语音伪装方法及其防御对策综述*

郑琳琳[1],孙 蒙[1],张雄伟[1],潘志欣[2]
(1.陆军工程大学,江苏 南京 210007;2.海军工程大学,湖北 武汉 430000)

摘 要:语音伪装是指以隐藏说话人身份为目的对说话人的个性特征进行的改变。近年来,随着智能语音交互技术和声纹认证产品的快速发展,语音伪装被应用到语音产品隐私保护中,但是也被不法分子用以实施违法犯罪行为。因此,语音伪装成为目前语音处理和信息安全领域非常有实用意义的研究问题。在简要介绍语音伪装的典型模型和方法的基础上,梳理了语音伪装威胁量化评估方案,总结了近几年伪装语音的防御对策,分析了目前防御对策中存在的问题及技术难点,并对语音伪装的未来研究发展方向进行了展望。

关键词:语音伪装;声纹识别;自动说话人确认;语音变换
中图分类号:TN912 **文献标识码**:A DOI:10.19358/j.issn.2096-5133.2020.08.007
引用格式:郑琳琳,孙蒙,张雄伟,等. 语音伪装方法及其防御对策综述[J]. 信息技术与网络安全,2020,39(8):33-42.

Voice disguise and its countermeasures for anti-disguise

Zheng Linlin[1], Sun Meng[1], Zhang Xiongwei[1], Pan Zhixin[2]
(1.Army Engineering University of PLA, Nanjing 210007, China;
2.Naval University of Engineering, Wuhan 430000, China)

Abstract: Voice disguise refers to the change of the personality characteristics of the speaker for the purpose of hiding the speaker's identity. In recent years, with the rapid development of intelligent voice interaction technology and voiceprint authentication products, voice disguise has been applied to the privacy protection of voice products, but it has also been used by criminals to commit illegal crimes. Therefore, voice disguise has become a very practical research issue in the field of voice processing and information security. On the basis of introducing the typical models and methods of voice disguise, the methods of quantifying voice disguise threat assessment methods are summarized in this paper, and the defense countermeasures of disguise voice in recent years are introduced. Then, the problems and technical difficulties still existing in defense countermeasures are analyzed. Finally, the prospects for the research development direction of voice disguise are given.

Key words: voice disguise; speaker recognition; automatic speaker verification; voice transformation

0 引言

语音是人们日常交流中的一种最直接、最有效和最常用的传递信息方式。由于说话人发音器官的生理差异和后天成长环境形成的行为差异,每个人的语音都带有强烈的个性特征,能够像虹膜、指纹、人脸等生物认证技术一样,成为身份验证的重要手段,称为声纹识别技术。声纹技术因其具有经济、可靠、交互自然等优势而备受关注,具有重要研究意义和广泛应用前景[1]。

虽然每个说话人的语音有自己的个性特征,但是语音也是可以被模仿和伪装的。目前,市面上流行的各类变声器及变声软件可以对说话人的语音进行个性化改变,致使人耳甚至部分声纹识别技术产品很难识别出说话人的身份[2]。犯罪分子利用特定手段来伪装自己的语音不被辨识出来,实施电话诈骗、恐吓、绑架勒索等相关新闻报道也是数见不

*基金项目:国家自然科学基金(61471394);江苏省优秀青年基金(BK20180080)

鲜。军事上,某些组织成员通过使用全新的电话号码和语音伪装的方式来逃脱政府监控的识别[3]。随着智能语音交互技术被广泛应用到商业活动和军事应用中,人们对信息安全的要求也越来越高。然而,语音伪装严重影响声纹识别效果,使犯罪分子有机可乘。

语音伪装(Voice Disguise)是指对于正常语音的任何改变、扭曲或者偏离[4]。它涵盖了故意伪装和非故意伪装两种形式。网络空间安全领域更多关注的是故意伪装,即"以掩盖真实身份为目的,有意识地改变声音,使其模糊、畸变、扭曲的发音方式"[5]。伪装语音的相关研究工作最早可追溯至20世纪六十年代初期的法庭说话人辨认,至今已有50多年的研究历史[6]。近年来,语音信号处理和互联网技术的进步,以及语音数据获取和共享的更加便捷,有力地推动了语音伪装技术的发展[7]。特别是基于机器学习和深度学习的语音合成技术[8]能够生成特定说话人的语音样本,对声纹识别接口的用户构成了严重的隐私威胁[9]。因此,语音伪装受到学术界和产业界的广泛关注,诸多国内外学者开展了与语音伪装相关的研究。日本东京国立资讯研究所、法国国家信息与自动化研究所以及美国伊利诺理工大学等开展了语音伪装方式的研究,进一步提高了伪装语音的匿名化程度;中国刑警学院和多地公安部门针对伪装语音变声规律及其对自动说话人确认系统(Automatic Speaker Verification,ASV)的影响展开了相关工作;清华大学、南京邮电大学以及中山大学等在伪装语音防御对策方面做了相关研究,并相继取得了一些研究成果。

本文在简要梳理语音伪装的典型模型和基本方法的基础上,介绍了语音伪装的威胁量化评估方法,归纳了语音伪装的防御对策,并总结了目前语音伪装防御对策研究中仍存在的问题和挑战,对未来的发展方向作出了展望。

1 语音伪装的典型模型和方法

语音的个性特征通常包括音色、音调、韵律特征和说话风格等方面,主要受到声道谱信息、共振峰频率和基音频率等参数的影响。语音伪装就是通过改变说话人的语音个性特征,故意隐藏或伪造说话人的身份。根据伪装方式的不同,语音伪装可以分为两种类型:人为伪装和电子伪装[10]。深入了解语音伪装的基本方法能够更好地防御伪装语音带来的安全威胁。

