说话人重识别中的基频和共振峰联合还原方法

魏春雨,孙　蒙,贾　冲
(陆军工程大学 指挥控制工程学院,江苏 南京 210007)

摘　要:说话人匿名技术的出现,对基于声纹的生物特征识别造成了巨大的安全威胁。对于利用各种变声工具实施的说话人匿名,匿名语音中的说话人个性特征相比原始语音发生了显著改变,会严重影响说话人识别的效果。针对现有说话人重识别方法存在的语音还原手段单一、在变声工具类型未知情况下的匿名语音还原效果尚不明确等问题,提出了一种基于基频和共振峰联合还原的说话人变声匿名重识别方法。该方法在基频逆变换变声还原的基础上,引入 McAdams 系数调整语音的共振峰,同时使用基于 x-vector 的说话人识别模型进行声纹相似度评分,提高了黑盒变声匿名条件下还原语音与真实语音的声学特征相似度,增强了说话人识别系统对不同类型变声匿名语音的重识别能力。实验结果表明,提出的方法对四种音频编辑软件和三种真实变声器材匿名语音的重识别效果均优于现有基线重识别方法。

关键词:说话人匿名;说话人重识别;McAdams 系数;共振峰;还原
中图分类号:TP391.9　　**文献标识码**:A　　**文章编号**:1673-629X(2023)06-0047-07
doi:10.3969/j.issn.1673-629X.2023.06.008

Joint Restoration of Pitch and Formant for Speaker Re-recognition

WEI Chun-yu, SUN Meng, JIA Chong
(School of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China)

Abstract: The emergence of speaker anonymization poses a huge security threat to biometric recognition based on voiceprint. For the speaker anonymization implemented by various voice changing tools, the personality characteristics of the speaker in the anonymous voice have changed significantly compared with the original voice, which will seriously affect the effect of speaker recognition. Aiming at the problems of single speech restoration means and unclear effect of anonymous speech restoration in the case of unknown voice modification tools in existing speaker re-recognition methods, a speaker re-recognition method of anonymous voices joint restoration of pitch and formant based is proposed. Besides the pitch inversion transformation, the proposed method introduces the McAdams coefficient to modify the formant characteristics of voices, and uses the speaker recognition model based on x-vector to calculate the utterance-level similarity score, which improves the acoustic similarity between the restored voice and the real voice under the condition of black box voice changing, and enhances the ability of speaker recognition system to recognize different kinds of anonymous voices. Experimental results show that the proposed method has better performance than the existing baseline method in restoring the anonymous voices generated by four audio editing software and three physical voice changing tools.

Key words: speaker anonymization; speaker re-recognition; McAdams coefficient; formant; restoration

0　引　言

近年来,随着说话人识别技术在精确度和鲁棒性方面不断进步,其在金融、安防、军事和娱乐消费等领域获得了广泛应用。然而,说话人匿名技术的出现对说话人识别的安全应用构成了严重威胁,提高说话人识别的可靠性变得异常迫切。

在司法取证领域,由于语音很容易被改变,目前基于声纹的证据并不具有与签名文件、指纹和 DNA 相同的法律效力,但说话人识别作为一种身份认证的辅助方法仍然可以在案件的调查阶段发挥重要作用。在类似电信诈骗的作案场景中,犯罪嫌疑人没有留下除语音之外的其他线索,通过仅有的电话录音进行可靠的说话人识别变得至关重要。为了隐藏自己的身份,逃避依据声纹开展的身份追踪,多数罪犯在拨打诈骗电话或发出恐怖威胁时,都会对自己的声音进行匿名化处理[1],如何准确识别经过匿名的语音成为了人们

收稿日期:2022-08-10　　**修回日期**:2022-12-16
基金项目:国家自然科学基金(62071484);江苏省自然科学基金(BK20180080)
作者简介:魏春雨(1993-),男,硕士研究生,研究方向为说话人识别、语音鉴伪;通信作者:孙　蒙(1984-),男,博士,副教授,研究方向为智能语音处理、机器学习。

日益关注的课题。

说话人匿名的目标是隐藏语音信号中的个人身份信息,同时尽可能保持语音内容的可理解性不受影响[2]。一些专业的语音处理软件,如 SoundTouch、Audition、Audacity、GoldWave 等,通过使用不同的变调算法对语音进行更改,可以很方便地隐匿说话人的年龄、性别、身份等生物特征信息[3]。随着移动互联网的迅速发展,QQ 变声器、安卓变声应用、各类直播平台的变声设备等可以将录音实时转换成具有不同变声效果的语音,这些变声器在实现音调变换的同时也改变了包括音色在内的其他声音特性。变声匿名语音的基频、元音共振峰频率、发音速率等声学特征发生了显著改变,不仅会严重干扰人耳的听觉辨识,还会大幅降低说话人识别的性能[4]。

