带应有的表现力发音风格, 使人听后容易产生理解上的偏差。因此, 基于上述分析面向多风格语音伪造的口语化韵律建模方法, 探索多风格协同促进生成将是未来发展趋势之一。

4.1.2 低成本语音伪造

自学习的低成本语音伪造是探索低资源下语音技术, 提出更加细粒度的说话人特征提取方法, 引入对抗学习机制, 构建高鲁棒性自适应声音模仿模型, 利用小样本数据增量式模仿目标语音, 达到较高的泛化性能。目前的声学建模所需要训练语料对于语音的音质要求高, 如常规的语音伪造训练数据要求录制环境相同。而在真实的生活环境下同一目标所能获取的且达到要求的数据很少。探索低资源下声音模仿技术, 提出更加精细化的说话人特征提

取方法, 引入对抗学习机制, 构建高鲁棒性自适应声音模仿算法, 快速利用有限量伪造目标语音, 达到较高的泛化性能, 伪造多种体裁内容的声音。因此, 基于上述分析面向低成本语音伪造的说话人特征提取方法与高鲁棒性自适应算法, 基于所用训练数据的自学习特征提取将是未来发展趋势之一。

4.1.3 鲁棒性语音伪造

自进化的高鲁棒性声音伪造, 旨在针对待模仿的语音目标采集所用信道的多样性, 选取抗干扰的声学参数特征作为模仿中介, 选取抗干扰的模仿算法, 对待模仿目标所采集声音进行分离, 实现清晰度可控的逼真声音模仿。对于目标说话人的语音采集渠道可以有多种方式, 但是每种方式获取的数据分布差异大, 并且含有较多的背景噪声, 所以可以考虑对于含有噪声的语音进行分解, 从而提取出纯正的语音。同时对背景噪声可以进行还原或篡改, 并且在对声学建模的过程中考虑到噪声对算法的精度下降, 不稳定和可懂度下降等影响。因此, 基于上述分析面向高鲁棒性语音伪造的抗干扰模仿算法及背景噪声建模与控制, 自进化形式语音伪造增强将是未来发展趋势之一。

4.2 语音鉴伪的趋势与挑战

现有声音防伪检测主要存在三方面的挑战: 一是缺乏真实声音与伪造声音间差异的理论研究; 二是防伪检测模型的通用性不足; 三是可解释性不足, 溯源性不足。一: 尽管现有研究取得了一定的进展, 但是缺乏对真实声音与伪造声音之间差异的理论研究, 不能从特征和信号层面解释真实声音和伪造声音之间的差异。二: 声音的防伪检测要求系统具有鲁棒性, 即能够检测出来自于多种不同的伪造系统的伪造声音。由于缺乏大规模数据集, 使得目前伪造的声音不具有多样性, 基于这样数据集训练出来的模型虽然能够鉴别出部分伪造声音, 但是普适性不足。三: 目前的神经网络模型, 本质上类似一个“黑盒子”, 虽然效果较传统方法有所提高, 但是溯源性不足。

4.2.1 多数据源声音鉴伪

声音伪造数据库的构建是指收集来自多种声音伪造模型的多个说话人的伪造声音。目前尚缺乏关于声音的伪造检测的数据库, 难以支撑声音伪造系统的落地应用。随着端到端的语音伪造与模仿技术的不断成熟, 伪造的声音已经足以以假乱真。现有伪造声音数据库已经不能适应目前的研究需求。因此, 全覆盖的声音伪造检测数据库构建研究将是未来发展趋势之一。

4.2.2 高普适性声音鉴伪

高区分性的语音鉴伪特征挖掘是指分析真实声音和伪造声音之间的差异, 提取能够反映这些差异的声音真伪特征用于模型训练。传统的声纹验证系统采用梅尔倒谱系数等特征来建立模型, 这些特征能够反映不同说话人之间的差异, 但是面对发音极其相似的伪造声音的时候, 上述特征难以具备鉴别度。为了准确辨识出真实声音和伪造声音, 针对真伪声音的发音进行系统的分析, 将设计提取更加具有鉴别度, 同时对于来自不同伪造系统的伪造声音具有区分度的语音鉴伪特征将是未来发展趋势之一。