1.1 人为伪装

人为伪装是说话人借助本身的技能实施的语音伪装,大致可分为两种情况,一是刻意模仿某人的声音,如冒充领导;还有一种是故意改变自己原有的发音习惯,如捏鼻、咬物等,来伪装自己的声音不被辨识出来。人为伪装的具体伪装类型主要有调音、改变音素、改变韵律及变形等方式[2]。在调音伪装中,有改变音调、紧喉音、吸气音及耳语伪装等;改变音素伪装主要有使用方言、变更方言、鼻音化和模仿说话等;改变韵律的伪装有语调的改变,重音位置的调整,音段的拉长和缩短以及言语节奏的变化等方式;而变形主要指依靠外力阻碍正常的发音,如捏鼻子、捂嘴、咬物以及嚼物等。

人为伪装虽然能达到一定的伪装说话人身份的目的,但伪装效果受制于说话人自身的伪装能力。张翠玲等[11]研究了10种刑侦情况下的伪装形式,发现不同说话人受自身的调音能力和发音习惯的影响,伪装水平差异是普遍存在的,没有伪装经验的人伪装后更容易暴露身份。即使是专业的模仿者,也是模仿目标说话人的某些特定特征,如方言、韵律或者说话风格等[12],虽然改变了人耳的听觉感受,但是对自动说话人确认系统的欺骗干扰作用并不是特别明显。

1.2 电子伪装

电子伪装是指采用电子设备或语音处理软件对说话人的原始语音进行的变声伪装。与人为伪装相比,电子伪装使用电子设备及内置算法对语音时域或频域特性进行变形,得到的伪装语音要更加自然。随着深度学习技术的不断进步,可用于语音伪装的模型也越来越多,产生的伪装语音能更有效地隐藏说话人身份。因此,电子伪装以其高质量的伪装效果和便捷的实现方式,得到了越来越广泛的应用。目前成熟的电子伪装技术主要分为三类:基于基频线性变换的电子伪装、基于频谱非线性变换的电子伪装以及使用复杂转换函数的语音转换。

1.2.1 基于基频线性变换的电子伪装

语音基音频率(Fundamental Frequency),简称基频,是指发浊音时声带振动所引起的周期性振动频率,它反映了语音激励源的重要特征。语音学中,人类心理对语音基音频率的感知量可以用音调(Pitch)来描述。基于基频线性变换的电子伪装主要是通过

简单地修改基音频率来达到修改音调的目的。提高音调，语音变得尖锐；降低音调，语音变得低沉[13]。根据伪装作用域不同，基于基频线性变换的电子伪装可以分为频域伪装和时域伪装。

(1)频域基频线性变换伪装

频域基频线性变换伪装是指通过直接在语音频域内拉伸或压缩频谱来改变基音频率，从而提高或降低音调的伪装方式，该方式可以改变语音的音调而保持语音节奏不变。其伪装步骤示意图如图1所示。

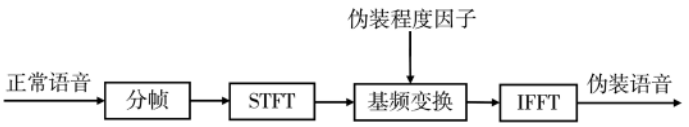

图1 频域伪装示意图

首先对语音信号分帧，然后对每一帧语言信号进行短时快速傅里叶变换(Short-Time Fourier Transform，STFT)，得到语音信号频域分析结果。对每一帧信号进行频谱的压缩伸展变换，同时利用插值法[14]对幅度谱进行相应处理。将变换后的频谱进行快速傅里叶逆变换(Inverse Fast Fourier Transform，IFFT)，即可得到频域伪装的语音信号。

频域基频线性变换电子伪装可以在很大伪装程度范围内对语音进行变形伪装，同时保持语音的自然度和可懂度。但是，利用基于频域的电子伪装方法对语音进行升调伪装时，频谱扩展，会造成语音高频部分缺失，伪装语音音频质量略显不足。

(2)时域基频线性变换伪装

时域基频线性变换伪装一般通过调整采样率和采用基音同步叠加(Pitch-Synchronous Overlap and Add Method，PSOLA)[15]相结合的方法来实现，这种伪装方式既改变了语音的音调，又改变了语速。调整采样率能够改变语音信号的基音频率从而改变音调。但是语音信号时频结构之间的约束性使得信号的时域特性和频域特性紧密相关，只利用调整采样率生成的伪装语音往往听起来不自然，需要结合PSOLA对语音进行进一步处理。PSOLA可以在误差最小准则下丢弃或重复部分语音帧，使伪装之后的语音与原来语音的频谱有着基本相同的包络，PSOLA工作原理如图2所示。

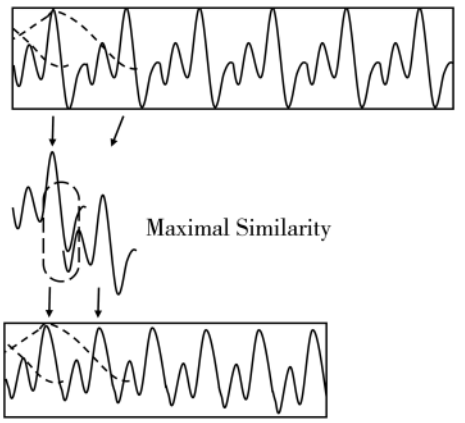

图2 基音同步叠加

由于时域基频线性变换伪装方法同时改变了语音音调和语速，因此，要保证伪装语音的自然度和可懂度，伪装程度的变化范围会受到限制，进而制约了伪装效果，故该方法在实际应用中有一定局限。