说话人重识别是指在与训练数据不匹配的条件下,对具有不同讲话场景、噪声环境以及经过变换处理的语音开展的说话人识别任务[5-7]。以往对变声匿名的研究,多集中在匿名语音的检测[8-10],关注的主要问题是语音是否经过了变声处理和语音基频的变化程度,一定程度上减少了变声匿名语音的输入危害,但却不能从根本上解决由语音匿名造成的说话人识别精度下降的问题。因此,针对匿名语音进行还原,在语音变声匿名背景下进行说话人重识别研究变得非常重要,对于提高说话人识别在用户安全访问、案件调查取证等场景的可靠性具有重要意义。由于语音变声匿名过程中声学特性的变化大致遵循线性规律[11],因此从变声匿名语音中恢复出原始语音具有理论的可行性。

1 相关工作

在针对变声匿名语音的说话人重识别算法设计方面,郑等[12]提出了基于 i-vector 相似度的匿名语音还原方案,他们利用 SoundTouch 软件对变声匿名语音进行遍历基频变换因子的预还原,将预还原的声音与注册说话人的真实语音进行对比,把其中说话人识别分数最高,即预还原语音与真实语音的同一人相似度最高的一条语音作为还原后的语音。该方法在判别匿名语音基频变换因子的同时实现了语音的还原,并在针对变声匿名语音的说话人确认任务中取得了较低的等错误率。然而,这种方法在还原阶段使用与匿名阶段相同的变声工具,这在实际的匿名语音还原场景中是不现实的。

张等[11]通过对变声匿名语音的声学特性进行分析,发现经过变声器处理的语音,基频和共振峰均与原始语音存在一定的线性关系。张等[13]研究发现经过逆变声恢复的匿名语音的 F1 共振峰频率与原始语音仍存在差异。Zhang 等[14]通过对 SoundTouch 音频编辑软件处理的语音进行研究,发现共振峰参数的变化遵循线性规律。Helianti 等[15]的研究表明,经过变声器变声的语音,基频和共振峰同时发生了改变,而 Lutsenko 等[1]的研究表明,一些变声器在改变语音基频之后为了使声音听起来更自然,会对语音共振峰进行额外的校正。

以上的研究结果表明,经过变声匿名的语音,共振峰和基频均会发生较显著的变化。基于某一种变声工具的基频逆变换还原方法,对不同类型变声软件和一些真实变声器材的有效性没有得到验证。现有的基于基频逆变换的变声匿名语音还原方法具有以下两个方面的不足:

(1)不同变声工具所采用的变声算法不同,使得变声匿名语音的种类具有多样性。当不知晓实际变声工具的类型时,即使准确执行了与变声过程相反的基频逆变换,恢复得到的语音频谱特征与原始语音仍存在差异。

(2)不同类型变声器所追求的变声效果不同,导致对语音共振峰的改变有所区别,仅依靠基频逆变换的还原方法,不能完全恢复语音共振峰体现的声音音色特性。

对说话人识别来说,共振峰代表语音所携带的声道信息,是与基频所代表的声门特性同等重要的说话人个性特征。因此,在基频还原的基础上对共振峰进行修复,可以提高变声匿名语音还原方法的鲁棒性。

该文引入 McAdams 系数[16]对变声匿名语音频谱包络进行调整,通过对说话人识别中两个重要的声学特征即基频和共振峰开展联合修复,进一步提高了变声匿名语音还原方法对不同变声工具的鲁棒性。

2 说话人重识别中的基频和共振峰联合还原方法

2.1 说话人识别系统的工作原理

说话人识别是一种基于语音的生物特征识别技术。前沿的说话人识别技术使用说话人嵌入(Speaker Embedding)将说话人的声学特性表示为固定维度的向量。典型的说话人嵌入是基于高斯混合模型(Gaussian Mixed Model,GMM)的 i-vector 和基于深度神经网络的 x-vector。

图 1 显示了说话人识别的一般体系结构,包括三个阶段:训练、注册和识别阶段。在训练阶段,使用大量的来自不同说话人的语音训练背景模型,学习与说话人无关的语音特征分布。在注册阶段,背景模型将每个注册说话人的语音映射到注册嵌入,作为注册说话人的唯一身份标识。在识别阶段,给定未知说话人的语音,从背景模型中提取语音嵌入进行评分。评分