4.2.3 可解释与溯源声音鉴伪

在信息安全领域, 利用伪造声音来进行欺骗、入侵的攻击者往往是针对人或系统构建出针对性的声音样本。这种样本迷惑性高, 特别是由于对系统有针对性, 使普通的异常检测模型难以识别出这种样本, 进而造成系统被攻破。因此, 如何构建出针对攻击样本的高鲁棒性异常检测模型就成为一个亟待解决的问题。另一方面, 攻击者往往会针对旧的防御手段使用新的攻击手法进行攻击, 比如构建新类型的伪声音样本, 使用新方法构建伪声音样本等。如何针对新类型的攻击样本种类快速地构建出高鲁棒异常检测模型, 也是未来发展趋势之一。

5 总结

本文对语音伪造与鉴伪的发展进行了梳理与阐释。针对语音伪造的适用场景与关键技术点, 分别对份风格伪造、音色与韵律伪造、语音模拟三大核心语音伪造技术的基本概念、发展历程、优势与不足进行梳理与分析。针对语音伪造的应对技术语音鉴伪技术, 首先介绍整理了针对性较强、面向三大语音伪造技术框架的语音鉴伪技术, 在此基础上介绍了具有普适性更强的基于深度鉴别网络语音鉴伪研究进展。此外, 结合已有研究现状的调研, 本文对语音伪造与鉴伪的趋势与挑战进行了多维度的阐述。语音伪造与鉴伪作为一组“攻”与“防”的耦合技术, 在未来会协同发展, 相互促进。

参考文献

[1] Zhizheng Wu, Pawel Swietojanski, Christophe Veaux, et al., A study of speaker adaptation for DNN-based speech synthesis[C]. *INTERSPEECH*, 2015.

[2] Fan Y C, Qian Y, Soong F K, et al. Speaker and Language Factorization in DNN-based TTS Synthesis[C]. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,

March 20-25, 2016. Shanghai. Piscataway, NJ: IEEE, 2016: 23-26.

[3] Hunt A J, Black A W. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database[C]. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Atlanta, GA, USA. Piscataway, NJ: IEEE, 1996 : 234-241

[4] H. Zen, K. Tokuda, , A. W. Black, Statistical parametric speech synthesis[J]. *Speech Communication*, 2009, 51(11): 1039-1064.

[5] Z. H. Ling, L. Deng , D. Yu, Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, 21(10): 2129-2139.

[6] H. Zen, A. Senior , M. Schuster, Statistical parametric speech synthesis using deep neural networks[C]. *ICASSP-2013-IEEE International Conference on Acoustics*, Speech, and Signal Processing (ICASSP), 2013: 7962-7966.

[7] H. Zen , H. Sak, Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis[C]. *ICASSP-2015-IEEE International Conference on Acoustics*, Speech, and Signal Processing (ICASSP), 2015: 4470-4474.

[8] Y. Fan, Y. Qian, F. L. Xie, et al, TTS synthesis with bidirectional LSTM based recurrent neural networks[C]. *INTERSPEECH 2014–Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.

[9] A. V. D. Oord, S. Dieleman, H. Zen, et al, WaveNet: A generative model for raw audio[EB/IT].2016: arXiv preprint arXiv:1609.03499.

[10] S. O. Arik, M. Chrzanowski, A. Coates, et al. Deep Voice: Real-time Neural Text-to-Speech[C]. *International Conference on Machine Learning (ICML)*, 2017:367-371.

[11] S. Arik, G. Diamos, A. Gibiansky, et al. Deep Voice 2: Multi-Speaker Neural Text-to-Speech[C]. *NIPS- Annual Conference on Neural Information Processing Systems (NIPS)*, 2017:65-71.

[12] J. Sotelo, S. Mehri, K. Kumar, et al, Char2Wav: End-to-end speech synthesis[C]. *ICLR2017 workshop submission*, 2017:34-41.

[13] Y. Wang, R. Skerry-Ryan, D. Stanton, et al, Tacotron: Towards End-to-End Speech Synthesis[C]. *INTERSPEECH 2017 –Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017: 4006-4010.

[14] J Shen, R Pang, R J Weiss, et al. Natural TTS Synthesis by Conditioning Wavnet on MEL Spectrogram Predictions[C]. *2018 IEEE International Conference on Acoustics*, Speech and Signal Processing (ICASSP), April 15-20, 2018. Calgary, AB. Piscataway, NJ: IEEE, 2018: 373-376.e

[15] 林焘等.《北京语音实验录》[M], 北京大学出版社, 1985.

T Lin. Beijing Speech experment record[M]. Peking Vniversity Press, 1985.

[16] Fujisaki W, Shimojo S, Kashino M, et al. Recalibration of Audiovisual Simultaneity[J]. *Nature Neuroscience*, 2004, 7(7): 773-778.