1.2.2 基于频谱非线性变换的电子伪装

基于频谱非线性变换的电子伪装方法是基于声道归一化(Vocal Tract Length Normalization，VTLN)技术实现的。人们认为，对于同样内容的语音，说话人声道长度的变化导致了语音波形的变化。VTLN可以通过翘曲函数(Warping Function)调整频谱的频率轴，来改变共振峰的位置和带宽，从而隐藏声道长度的个性特征[16]。从理论上讲，任何从定义域[0，π]到值域[0，π]的映射函数都可以作为VTLN中的翘曲函数，前提是翘曲函数需要保持伪装后语音的自然度和可懂度。

基于频谱非线性变换的电子伪装方法主要分为6个步骤：音调标记、帧分割、FFT、VTLN、IFFT和PSOLA。音调标记和帧分割的目的是将语音信号分割成与语音基音频率所决定的浊音伪周期性相匹配的帧，从而使输出的伪装语音具有最佳的音质。VTLN是频率弯折伪装中的关键步骤，它使用频率翘曲函数来修改每一帧的频谱。常用的翘曲函数包括对称分段线性函数、幂函数、二次函数及双线性函数等[17]。

为了抵御去匿名化攻击(De-anonymization Attacks)，提高伪装语音的伪装效果，基于频谱非线性变换的电子伪装方法的研究经历了从单帧变换到音段变换、从单一方法到多方法融合的过程，伪装质量不断提升。目前，基于频谱非线性变换的电子伪装方法研究主要集中在鲁棒性的频谱参数伪装变换函数方面。文献[17]提出了分段VTLN的方法，这种方

法的翘曲函数参数是可变的，随着时间的推移将频率轴向不同的方向变形。文献[18]通过随机选取翘曲函数参数、复合多种翘曲函数等方法来提高语音伪装机制的鲁棒性。

研究显示，基于频谱非线性变换的电子伪装方法能够最大程度地保持语音自然度，且鲁棒性能较好，但是其在伪装语音质量方面略显不足，还需结合其他方法以获得进一步提升。

1.2.3 基于语音转换的电子伪装

语音转换(Voice Conversion，VC)是指在保持说话内容信息不变的情况下，将一个人的声音特征通过修改变换，使其听起来像另一个人的声音。基于语音转换的电子伪装方法就是利用语音转换技术来隐藏源说话人身份信息，其原理如图3所示。与基频变换伪装方法和频谱变换伪装方法相比，基于语音转换的电子伪装方法需要目标说话人信息，伪装转换模型更加复杂。

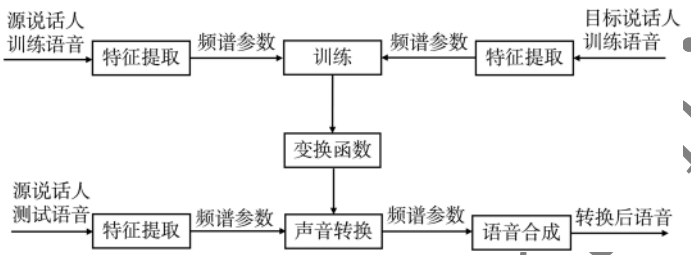

图3 基于语音转换的电子伪装方法原理图

基于语音转换的伪装方法核心思想是说话人的身份信息在整个说话过程中是静态的，而内容信息是动态的[19]。该方法可以基于神经网络变换，使用说话人编码器和内容编码器来分离身份信息和内容信息，对身份信息进行匿名化处理，然后再利用语音合成模型生成伪装语音[20]。这样，源说话人身份信息被压制，取而代之的是一种匿名的伪身份信息。

随着神经网络模型的不断改进和发展，结合不同语音特征采用不同的网络转换模型的伪装方法不断提出。文献[21]提出了利用目前最先进的说话人身份特征x-vector[22]和神经波形模型相结合的伪装方法。该方法使用基于深度神经网络(Deep Neural Network，DNN)的自动语音识别系统(Automatic Speech Recognition，ASR)以音素后验图(Phoneme Posteriorgram，PPG)[23]的形式捕获语言信息，并使用预先训练的x-vector系统对说话人身份进行编码，然后通过多个随机x-vector组合派生出匿名的伪说话人身份。在给定PPG和伪装的x-vector的情况下，通过神经声学和波形模型[24]来生成伪装语音。实验结果表明，该方法能在保持伪装语音高质量的同时，有效地隐藏说话人身份。

基于语音转换的电子伪装方法的本质是参数的多元回归模型，通过添加网络层数、高维特征序列和训练数据量等多种手段可以有效提升伪装语音的质量。随着参数的增多，伪装转换模型的表示能力会不断增强。但当训练数据不充分时，就会发生过拟合现象，导致性能急速下降。同时，目标说话人语音信息的依赖也成为制约此类方法伪装效果的一个重要因素。

2 威胁量化评估

语音伪装研究的主要目的是增大说话人身份识别系统的辨识难度，同时保持伪装语音的可懂度及算法的低复杂度。由此可知，如果想要对语音伪装方法进行威胁量化评估，可以利用说话人识别系统的性能指标来实现：利用某种语音伪装方法对待测语音进行伪装处理，然后利用说话人识别系统进行身份识别，说话人识别系统性能下降越明显，说明该语音伪装方法威胁越大。目前，语音伪装方法的威胁评估测试主要有主观和客观两种手段。