模块测量注册嵌入和测试语音嵌入之间的相似度,根据相似度分数给出识别结果。

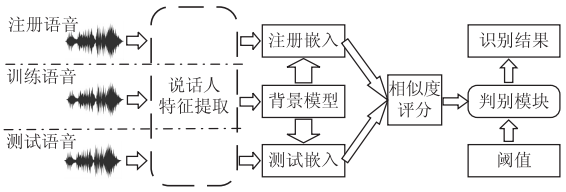

图 1　说话人识别体系结构

说话人确认作为说话人识别中的典型任务,根据预设阈值验证输入语音是否由唯一注册的说话者发出,该任务可以被用来确认嫌疑人声称的语音所代表的身份是否属实。

2.2　基频还原原理

语音变声匿名的典型原理是通过各种方法拉伸或压缩频谱来提高或降低音调。在语音学中,音高最多可以提高或降低 12 个半音,即音高是用 12 个半音来衡量。因此,通常将音高的半音变化量作为基频变换因子。假设语音的基频为 p_0,基频变换因子为 β 半音,修改后的语音的基频为 p ,则有:

$$p = 2^{\beta/12} p_0 \tag{1}$$

如果 β 为正,则音调升高,频谱展宽。否则,频谱被压缩,音调降低。使用基频变换因子 $+\beta$ 来表示音调升高的匿名,使用 $-\beta$ 来表示具有 β 半音的音调降低的匿名。

对变声匿名语音的基频进行还原,文献[12]介绍了一种基于 i-vector 的变声匿名语音还原方法。与文献[12]不同,该文使用当前说话人识别领域最先进的基于 x-vector 的说话人识别模型。相比 i-vector 模型,x-vector 模型具有更高的准确率和更好的鲁棒性。

该文使用基频变换因子 β 来度量变声匿名语音基频的变化程度,通过遍历基频变换因子的理论取值,逐一改变待测语音的基频,得到多条具有不同基频改变程度的还原语音。当基频还原过程中的基频变换因子与语音实际匿名过程中的取值互为相反数,此时对应的还原语音与原始语音具有最高的声纹相似度。对每一条预还原语言和原始注册语音的 x-vector 余弦相似度进行评分,相似度分数最高的预还原语音即为基频被准确还原的语音。

2.3　共振峰还原原理

利用 McAdams 系数进行语音共振峰位置调整,源于在音乐信号处理领域中比较流行的一种声音加法合成技术[17],该技术通过重新合成多个余弦波来产生音色:

$$y(t) = \sum_{k=1}^{K} r_k(t)\cos(2\pi (kf_0)^{\alpha} t + \varphi_k) \tag{2}$$

其中, k 是谐波指数, $r_k(t)$ 是幅值, φ_k 是相位, t 是时间, α 是 McAdams 系数。函数是将谐波余弦信号与逆傅里叶级数相结合,每个谐波都有一定的幅度和一定的相移。McAdams 系数用于调整各次谐波的频率,进而改变声音的音色。通过改变 McAdams 系数调整频谱包络,对语音中的共振峰位置进行变换。共振峰位置的改变程度由 McAdams 系数控制,原理见图 2。