[17] Kochanski G, Shih C. Prosody Modeling with Soft Templates[J]. *Speech Communication*, 2003, 39(3/4): 311-352.

[18] Sun L F, Li K, Wang H, et al. Phonetic Posteriorgrams for Many-to-one Voice Conversion without Parallel Data Training[C]. *2016 IEEE International Conference on Multimedia and Expo (ICME)*, July 11-15, 2016. Seattle, WA, USA. Piscataway, NJ: IEEE, 2016: 564-571.

[19] Kaneko T , Kameoka H, Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks[OB/LE], 2017: arXiv:1711.11293.

[20] Satoh T., Masuko T., Kobayashi T. , Tokuda K., A robust speaker verification system against imposture using an HMM-based speech synthesis system[C]. *European Conference on Speech Communication and Technology (Eurospeech)*, 2001:879-881.

[21] Chen L W, Guo W, Dai L R. Speaker Verification Against Synthetic Speech[C]. *2010 7th International Symposium on Chinese Spoken Language Processing*, November 29-December 3, 2010. Tainan, Taiwan, China. Piscataway, NJ: IEEE, 2010: 345-351.

[22] Shang W, Stevenson M. Score Normalization in Playback Attack Detection[C]. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 14-19, 2010. Dallas, TX, USA. Piscataway, NJ: IEEE, 2010: 32-36.

[23] Villalba J, Lleida E. Detecting Replay Attacks from Far-Field Recordings on Speaker Verification Systems[M]. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011: 274-285.

[24] Wang Z.F., Wei G. , He Q.H., Channel pattern noise based playback attack detection algorithm for speaker recognition[C]. *IEEE Int. Conf. Machine Learning and Cybernetics (ICMLC)*, 2011:87-92.

[25] Zen H., Tokuda K. , Black A.W., Statistical parametric speech synthesis[J]. *Speech Commun*, 2009, 51: 1039–1064.

[26] De Leon P.L., Hernaez I., Saratxaga I., et al. Detection of synthetic speech for the problem of imposture[C]. *IEEE Int. Conf. on Acoustics*, Speech, and Signal Processing (ICASSP), 2011.

[27] Wu Z., Chng E.S. , Li H., Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition[C]. *INTERSPEECH* , 2012.

[28] Wu Z., Kinnunen T., Chng E.S., et al. Astudy on spoofing attack in state-of-the-art speaker verification: the telephone speech case[C]. *Asia-Pacific Signal Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2012:34-41.

[29] Alegre F., Vipperla R., Evans N., et al. Spoofing countermeasures

for the protection of automatic speaker recognition systems against attacks with artificial signals[C]. *INTERSPEECH*, 2012

[30] Alegre F., Amehraye A., Evans N., Spoofing countermeasures to protect automatic speaker verification from voice conversion[C]. *IEEE Int. Conf. on Acoustics*, Speech, and Signal Processing (ICASSP), 2013.

陶建华 于 2001 年在清华大学获得博士学位。现任中国科学院自动化研究所模式识别国家重点实验室研究员。研究领域为语音识别与合成、人机交互、多媒体信息。研究兴趣包括: 智能协同控制、自然口语语音交互、情感计算。Email: jhtao@nlpr.ia.ac.cn

傅睿博 于 2015 年在北京航空航天大学自动化专业获得学士学位。现在中国科学院大学模式识别与智能系统专业攻读博士学位。研究领域为语音合成。研究兴趣包括: 声学模型、迁移学习。Email: ruibo.fu@nlpr.ia.ac.cn

易江燕 于 2018 年在中国科学院大学模式识别与智能系统获得博士学位。现任中国科学院自动化研究所模式识别国家重点实验室助理研究员。研究领域为语音识别与合成。研究兴趣包括: 声学模型、迁移学习。Email: jiangyan. yi@nlpr.ia.ac.cn

王成龙 于 2018 年在合肥工业大学医学信息工程专业获得学士学位。现在中国科学技术大学控制科学与工程专业攻读硕士学位。研究领域为声纹识别、语种识别。研究兴趣包括: 声音防伪检测。Email: chenglong.wang@nlpr.ia.ac.cn

汪涛 于 2018 年在山东大学自动化专业获得学士学位。现在中国科学院自动化研究所攻读博士学位。研究领域为语音合成、语音转换。研究兴趣包括: 个性化语音合成、语音转换等。Email: wangtao2018@ia.ac.cn