2.1 主观评估

主观评估就是以人为主体，通过人的主观感受来对语音进行测试。由于语音伪装最直接的目的是改变人耳的听觉感受，因而主观评估是最基本的评估方法。检测错误率(Detection Error Rate，DER)是语音伪装的主观威胁量化评估的常用标准之一。这种测试方法使用若干组语音对进行测试，每对语音有50%的概率来自同一个说话人。测评人需要判断所听到的每对语音是否来自同一个说话人，全体测评人判断错误的百分比就是DER得分，包含虚警(False Alarm)和误识(False Rejection)。

主观评估是建立在人的感觉的基础上，测试结果可能因人而异。为了尽可能减小个体差异的影响，主观评估的方案设计必须要周密，参加测试的测评人要足够多，测试环境应该尽量保持相同，所测语音音频也要足够丰富。测试语音必须仔细地选择发音，以保证所选样本具有代表性，同时还要保证能够覆盖所有类型的语音。例如，有的语音伪装方

法在浊音的处理上比较好，但伪装后的清音则太模糊；而有的语音伪装方法在低频段的性能较好，甚至会直接将高频段丢弃。所以，在选择测试样本时，不仅要包含男声、女声，同时还应该选择不同年龄段的语音。

2.2 客观评估

通过以上对主观评估方法的简单介绍可以看出，主观评估虽然是语音伪装威胁量化评估最基本的方法，但它的缺点也很明显：灵活性差、费时费力以及可重复性差等。针对主观评估方法的不足，基于主观测度的客观评估方法被提出。

目前，说话人身份伪装效果的主要客观衡量指标是自动说话人确认系统的等错误率(Equal Error Rate，EER)。在自动说话人确认系统中，系统可能把伪装者误认为目标说话人而错误地接受，为错误接受率(False Acceptance Rate，FAR)；也可能把目标说话人误认为伪装者而错误地拒绝，为错误拒识率(False Rejection Rate，FRR)。两个指标对应的公式如下[25]：

$$FAR=\frac{\text{伪装者中误识的语音个数}}{\text{来自伪装者的语音总数}}\times 100\%$$

$$FRR=\frac{\text{目标说话人语音中漏识的语音个数}}{\text{来自目标说话人的语音总数}}\times 100\%$$

FAR 和 FRR 是两个矛盾的参量指标，一个指标降低会导致另一参量上升。自动说话人确认系统的性能指标用 EER 来表示，它是 FAR 和 FRR 相等时系统的性能，代表了 FAR 和 FRR 的一个平衡点。当利用 EER 评估伪装效果时，EER 的数值越大，说明自动说话人确认系统的识别效果越差，同时也说明了语音伪装造成的威胁越大，伪装效果越好。

3 防御对策

语音伪装防御系统具有的先验知识的不同，造成了语音伪装防御效果的很大差别。根据语音伪装防御系统对语音伪装方式及其参数的知情程度，可以将语音伪装防御场景分为三种类型：

(1)黑盒系统。语音伪装防御系统完全不知道测试语音经过了语音伪装处理。

(2)白盒系统。语音伪装防御系统知道测试语音采用的完整伪装策略，包括伪装处理方法和确切参数值。

(3)灰盒系统。在以上两种极端情况之间，可以定义第三种语音伪装防御系统，该系统知道测试语音采用的部分伪装策略。例如，灰盒系统知道语音伪装方法，但不知道它的参数值。这种防御场景可能更实际，因为语音伪装处理方法可能是开源的，但伪装者使用的具体参数策略则不太容易获取到。

语音伪装技术的出现给说话人识别系统带来很大的困难，实验发现，不采取伪装防御策略，利用当前最先进的基于 x-vector 的自动说话人确认模型对电子伪装后的语音进行识别，EER 高达 30%以上，几乎无法辨认出伪装者的身份。但是，采用了语音伪装防御策略的说话人识别系统 EER 明显降低，白盒语音伪装防御系统的 EER 可以降至 3.9%[26]。

随着智能语音交互应用的不断发展，语音代表个人身份特征的场景日益广泛，急需有效的语音伪装防御对策的出现。本节将概括目前已有的语音伪装判别策略，并分别介绍针对人为伪装语音和电子伪装语音的身份辨识对策。

3.1 语音伪装判别策略

在进行说话人身份鉴定之前，有效判断待测语音是否经过伪装以及经过何种类型的伪装，是后续选择合适说话人身份辨识系统的前提，可有效提高声纹识别的识别率。

语音伪装判别的研究主要基于语音伪装能够对音质和部分语音特征产生一些重要的影响。研究人员在仔细分析了伪装语音的生成原理后发现，语音伪装过程可能会导致生成的伪装语音与自然语音在某些语音特征方面存在差异，因此可以利用这些不一致性构建检测特征。例如，文献[27]提出了 MGDCC(Modified Group Delay Cepstral Coefficients)特征，它同时综合了语音频谱中的幅度和相位信息；文献[28]根据伪装语音与正常语音基音周期之间的差异，利用 PP(Pitch Pattern)特征进行语音伪装鉴定。