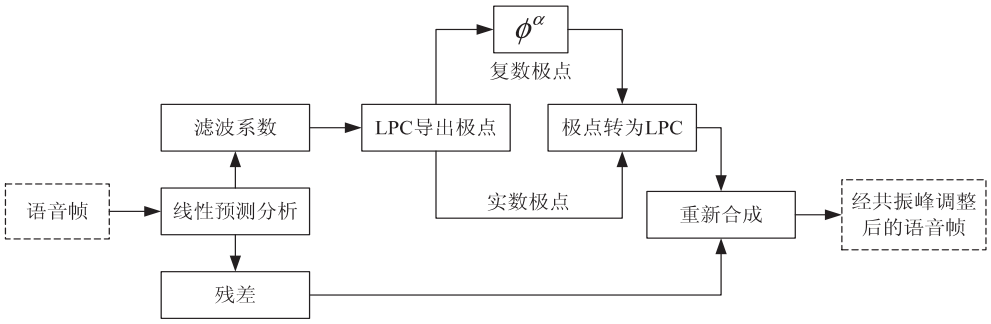

图 2　基于 McAdams 系数的语音共振峰调整

首先,使用线性预测编码(Linear Predictive Coding,LPC)[18]对每一帧进行源滤波器分析。源或残差被用作稍后的再合成,而滤波器系数被用来导出极点位置集。实值极点位置(具有零值虚部)保持不变,而复数极点(具有非零虚部)根据图 2 中指示的方法进行移位。通过公式(2)对正实轴和从 z 平面的原点到复极点的向量之间的夹角进行操作。该角度对应于频率,单位圆的上半部分(π 弧度)对应于采样频率。对复极点位置根据 φ^{α} 进行移位,会导致极点位置的顺时针或逆时针变化。对于 φ 角小于 1 弧度的极点, φ^{α} 的位移 $\alpha<1$ 是逆时针方向,而 $\alpha>1$ 的值则引起顺时针方向的位移。φ 角大于 1 弧度时会朝相反的方向移动。对于实验所涉及到的 16 kHz 采样音频数据, $\varphi=1$ 弧度的值对应于大约 2.5 kHz 的频率,在这个频率两侧共振峰位置移动具有相反的方向,位移的大小取决于 φ^{α} 和 $\varphi=1$ 弧度之间的间距,间距越大,位移越大。

将原始实值极点和修改后的新极点集合转换回 LPC 系数,并与来自原始语音信号的残差组合,用于在时域中重新合成语音帧。然后,通过重叠相加(Overlap and Add,OLA)技术[19]组合每个根据上述过程处理的语音帧,以产生最终的语音信号。

经过 McAdams 系数调整的声音频谱包络的变化情况见图 3。图中显示了发元音“a”时的 5 条包络曲线,包括当 $\alpha \in \{0.85, 0.95, 1.05, 1.15\}$ 时共振峰移位后的频谱。每个极点和峰值大致对应于一个语音共振峰,可以看到随着系数值的变化,频谱包络中相应的峰值也出现了明显的移动。

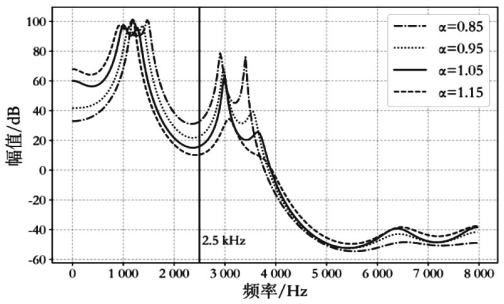

图 3 语音频谱包络移位情况

在利用 McAdams 系数对语音共振峰进行修复的过程中,遍历 McAdams 系数在指定范围内的取值,得到多条共振峰预还原语音,然后利用说话人识别模型对每一条共振峰预还原语音和注册语音的相似度进行打分,相似度分数最高的预还原语音即为最终得到的共振峰修复语音,原理图见图 4。

2.4 联合优化还原方法

由于使用的变声算法不同,即使是在基频变换因子相同的情况下,不同变声工具对语音频谱包络的改变也是不同的。在实际的变声匿名还原场景中,用于变声的设备类型未知,还原的难度进一步增大。即使是使用相同的变声工具对匿名语音进行基频逆变换,变换后得到的语音频谱结构与原始语音也并不完全一致。

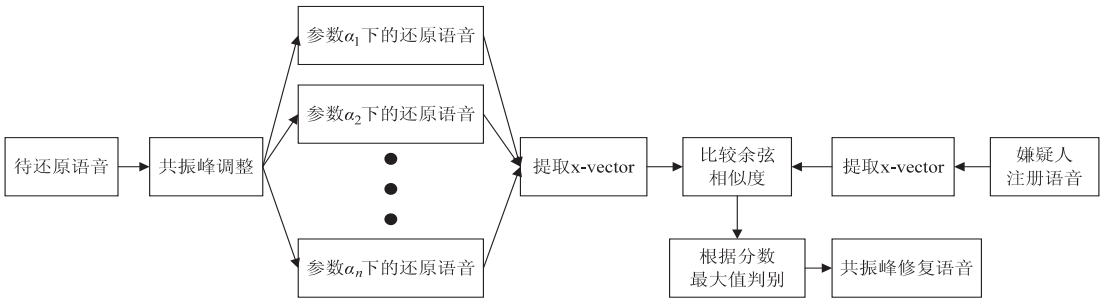

图 4 基于 McAdams 系数的变声匿名语音共振峰修复原理

人类的发声器官在工作时,首先通过声带的震动产生声音基频。随后,声音经过口腔进行传播,不同人口腔生理结构的差异会形成不同的共振峰频率,而声音的基频在口腔传播过程中没有改变。也就是说,不同人声音共振峰的差异受声带振动和声道的共同影响,而基频只与声带振动有关。语音基频的变化会自然引起语音共振峰的改变,而从公式(2)中可以看出,基于 McAdams 系数的共振峰调整并不会带来语音基频的改变。