目前，关于语音伪装鉴定方法的研究已经取得了不错的成果，采用图 4 所示的特征参数与分类器相结合的方法能达到较高的检出率。HUANG J W 等[29-31]提出了利用 MFCC 作为声学特征，采用 SVM 分类器从真实语音中检出电子伪装语音的算法。采用交叉伪装法和交叉语料库对算法进行测试，伪装语音的检出率均可达到 90%以上。李燕萍等[32]在前人工作的基础上提出了一种 SVM 分类器结合高斯混合模型(Gaussian Mixture Model，GMM)均值组合

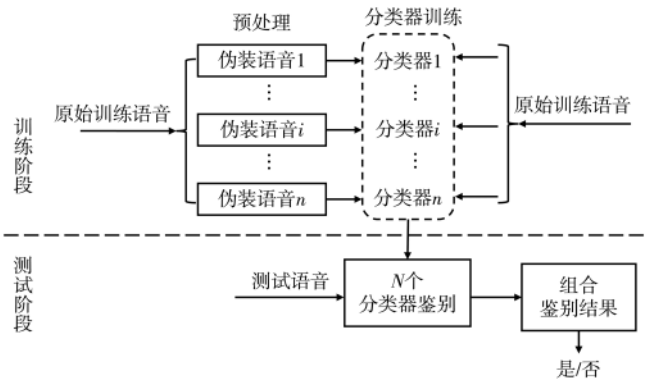

图 4 电子伪装鉴定原理框图

特征参数的电子语音伪装鉴定方法，通过 GMM 模型对电子伪装语音建模，将其均值矢量构成组合特征向量作为SVM 分类器训练和鉴别的特征参数。实验结果证明，这种方法对于电子伪装语音的鉴定率达到 90%。

为了提高对低失真语音的检测，一些机器学习及深度学习模型也被应用到语音伪装判别工作中。文献[33]利用卷积神经网络(Convolutional Neural Network，CNN)来识别检测电子伪装语音信息，准确率高于 95%。文献[34]提出了一种基于稠密卷积网络的伪语音检测方法，通过对核约简的优化和对瓶颈层的利用，达到了较高的计算效率，对数据库内部和跨数据库的平均准确率为96.45%，优于目前已有的方法。

当前关于语音伪装鉴定算法的研究主要集中于降低特征参数维度及计算复杂性、提高跨库交叉检出率、减轻对后续声纹识别系统的冗余影响等方面。

3.2 人为伪装语音身份辨识对策

由于人为伪装的伪装效果受到伪装者自身伪装能力的影响，即使是专业的模仿者，也仅仅是模仿目标说话人的部分特征。因此，人为伪装的防御主要集中在研究伪装过程中不变的语音特征参数。

为了探求人为伪装与语音特征参数之间的关系，研究人员针对不同的伪装方法及不同的语音特征参数做了相关研究。例如，文献[35]分析了咬物伪装对元音共振峰的影响，并详细描述了共振峰的比例变化；文献[36]研究了改变音调及捏鼻子等非电子语音伪装对语音基音频率的影响；而文献[37]研究了耳语伪装对基音频率、语音强度及音质的影响。研究发现，说话人识别中常用的特征参数会受到人为伪装的干扰，一定程度影响 ASV 系统的识别效果。

清华大学信息技术研究院语音和语言技术中心(CSLT)王东在研究中发现，人与人对话中无处不在的琐碎事件，如咳嗽、大笑、“喂”等，虽然时长较短且不清晰，但在伪装语音身份鉴定的情况下是非常有价值的。因为它们较少受到人为故意改变，所以可以用来从伪装语音中识别说话人身份[38]。实验发现，利用琐碎事件对人为伪装进行声纹识别，识别效果有了很大改进[39]。

3.3 电子伪装语音身份辨识对策

对于电子伪装语音的身份辨识主要考虑两种思路：一种是将伪装语音还原得到正常语音，然后利用目前发展成熟的 i-vector 或 x-vector 等自动说话人确认系统进行识别；另一种是设计伪装语音特征补偿算法，对现有的自动说话人确认系统进行改进。

3.3.1 电子伪装语音的还原

电子伪装语音的还原是指通过一定的算法来消除语音中的电子伪装特征，生成更为接近原始音频的语音。电子伪装语音还原最直接的方法是推导出变声算法的逆运算，然后根据逆运算算法处理伪装语音，从而得到原始正常语音。然而这种方法受到伪装算法的封闭性和多样性制约，很难得到推广。但是原始语音转换为电子伪装语音的过程存在一定的变化规律，因此可以通过统计对比原始语音与电子伪装语音之间的声纹偏差特征，为电子伪装语音的还原提供依据。目前，伪装语音还原算法可分为基于特征变化规律的传统还原方法以及基于深度学习技术的还原方法[40]。

(1)基于特征变化规律的还原方法

南京邮电大学林乐[41]根据电子伪装语音的变声规律，利用语音信号重采样技术和基音同步叠加方法实现了电子伪装语音的还原。该方法首先采用重采样技术将电子伪装语音的基音频率调整至与正常语音相接近的程度，然后利用 PSOLA 的方法在保持基音频率相对稳定的情况下，将语音时长调整至正常水平。

实验发现，利用这种方法还原后的电子伪装语音虽然丢失了部分语音细节，但仍然保留了用于辨认说话人身份的信息(主要集中在低频部分)，因此可以通过还原处理后的电子伪装语音识别出该语音的说话人。