因此,在具体的变声匿名语音还原过程中,在 2.2 基频逆变换还原方法基础之上,利用 2.3 中介绍的 McAdams 系数对语音共振峰进行参数化调整,对实施基频预还原后语音的共振峰进行修复。选择先基频、再共振峰的修复顺序,以避免在同时还原或先共振峰、再基频的修复过程中,对基频的调整会伴随语音共振峰的改变,进而影响最终对变声匿名语音的还原效果。相比文献[12]的还原方法,文中方法先后对变声匿名语音的基频和共振峰进行修复,更好地实现了对语音变声匿名逆变换过程的近似。变声匿名语音基频、共振峰联合还原的原理见图 5。

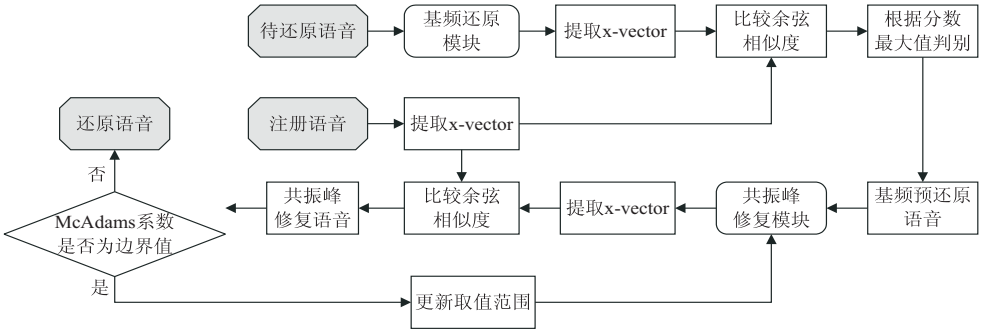

图 5 变声匿名语音基频和共振峰联合还原框架

在基频还原阶段,基频变换因子在理论取值范围内进行遍历,包括+3 ~ +11 和−3 ~ −11 共 18 个参数,这些参数包含了变声匿名语音基频的实际变化范围。对于在共振峰修复过程中 McAdams 系数的取值范围,由于不同变声匿名语音的共振峰差异波动较大,且较小的 McAdams 系数改变也会带来共振峰较大的偏移

量。因此,在 McAdams 系数取值过程中,首先在长度为 0.1 的 0.95 ~ 1.05 的范围内以 0.01 的间隔对 McAdams 系数进行取值,共有 11 个参数。通过遍历这 11 个参数对语音的共振峰进行调整。对于在第一轮的遍历调整中找出的与原始语音相似度分数最高的语音,如果其 McAdams 系数的取值不为 0.95 或 1.05,即不在取值范围的边界,则这条语音就作为最终的还原语音。若找出的相似度分数最高的语音 McAdams 系数取到了边界值,就在该边界值的一侧长度为 0.1 的范围内,以 0.01 为间隔再次对 McAdams 系数进行取值。例如,若在首次修复时 McAdams 系数取值为 0.95,则对 0.85 ~ 0.95 范围内的 11 个参数值进行遍历,以对基频预还原语音进行共振峰修复,直到 McAdams 系数没有取到区间边界值为止。

联合还原具体步骤如下:

(1)使用测试说话人的正常语音在预训练的 x-vector 说话人识别模型进行注册,得到测试说话人的真实语音特征。

(2)利用 2.2 的方法对变声匿名语音的基频进行还原,得到与原始语音基频特征最相似的预还原语音。

(3)对得到的基频预还原语音,在 0.95 ~ 1.05 的范围内取值 McAdams 系数,调整语音共振峰位置,对应每一条基频预还原语音得到 11 个待测的共振峰修复语音。

(4)利用说话人识别模型分别提取每一个共振峰修复语音和注册语音的 x-vector,进行相似度打分,找出得分最高的共振峰修复语音。

(5)如步骤(4)中找出的声纹相似度分数最高的共振峰修复语音对应的 McAdams 系数没有取到边界值,则这条语音作为最终的还原语音。否则,在上一次取边界值的一侧长度为 0.1 的范围内,以 0.01 为间隔,再次取值 11 个参数对步骤(2)中得到的基频还原语音共振峰进行遍历调整。

3　实验与结果分析

3.1　实验设置

3.1.1　变声匿名器材设置

为了验证文中变声匿名语音还原方法的有效性,使用四种音频编辑软件和三种真实的商用变声器对语音进行匿名。四种音频编辑软件包括 SoundTouch、Audition、Audacity、GoldWave,这四种音频编辑软件都可以按照设定对语音进行不同程度的基频变换,变换的程度以半音为单位,对应不同的基频变换因子。实验中使用 SoundTouch 软件进行基频预还原,所有音频编辑软件生成的匿名语音均为 16 kHz 采样、16 bit、单声道、PCM-WAV 格式。