(2)基于深度学习技术的还原方法

随着深度学习技术的广泛应用,华东政法大学王永全提出了一种基于扩大的因果卷积神经网络(Dilated Casual Convolution Neural Network, DC-CNN)的电子伪装语音还原模型[42]。该还原模型具有非线性映射性、扩展性、多适应性与条件性、并发性等明显特点,能有效削减语音中的电子伪装特征。将还原语音与原始语音进行声纹特征比对、LPC数据分析和语音同一性的人耳测听辨识,结果表明,还原语音与原始语音的声纹特征十分吻合,且实现了高质量的共振峰波形复原,钢琴曲和英文语音的共振峰参数总体还原拟合率分别达到79.03%和79.06%,远超电子伪装语音与原始语音35%的相似比例,较好地实现了电子伪装的钢琴曲和英文语音的还原。

3.3.2 自动说话人确认系统补偿策略

(1)基于DTW模型补偿的识别方法

南京邮电大学陶定元[43]提出了基于DTW模型补偿的电子伪装语音说话人识别方法,如图5所示。该方法提取语音的梅尔倒谱系数(Mel Frequency Cepstral Coefficients, MFCC)作为特征参数,通过动态时间规整(Dynamic Time Warping, DTW)模型进行伪装程度鉴定,再利用矢量量化(Vector Quantization, VQ)模型进行说话人识别,从而设计了DTW与VQ相结合的电子伪装语音说话人识别系统。实验结果表明,该系统一定程度上缓解了VQ说话人识别系统对电子伪装语音识别率过低的问题,识别效果得到了明显改善。

(2)利用基频比补偿特征参数的识别方法

针对基于基频线性变换的电子伪装语音,文献[44]提出用基频比来估计伪装程度,进而还原语音

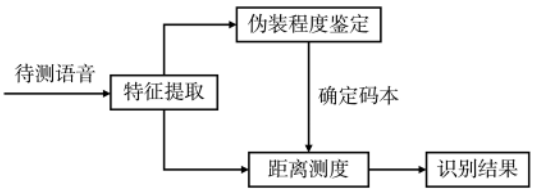

图5 基于DTW模型补偿的伪装语音说话人识别框图

特征参数的抗伪装攻击的说话人识别系统,其原理框图如图6所示。该方法根据待测语音与注册语音的基频比估计伪装程度,利用估计出的伪装程度修正待测语音的MFCC,从而得到还原后的MFCC特征。将提出的方法作为特征还原工具应用于GMM-UBM说话人识别系统的前端,可提高电子伪装语音伪装者的识别准确率,EER仅为3%~4%,明显优于未经还原的MFCC特征的40%。

4 防御对策中存在问题及发展趋势

4.1 存在问题

虽然关于伪装语音防御对策研究经过了几十年的发展,但是目前仍然存在一些问题和挑战,归纳起来有以下几个方面:

(1)对于伪装语音的语料质量要求过于苛刻。研究发现,当伪装语音含有噪声或者伪装语音由多种伪装方式组合生成时,利用现有防御对策得到的说话人识别EER明显增大,这说明当前存在的语音伪装防御对策在应对复杂情况下的伪装语音语料时失效。伪装语音的说话人身份鉴定技术真正应用到实际中时,通常情况下不可避免地受到各种噪声的污染,很难直接获取高质量的伪装语音。由于录音环境及伪装手段未知,噪声及其统计特性都难以获取,给伪装语音的研究带来了新的问题。

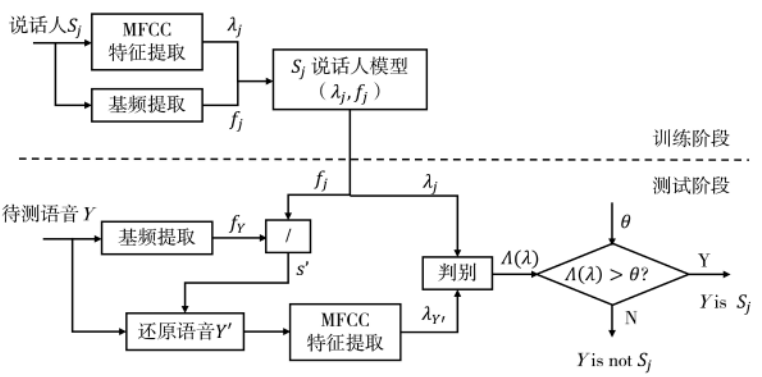

图6 利用基频比补偿特征参数的电子伪装语音说话人识别框图

(2)伪装语音还原算法的研究还有待发展。虽然目前伪装语音还原算法取得了一定发展和改善,但是与原始正常语音相比还是存在一定差距。例如,基于基频线性变换的电子伪装语音的高频部分会存在缺失,目前的还原算法侧重于还原人耳听觉系统敏感的低频部分,对高频部分的还原质量不高,会引入一些不必要的噪声,因此还需进一步提升还原语音与原始语音的相似度。另外,当前现有的还原方法过于依赖先验知识,只针对特定的伪装方式,这显然不符合实际要求。

(3)伪装语音防御策略通用性不强。当前伪装语音防御策略的相关研究针对的伪装语音伪装方式都比较单一,但现实应用中,伪装软件种类繁多,伪装手段不尽相同,伪装者可能会将人为伪装方法和电子伪装方法结合运用。因此,需要提出一个具体的、稳健的、普遍的解决方案,即使不知道语音伪装方法,依然能够有效鉴别伪装语音的说话人身份。

4.2 发展趋势

目前语音伪装防御策略还存在很多问题和挑战,语音伪装的相关研究也一直是语音信号处理领域以及网络空间安全领域的热点问题。本文认为,未来伪装语音身份鉴定相关研究也必将着力解决当前伪装语音中存在的现实问题,朝着下述方向不断发展:

(1)普适的伪装语音防御方法

目前伪装语音的身份鉴定容易受到伪装方式的影响,未来的研究方向必定是研究具有通用性、高效性的伪装身份鉴定方式,提升伪装语音身份鉴定效果。针对由人为伪装和电子伪装结合产生的伪装语音,可以考虑在进行电子伪装语音鉴定前消除非电子伪装方式的影响;对于复杂多变的电子伪装语音还原方法,可以试图寻找一个通用的非线性还原函数,通过调节还原函数的参数,来逼近伪装函数的反函数,从而实现电子伪装语音的还原,为伪装语音身份识别奠定基础。

(2)鲁棒的伪装语音防御方法

伪装语音的研究最终将会真正运用到实际,而真实情况下,伪装语音噪声信号混杂,语言也会出现各种各样的情况。针对目前伪装语音防御策略在真实含噪数据集上效果不理想问题,下一步的研究可以结合当前发展比较成熟的语音信号预处理和语音增强技术,在不损失待测语音音质的条件下有效去除噪声,然后再进行语音伪装判别及伪装语音身份鉴别,这将会是另一个提高伪装语音防御系统性能的重要方式。

(3)可靠的伪装语音防御方法

伪装语音的防御对策研究的最终目标是要保证识别结果准确。近年来,神经网络、机器学习以及深度学习的发展使得语音增强、语音合成以及说话人识别等相关技术取得了较大进展。未来可以尝试采用这些先进技术相结合,选取更加优秀的匹配方法来提高伪装语音身份鉴定准确度。另外,有效地提高训练鉴定速度和系统的稳定性是可靠的伪装语音防御模型的必备条件,这也将是以后的研究重点之一。

5 结束语

声纹识别技术的普及给人们的生活带来了极大的便利,同时人们对于信息安全有着越来越高的需求和期望。然而,语音伪装技术的出现给声纹认证产品带来了极大挑战。本文概括了常用的语音伪装方法,介绍了伪装语音的威胁量化评估指标,讨论了语音伪装防御对策目前存在的问题并给出研究方向。未来的伪装语音防御对策会朝着普适性、鲁棒性、可靠性方向发展,同时,抗伪装的说话人识别技术的发展也必将进一步推动声纹识别技术的落地应用和发展。

参考文献

[1] KINNUNEN T, LI H Z. An overview of text-independent speaker recognition: from features to supervectors[J]. Speech Communication, 2010, 52(1): 12-40.

[2] PERROT P, AVERSANO G, CHOLLET G. Voice disguise and automatic detection: review and perspectives[C]. Progress in Nonlinear Speech Processing. Heidelberg, Berlin: Springer, 2007: 101-117.

[3] 石军. 智能音频检索技术在侦收系统中的应用研究[J]. 通信技术, 2016, 49(10): 1415-1418.

[4] FARRÚS M. Voice disguise in automatic speaker recognition[J]. ACM Computing Surveys, 2018, 51(4): 1-22.

[5] PERROT P, CHOLLET G. The question of disguised voice[J]. Journal of the Acoustical Society of America, 2008, 123(5): 3878.

[6] 王志飞. 数字音频司法鉴定技术研究[D]. 厦门: 厦

门大学, 2014.

[7] SRIVASTAVA B M L, BELLET A, TOMMASI M, et al. Privacy-preserving adversarial representation learning in ASR: reality or illusion? [EB/OL]. (2019-11-12) [2020-04-28]. https://arxiv.org/abs/1911.04913.

[8] LORENZO-TRUEBA J, FANG F M, WANG X, et al. Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data[EB/OL]. (2018-03-02)[2020-04-28]. https://arxiv.org/abs/1803.00860.

[9] NAUTSCH A, JIMÉNEZ A, TREIBER A, et al. Preserving privacy in speaker and speech characterization[J]. Computer Speech & Language, 2019, 58: 441-480.

[10] ZHANG C L, LI B, CHEN S, et al. Acoustic analysis of whispery voice disguise in Mandarin Chinese[C]. Annual Conference of the International Speech Communication Association, Hyderabad, India: [s.n.], 2018.

[11] ZHANG C L, TAN T J. Voice disguise and automatic speaker recognition[J]. Forensic Science International, 2008, 175(2-3): 118-122.

[12] ZETTERHOLM E. Voice imitation: a phonetic study of perceptual illusions and acoustic success[M]. Sweden: Lund University, 2003.

[13] 张桂清, 金怡珠, 刘红伟, 等. 电子伪装语音的变声规律研究[J]. 证据科学, 2010, 18(4): 503-509.

[14] LAROCHE J. Time and pitch scale modification of audio signals[M]. Applications of Digital Signal Processing to Audio and Acoustics. Boston, MA: Springer, 2002: 279-309.

[15] RIBARIC S, ARIYAEEINIA A, PAVESIC N. De-identification for privacy protection in multimedia content: a survey[J]. Signal Processing: Image Communication, 2016, 47: 131-151.

[16] COHEN J, KAMM T, ANDREOU A G. Vocal tract normalization in speech recognition: compensating for systematic speaker variability[J]. The Journal of the Acoustical Society of America, 1995, 97(5): 3246-3247.

[17] SUNDERMANN D, NEY H. VTLN-based voice conversion[C]. Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology. Darmstadt, Germany: IEEE Press, 2003: 556-559.

[18] QIAN J W, DU H H, HOU J H, et al. Hidebehind: enjoy voice input with voiceprint unclonability and anonymity[C]. Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems. New York, USA: ACM Press, 2018: 82-94.

[19] CHOU J, YEH C, LEE H. One-shot voice conversion by separating speaker and content representations with instance normalization[C]. Annual Conference of the International Speech Communication Association, Graz, Austria: [s.n.], 2019.

[20] ULYANOV D, VEDALDI A, LEMPITSKY V. Improved texture networks: maximizing quality and diversity in feed-forward stylization and texture synthesis[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE Press, 2017: 6924-6932.

[21] FANG F M, WANG X, YAMAGISHI J, et al. Speaker anonymization using x-vector and neural waveform models[EB/OL]. (2019-05-30)[2020-04-28]. https://arxiv.org/abs/1905.13561.

[22] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: robust DNN embeddings for speaker recognition[C]. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE Press, 2018: 5329-5333.

[23] 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). Seattle, USA: IEEE Press, 2016: 1-6.

[24] WANG X, TAKAKI S, YAMAGISHI J. Neural source-filter-based waveform model for statistical parametric speech synthesis[C]. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom: IEEE Press, 2019: 5916-5920.

[25] DU S. Robust speaker verification system with anti-spoofing detection and DNN feature enhancement modules[D]. Singapore: the School of Computer Engineering of the Nanyang Technological University, 2015.

[26] SRIVASTAVA B M L, VAUQUIER N, SAHIDULLAH M, et al. Evaluating voice conversion-based privacy protection against informed attackers[C]. 2020 IEEE International Conference on Acoustics, Speech and

Signal Processing(ICASSP). Barcelona, Spain: IEEE Press, 2020: 2802-2806.

[27] WU Z Z, XIAO X, CHNG E S, et al. Synthetic speech detection using temporal modulation feature[C]. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Vancouver, Canada: IEEE Press, 2013: 7234-7238.

[28] LEON P L D, STEWART B, YAMAGISHI J. Synthetic speech discrimination using pitch pattern statistics derived from image analysis[C]. Thirteenth Annual Conference of the International Speech Communication Association. Portland, OR, USA: [s.n.], 2012.

[29] WANG Y, DENG Y H, WU H J, et al. Blind detection of electronic voice transformation with natural disguise[C]. International Workshop on Digital Watermarking. Heidelberg, Berlin: Springer, 2012: 336-343.

[30] WU H J, WANG Y, HUANG J W. Blind detection of electronic disguised voice[C]. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Vancouver, Canada: IEEE Press, 2013: 3013-3017.

[31] WU H J, WANG Y, HUANG J W. Identification of electronic disguised voices[J]. IEEE Transactions on Information Forensics and Security, 2014, 9(3): 489-500.

[32] 李燕萍, 林乐, 陶定元. 基于GMM统计特性的电子伪装语音鉴定研究[J]. 计算机技术与发展, 2017, 27(1): 103-106.

[33] LIANG H X, LIN X D, ZHANG Q, et al. Recognition of spoofed voice using convolutional neural networks[C]. 2017 IEEE Global Conference on Signal and Information Processing. Montreal: IEEE Press, 2017: 293-297.

[34] WANG Y, SU Z. Detection of voice transformation spoofing based on dense convolutional network[C]. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Brighton, United Kingdom: IEEE Press, 2019: 2587-2591.

[35] DE FIGUEIREDO R M, BRITTO H S. A report on the acoustic effects of one type of disguise[J]. International Journal of Speech, Language and the Law, 1996, 3(1): 168-175.

[36] KÜNZEL H J, GONZALEZ-RODRIGUEZ J, ORTEGA-GARCÍA J. Effect of voice disguise on the performance of a forensic automatic speaker recognition system[J]. IEEE Personal Communications, 2004, 7(6): 32-35.

[37] ORCHARD T L, YARMEY A D. The effects of whispers, voice-ample duration, and voice distinctiveness on criminal speaker identification[J]. Applied Cognitive Psychology, 1995, 9(3): 249-260.

[38] ZHANG M, KANG X F, WANG Y Q, et al. Human and machine speaker recognition based on short trivial events[C]. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Calgary, Canada: IEEE Press, 2018: 5009-5013.

[39] ZHANG M, CHEN Y X, LI L T, et al. Speaker recognition with cough, laugh and" Wei"[C]. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference(APSIPA ASC). Kuala Lumpur, Malaysia: IEEE Press, 2017: 497-501.

[40] 王永全. 声像资料司法鉴定实务[M]. 北京: 法律出版社, 2013.

[41] 林乐. 基于高斯超向量和支持向量机的电子伪装语音鉴定方法研究[D]. 南京: 南京邮电大学, 2016.

[42] 王永全, 施正昱, 张晓. 基于DC-CNN的电子伪装语音还原研究[J]. 计算机科学, 2019(8): 183-188.

[43] 陶定元. 电子伪装语音下的说话人识别方法研究[D]. 南京: 南京邮电大学, 2016.

[44] WANG Y, WU H J, HUANG J W. Verification of hidden speaker behind transformation disguised voices[J]. Digital Signal Processing, 2015, 45: 84-95.

(收稿日期: 2020-05-29)

作者简介:

郑琳琳(1992-), 女, 硕士, 主要研究方向: 语音处理与网络安全。

孙蒙(1984-), 通信作者, 男, 博士, 副教授, 主要研究方向: 智能语音处理、机器学习。E-mail: sunmengccjs@163.com。

张雄伟(1965-), 男, 博士, 教授, 主要研究方向: 语音与图像处理、智能信息处理。

版权声明