三种真实变声器包括一款直播用硬件变声设备、QQ 变音 APP 和基于 Windows 平台的变声精灵。其中硬件变声设备具有四档不同的变声模式,包括音调变高和变低各两种不同程度的变声设置。变声精灵具有“男声变女声”“男声变女孩”“女声变男声”三种变声模式。QQ 变音包括“萝莉音”“大叔音”“惊悚”“搞怪”“空灵”五种变声模式。实验中将变声器匿名的录音均转换为 16 kHz 采样、16 bit、单声道、PCM-WAV 格式。

3.1.2　数据集

为了公平比较,在四种音频编辑软件的匿名还原实验中,利用与文献[12]相同的方式选取 VoxCeleb1[20] 数据集的语音进行注册和匿名还原测试。VoxCeleb1 数据集取自 YouTube 网站的英文语音。语音中含有各种真实场景噪声,在每一段语音中随机出现的噪声包括环境噪声、周围人的话音、交谈中的笑声、多段话音的混叠、录音设备产生的噪声和回声等,说话场景包括明星采访、公开演说、真人访谈、运动赛事解说等。

在针对真实变声器材进行的匿名还原实验中,为了贴近真实的中文变声录音场景,选取 CN-Celeb[21] 数据集中的中文语音作为与测试说话人不匹配的混淆语音在说话人识别模型中进行注册。CN-Celeb 中文语音数据集包括娱乐节目、访问、唱歌、戏剧、电影、视频博客、现场直播、演讲、戏剧、朗诵和广告共 11 种语音场景,相比 VoxCeleb1 数据集,CN-Celeb 数据集含噪语音场景更加多样。

3.1.3　说话人识别系统设置

说话人识别系统使用在语音识别平台 Kaldi 预训练的基于 TDNN 的 x-vector 模型[22],该开源模型的训练数据集为 Voxceleb1。在说话人确认任务下进行了匿名还原实验,验证待测的变声匿名语音是否来自于注册说话人。

在四种音频编辑软件匿名语音还原实验中,在 VoxCeleb1 数据集中随机选取 400 名说话人,每个说话人各 11 条语音,其中 10 条语音在 x-vector 说话人识别模型上进行注册,1 条语音用作匿名还原测试。每一条测试语音都利用这四种音频编辑软件进行不同程度的音调变换,产生基频变换因子为 +3 ~ +11 和 -3 ~ -11 的 18 组匿名语音。

在真实变声器匿名还原实验中,为了贴近中文语音匿名还原场景,在 CN-Celeb 数据集中随机选择 100 名中文说话人的各 10 条语音在说话人识别模型上进行注册。在实验室录制 3 名说话人的各 12 条语音,其中 2 条语音在说话人识别模型上注册,其余 10 条语音在三种变声器的不同变声设置下进行匿名处理,得到

具有不同变声效果的匿名语音。

3.2 评估指标

在实验中使用等错误率(Equal Error Rate,EER)来评估还原语音的说话人确认效果。错误接受率(False Acceptance Rate,FAR)定义为错误接受的数量除以错误样本的总数,而错误拒绝率(False Rejection Rate,FRR)定义为错误拒绝数量除以正确样本的总数。当阈值变化时,FRR 和 FAR 变化的趋势相反,FRR 与 FAR 相等时的错误率即为 EER。EER 越低,说话人识别的效果越好。

3.3 实验结果

3.3.1 四种音频编辑软件匿名的还原效果

为了验证提出的变声匿名语音还原方法在说话人确认任务中的效果,在四种音频编辑软件变声条件下,对比了文中方法和文献[12]的方法对 VoxCeleb1 匿名语音还原后的说话人确认等错误率,结果见图 6。从图中可以看出,相比文献[12]的方法,文中方法在绝大多数基频变换因子下的说话人确认等错误率均有所降低。在音调变低和变高两种情况下,当语音匿名与基频预还原使用相同的音频编辑软件 SoundTouch 时,等错误率平均降低 0.1 和 0.25 百分点,使用 GoldWave 匿名时平均降低 0.17 和 0.31 百分点,使用 Audition 匿名时平均降低 0.13 和 0.20 百分点,使用 Audacity 匿名时平均降低 0.26 和 0.35 百分点。实验结果表明,文中方法对不同音频编辑软件的鲁棒性较好,在未知变声软件型号的匿名语音黑盒还原情境下,具有比基线方法更好的说话人重识别效果。

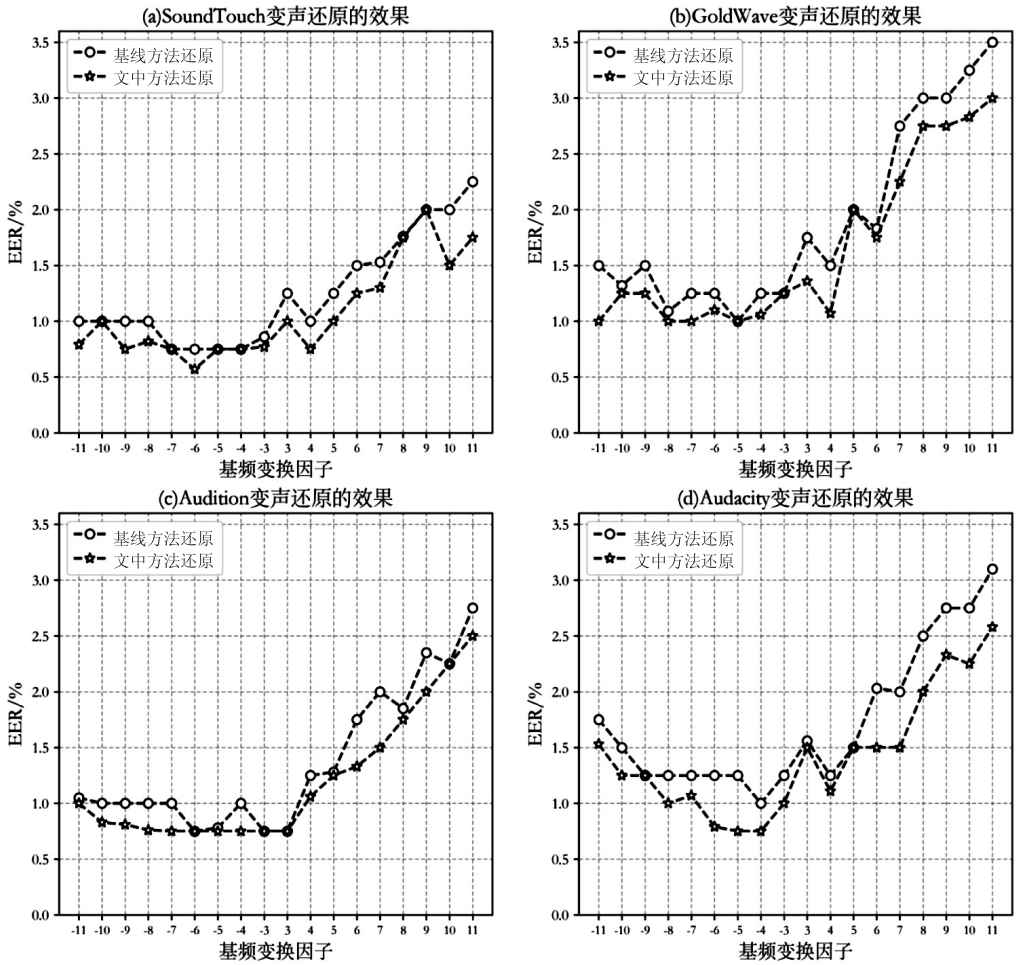

图 6 不同变声软件匿名还原后的等错误率

当变声匿名语音的基频变换因子为-5 和+5 时,从测试语音中随机选取一条原始语音和其对应的变声匿名语音、基频预还原语音以及经过共振峰修复的语音,画出它们的频谱包络曲线,见图 7 和图 8。

可以发现,文献[12]的方法还原之后的语音与原始语音的频谱包络拟合度仍然不高,特别是在高频部分的差异更加明显,在利用文中方法对共振峰进行修复后,语音频谱包络与原始语音更加接近。

3.3.2 真实变声器匿名的还原效果

对采用变声精灵、QQ 变音、硬件变声设备三种变声器材匿名的语音进行还原,说话人确认的效果见图 9。

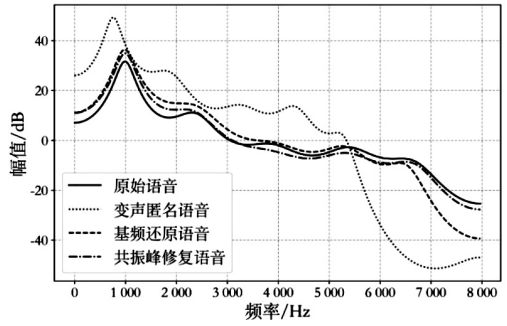

图 7　基频变换因子为-5 时的语音频谱包络对比

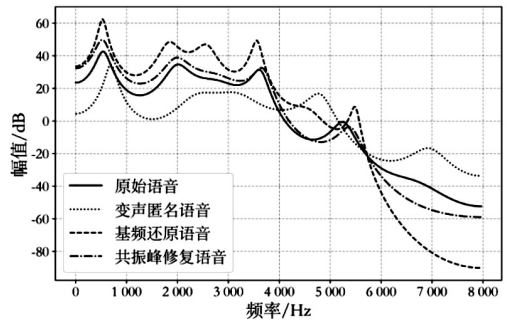

图 8　基频变换因子为+5 时的语音频谱包络对比

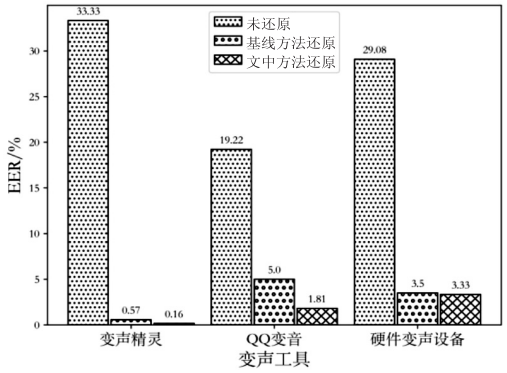

图 9　说话人确认等错误率对比

从图中可以看出,相比文献[12]的还原方法,利用文中方法还原后说话人确认等错误率更低。硬件变声设备的降幅最小,从 3.5% 下降到 3.33%,变声精灵匿名还原之后的等错误率在三种变声器材中最低,从 0.57% 下降到 0.16%。QQ 变音的等错误率降低幅度最大,从 5% 下降到 1.81%,分析原因可能是在 QQ 变音的变声效果中,对声音音色的改动更大,语音共振峰位置出现较大的变化,使得文中方法获得了更加显著的效果。

实验结果表明,文中方法在真实商用变声器匿名条件下同样具有更好的说话人重识别效果。

4　结束语

针对现有变声匿名说话人重识别方法仅修复语音基频,对未知变声软件的黑盒匿名以及在真实变声器匿名条件下的语音还原效果尚不明确等问题,提出了一种基于基频和共振峰联合还原的变声匿名说话人重识别方法。在基频逆变换语音还原的基础上,利用 McAdams 系数对语音的共振峰位置进行调整,结合基于 x-vector 的说话人识别模型给出的相似度得分,得到与原始语音声学特征最为接近的还原语音。实验结果表明,在黑盒变声匿名条件下,该方法对四种变声软件、三种真实变声器材匿名的语音具有比基线方法更好的重识别效果。

参考文献:

[1] LUTSENKO K, ROMAN A, GRIGORYAN S, et al. Research on a voice changed by distortion[J]. Theory and Practice of Forensic Science and Criminalistics, 2021, 23(1):188-202.

[2] TOMASHENKO N, SRIVASTAVA B M L, WANG X, et al. Introducing the VoicePrivacy initiative[C]//Proceedings of the 21st annual conference of the international speech communication association (INTERSPEECH). Shanghai: ISCA, 2020:1693-1697.

[3] PERROT P, AVERSANO G, CHOLLET G. Voice disguise and automatic detection: review and perspectives[M]//Progress in nonlinear speech processing. Berlin: Springer, 2007: 101-117.

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

[5] JIN Q, TOTH A R, SCHULTZ T, et al. Voice convergin: speaker de-identification by voice transformation[C]//2009 IEEE international conference on acoustics, speech and signal processing (ICASSP). Taipei, China: IEEE, 2009:3909-3912.

[6] SHI Y, HUANG Q, HAIN T. Speaker re-identification with speaker dependent speech enhancement[C]//Proceedings of the 21st annual conference of the international speech communication association (INTERSPEECH). Shanghai: ISCA, 2020:1530-1534.

[7] MAGARINOS C, LOPEZ-OTERO P, DOCIO-FERNANDEZ L, et al. Reversible speaker de-identification using pre-trained transformation functions[J]. Computer Speech & Language, 2017, 46:36-52.

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

[9] CAO W, WANG H, ZHAO H, et al. Identification of electronic disguised voices in the noisy environment[C]//International workshop on digital watermarking. [s. l.]: Springer, 2016:75-87.

[10] WANG L, LIANG H, LIN X, et al. Revealing the processing history of pitch-shifted voice using CNNs[C]//2018 IEEE international workshop on information forensics and security (WIFS). Hong Kong: IEEE, 2018:1-7.

(下转第 60 页)